An Investigation of Weighting Procedures for Unit-Nonresponse

Hyunshik Lee[1] and Jin Kim[2]

[1]Westat, 1600 Research Blvd., Rockville, MD 20850
[2]Statistics Korea, 189 Cheongsa-ro, Seo-gu, Daejeon 35208, Republic of Korea

Abstract

Unit-nonresponse is usually compensated for by weighting adjustment, and calibration techniques are frequently used for benchmarking purposes and for improving survey estimates. Traditionally, nonresponse adjustment and calibration are performed in two separate steps, and this weighting procedure is called the two-step procedure. Lately, there has been considerable discussion on whether the calibration step can take care of the nonresponse adjustment as well in a single step (called the one-step procedure). The answer to this question is not clear cut, and in this paper we want to address this issue for various situations arising in practice in terms of relationship between auxiliary information and survey variables and response mechanisms. This is an empirical investigation built upon the work by Haziza and Lesage (2016).

Key words: Unit-nonresponse, nonresponse bias, calibration, one-step procedure, two-step procedure, response mechanism

1. Introduction

Unit nonresponse in survey data is usually treated through weight adjustment of the base weight, which is the inverse of the selection probability. Assuming that nonresponse is another phase of sampling, the weighting class method, where the weighting classes are formed using categorical auxiliary variables, is often employed for nonresponse weight adjustment. Another commonly used method is the propensity score method, where the response propensity is estimated using a regression model such as the logistic regression model that describes the relationship between the response status and auxiliary variables. To reduce excessive variability of estimated propensity score adjusted weight when directly used, the latter also use the weighting class method by forming the weighting classes based on the estimated propensity scores. To distinguish the latter method from the traditional weighting class (TWC) method, it is called the propensity score based weighting class (PWC) method in this paper. The PWC method is more flexible in using available auxiliary variables than the TWC method because it is limited in using categorical auxiliary variables to form the weighting classes, and also the PWC method enjoys the robustness of the weighting class method. The weighting may end here.

However, if the population totals of some auxiliary variables are available, the nonresponse adjusted weight can be calibrated to the population totals to enhance the efficiency of the survey estimates and provide the credibility of the survey weights by making their weighted sums equal to the population totals of the auxiliary variables. Commonly used calibration weighting methods are post-stratification, raking, the generalized regression (GREG) estimator, and some variants of these (see Deville and Särndal, 1992). This is the usual two-step weighting procedure.

However, recently, several authors proposed the one-step procedure, which skips the first step but uses the calibration step only (e.g., Lundström and Särndal, 1999; Folsom and Singh, 2000; Bethlehem, 2002; Särndal and Lundström, 2005; Kott, 2006; D'Arrigo and Skinner, 2010; Kott and Liao, 2015). The potency of the one-step procedure is also shown by Flores Cervantes and Brick (2008), who compared the two procedures using the 2007 California Health Interview Survey data and concluded that the two procedures are virtually the same for over 700 survey estimates they studied. Under certain conditions, the one-step procedure is (nearly) unbiased. However, Haziza and Lesage (2016) shows that the one-step procedure can be severely biased. Extending their study, we investigate the same issue with more general and realistic situations.

In the section that follows, we discuss the one- and two-step procedures in detail. In Section 3, these two weighting procedures are compared. Some concluding remarks are provided in Section 4.

2. One- and Two-step Weight Adjustment Procedures for Unit Nonresponse

We are interested in estimating the population total of a survey variable y. The population U has N units indexed by i, from which a sample of size n is selected with a pre-determined probability π_i (> 0) for unit i. The base weight is given by $d_i = \pi_i^{-1}$, and if there is no nonresponse, the population total $T_y = \sum_{i=1}^{N} y_i$ can be estimated without bias by

$$\hat{T}_y = \sum_{i=1}^{n} d_i y_i$$

However, when some sampled units do not respond, the unadjusted estimator based on the respondent sample, R, of size n_R,

$$\hat{T}_{Uy} = N \sum_{i=1}^{n_R} d_i y_i \Big/ \sum_{i=1}^{n_R} d_i$$

is biased unless the response mechanism is uniform (i.e., every sampled unit has the same propensity of response). Note that it still uses a simple adjustment by a single factor, so we mean by "unadjusted" that the adjustment is too simplistic and insensitive to differential response propensity. To address the nonresponse bias issue, as mentioned in the introduction, two commonly used methods are the traditional weighting class (TWC) method and the propensity score based weighting (PWC) class method. The TWC method is based on the assumption that the sampled units respond under the quasi-randomization mechanism (Oh and Scheuren, 1983), and that the sampled units have the same response probability within the weighting classes. The weighting classes are formed as cross-classes of categorical auxiliary variables available for both respondents and nonrespondents, which are supposed to be predictive of the response probability. The PWC is based on a somewhat different assumption that sampled units respond independently of each other according to a Poisson process. This requires estimation of the individual response probability (propensity), and this is usually done by the logistic regression that relates the response status as the dependent variable and all auxiliary variables (either categorical or continuous) available for both respondents and nonrespondents as dependent variables. The propensity score modelling admits continuous auxiliary variables, whereas the TWC method has to use only categorical variables. Of course, a continuous variable can be categorized to use as a class defining

variable but some loss of information results through categorization. Furthermore, the number of categorical variables should be limited to avoid defining too many weighting classes with small sizes – when this happens, small cells are collapsed but it still limits the number. In any case, the PWC method is more flexible in terms of using the auxiliary variable for nonresponse adjustment. However, estimated propensity score can be very small, and this causes very large adjusted weights, resulting in unstable variance. This can be prevented by grouping respondents with similar estimated propensity scores into a number of classes and using the weighting class method to perform the nonresponse adjustment. In this paper, we always use the PWC method for the nonresponse adjustment.

Denoting the adjustment factor by a_i for respondent i, the nonresponse adjusted estimator is given by

$$\hat{T}_{Ay} = \sum_{i=1}^{n_R} a_i d_i y_i$$

If the population totals are available for some auxiliary variables, calibration weighting can be applied to the nonresponse-adjusted weight. In this case, calibration is regarded as a weighting tool (rather than means of nonresponse adjustment) to enhance the efficiency of the estimates and the credibility of the final survey weight by forcing the weighted sums of the auxiliary variables from the respondent sample equal (calibrated) to the population totals. Let $\mathbf{x_i} = (x_{1i}, x_{2i}, \dots, x_{pi})^T$ be a p-dimensional vector of auxiliary variables with the population total $\mathbf{T_x} = (T_{x_1}, T_{x_2}, \dots, T_{x_p})$. Then the calibration weight, w_i, establishes the following equation:

$$\sum_{i \in R} w_i \mathbf{x}_i = \mathbf{T_x}$$

The calibration weight is obtained from the calibration equation given by

$$w_i = a_i d_i F(\hat{\boldsymbol{\lambda}}^{\mathrm{T}} \mathbf{x}_i)$$

where $F(\cdot)$ is a monotonic and twice-differentiable function such that $F(0) = F'(0) = 1$ (F' is the derivative of F) and $\hat{\boldsymbol{\lambda}}$ is a p-vector of calibration coefficients (Deville and Särndal, 1992). Commonly used calibration functions are: (1) the linear function (from which the GREG estimator is derived) given by

$$F(v) = 1 + v$$

(2) the exponential function given by

$$F(v) = \exp(v)$$

(3) the logit function given by

$$F(v) = \frac{L(U-1) + U(1-L)\exp(Av)}{U - 1 + (1-L)\exp(Av)}$$

where $L < 1 < U$ are the user-specified lower and upper bounds to limit the calibration weight and $A = (U-L)/(1-L)(U-1)$. This is a bounded version of (2) as $F(v)$ in (3) converges to $F(v)$ in (2) as $L \to 0$ and $U \to \infty$. The weighting method described above is a typical two-step weighting procedure. The first step is intended to eliminate the nonresponse bias, and the second step is to improve the efficiency and credibility of the survey estimates.

However, some authors started seeing the calibration weighting as a tool for nonresponse

adjustment in a single step bypassing the usual nonresponse adjustment step (Fuller et al., 1994; Lundström and Särndal, 1999; Bethlehem, 2002; Särndal and Lundström, 2005; Kott, 2006; D'Arrigo and Skinner, 2010; Kott and Liao, 2012; Kott and Liao, 2015). The original form was based on the liner calibration function, where the one-step procedure is asymptotically unbiased if the survey variable is linearly related with the auxiliary variables. Using the same concept that Kim and Park (2006) use in the context of imputation, Kott and Liao (2015) demonstrate that the one-step procedure can be doubly protected against the nonresponse bias if either the model for the survey variable is linear or the response mechanism (the response propensity) is inversely related with the calibration function. The first condition (referred to as A1) is stated as:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \boldsymbol{\epsilon}_i$$

where $\mathrm{E}(\boldsymbol{\epsilon}_i|\mathbf{x}_i) = 0$. Haziza and Lessage (2016) provide a weaker condition than this, which is a special case. The second condition (referred to as A2) is given by:

$$F(\boldsymbol{\lambda}^T \mathbf{x}_i) = \phi_i^{-1}$$

where ϕ_i is the response propensity for unit i. If a survey is multi-purpose with many survey variables, then not all survey variables would satisfy A1. Therefore, A2 is particularly important in this case because if A2 is satisfied, the one-step procedure is asymptotically unbiased for all survey variables. However, Haziza and Lessage (2016) shows that the double protection can fail with a serious bias consequences and emphasize that the two-step procedure is free to choose a nonresponse adjustment procedure for that purpose only and a calibration procedure for calibration only, whereas the one-step procedure has to take care of both nonresponse adjustment and calibration in one shot, and hence more burden is imposed on the procedure. They used an artificially generated population with a single auxiliary variable that follows a uniform distribution. In the following section, we reproduce their results and expand the simulation with other distributions for the single auxiliary variable. We also examine the performance of the two estimators using the 2014 public use micro sample (PUMS) provided by the Census Bureau based on the 5 year American Community Survey (ACS).

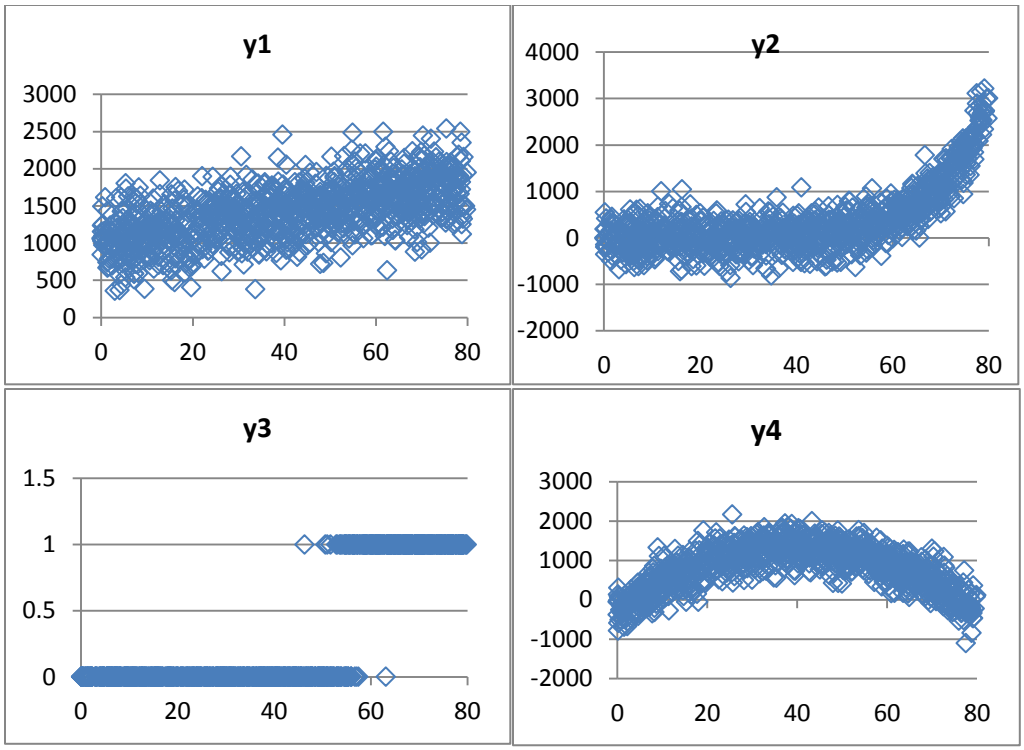
3. Comparison of the One- and Two-Step Procedures

3.1 Replication of the Simulation by Haziza and Lessge (2016)

To compare the one- and two-step procedures through simulation, Haziza and Lessage (2016) generated four artificial populations of size $N = 1{,}000$, using the following four models:

(M1) Linear: $y_1 = 1{,}000 + 10x + \varepsilon_1$;
(M2) Exponential: $y_2 = \exp(-0.1 + 0.1x) + \varepsilon_2$;
(M3) Logistic: $y_3 \sim B(p)$, which follows a Bernoulli distribution with $p = [\exp\{-0.5(x - 55)\} + 1]^{-1}$;
(M4) Quadratic: $y_4 = 1300 - (x - 40)^2 + \varepsilon_4$

where x follows a uniform distribution over (0, 80), ε_k follows a normal distribution with mean 0 and variance 300, for $k = 1,2,4$. Focusing on the nonresponse error, they used the census case (i.e., $n = N = 1{,}000$), and thus $d_i = 1$ for all $i = 1,2, \ldots, N$. The scatter plots of the four y-variables are shown below:



F1gure 1. Plot of *y*-variables against the uniform *x*-variable

Respondents were generated using the following response mechanisms:

(R1) Inverse linear: $\phi_1 = (1.2 + 0.024x)^{-1}$
(R2) Exponential: $\phi_2 = \exp(-0.2 - 0.014x)$
(R3) Logistic type: $\phi_3 = 0.2 + 0.6\{1 + \exp(-5 + x/8)\}^{-1}$
(R4) Quadratic: $\phi_4 = 0.7 + 0.0025x - 0.45(x/40 - 1)^2$

Response indicator, $R_k \sim B(\phi_k)$, for k = 1, ...,4, was generated with the expected overall response rate of 50%. Because the response indicator was generated independently (i.e., Poisson sampling), the respondent sample size fluctuates around 500. These response mechanisms are depicted in the following graph.

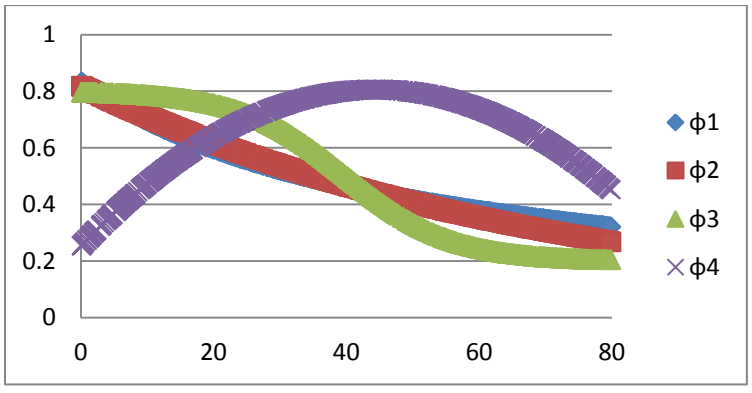


Figure 2. Graphs of four response mechanisms

Four estimators were compared, the PWC estimator and three calibration estimators. Haziza and Lessage (2016) omitted the calibration step for the two-step procedure to compare more clearly the bias removing properties of the procedures. Therefore, the PWC estimator, which is supposed to represent the two-step procedure, was computed without the calibration step. The PWC estimator is defined as follows:

- Run logistic model, $\text{logit}(R) = \beta_0 + \beta_1 x$;
- Partition the estimated response propensities, that is, predicted values of $\exp(\hat{\beta}_0 + \hat{\beta}_1 x)/\{(1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)\}$ into 20 equal size weighting classes;
- Post stratification weight adjustment using 20 weighting classes as post-strata.

Usually a smaller number of weighting classes (between 5 and 10) is recommended but the larger number of 20 was used by Haziza and Lessage (2016) to handle severely nonlinear response mechanisms. We also tried 10, and the result (not shown) was slightly more biased but very close to the result with 20.

The one-step procedure applies the calibration estimation directly without the nonresponse adjustment step. We use three calibration functions discussed in Section 2:

- Linear (CAL1): $1 + v$
- Exponential (CAL2): $\exp(v)$
- Logit (CAL3): $\frac{L(U-1)+U(1-L)\exp(Av)}{U-1+(1-L)\exp(Av)}$

We used the calibration software, CALMAR, developed by INSEE of France, with auxiliary vector $\mathbf{x} = (1, x)$. This means that the calibration estimators are calibrated to the population size (N) and the population total of x (T_x). The four estimators are denoted as $\hat{T}_{Py}$, $\hat{T}_{C1y}$, $\hat{T}_{C2y}$, and $\hat{T}_{C2y}$, or referred to as P-estimator, C1-estimator, C2-estimator, and C3-estimator. To measure the bias of an estimator $\hat{\theta}$, we compute the Monte Carlo percent relative bias(RB) and the percent relative root mean square error(RRMSE) of $\hat{\theta}$ with M being the simulation size as follows:

$$RB_{MC}(\hat{\theta}) = \frac{100}{M}\sum_{m=1}^{M}\frac{(\hat{\theta}_{(m)} - \theta)}{\theta}$$

$$RRMSE_{MC}(\hat{\theta}) = 100 \times \frac{\{M^{-1}\sum_{m=1}^{M}(\hat{\theta}_{(m)} - \theta)^2\}^{1/2}}{\theta}$$

The simulation results are summarized in Tables 1-4, where these abbreviations are used:

- R: Response mechanism
- Var: y-Variable
- U: Unadjusted estimator ($\hat{T}_{Uy}$) based on respondents without weight adjustment
- P: One-step PWC estimator based on estimated propensity scores
- C1: One-step calibration estimator based on the linear function
- C2: One-step calibration estimator based on the exponential function
- C3: One-step calibration estimator based on the logit function

C2 and C3 performed almost identically with the upper bound (U) we used to avoid the situation, where CALMAR could not find a solution during simulation. So, the result for C3 is omitted from the tables.

Table 1. Simulation results with the single uniform auxiliary variable

R	Var	Relative Bias (RB)				Relative Root Mean Square Error (RRMSE)			
		U	P	C1	C2	U	P	C1	C2
1	y1	-4.93	0.03	0.01	0.01	5	0.77	0.76	0.75
1	y2	-32.35	-0.16	-0.56	2.68	32.87	2.95	5.57	6.02
1	y3	-29.8	-0.22	-0.35	1.66	30.18	2.43	3.67	3.82
1	y4	-4.23	-0.13	0.23	-2.14	4.77	1.36	2.83	3.34
2	y1	-5.7	-0.01	-0.02	-0.01	5.76	0.83	0.81	0.81
2	y2	-39.17	-0.06	-4.21	0.16	39.6	3.02	7.61	6.28
2	y3	-35.95	-0.27	-2.86	-0.04	36.23	2.47	4.66	3.42
2	y4	-2.85	-0.05	3.24	-0.05	3.71	1.39	4.5	2.88
3	y1	-7.86	0.02	-0.01	-0.01	7.9	0.94	0.86	0.88
3	y2	-55.56	0.05	-10.63	-0.63	55.8	3.48	12.72	6.82
3	y3	-54.91	-0.13	-13.12	-6.32	55.08	3.09	13.89	7.42
3	y4	0.81	-0.02	12.47	4.67	2.27	1.55	13.04	5.6
4	y1	2.2	-0.01	0	0.01	2.35	0.83	0.67	0.67
4	y2	-11.57	-0.06	-28.57	-27.56	13.05	3.19	28.85	27.85
4	y3	-0.92	-0.08	-17.8	-17.17	4.84	1.47	18.02	17.4
4	y4	19.05	0.32	20.85	20.19	19.21	1.74	21	20.35

We obtained similar results as shown in Haziza and Lesage (2016). Some highlights are:

- The PWC estimator is virtually unbiased and almost always better than calibration estimators except for y_1 variable, which is linear in x, and for which all estimators are unbiased as the theory predicts.
- The C1 estimator works well in terms of the bias under the inverse linear response mechanism (R1) because the inverse of the calibration function estimates the response propensity. Likewise, the C2 estimator does well under the exponential response mechanism (R2).
- Under other response mechanisms (R3 and R4), the calibration estimators do not perform well except for y_1, for which any estimator is supposed to do well.
- When the double protection is present, the calibration estimators are safe from the nonresponse bias but in general less efficient and vulnerable in the absence of double protection.

Considering the uniform auxiliary variable is unusual, we also ran the simulation using the same set-up but with non-uniform auxiliary variables. To see how the performance of the estimators changes as the auxiliary variable moves away from uniform to a symmetric distribution with thinner tails, we used the Trapezoidal distribution over (0, 80) defined by the following density function with an arbitrary number a in (0, 40) and $b = \frac{a}{2(80-a)}$:

$$f(x) = \begin{cases} bx/a & \text{if } x < a \\ b & \text{if } a \le x < 80 - a \\ b(80-x)/a & \text{if } x \ge 80 - a \end{cases}$$

The density function of the Trapezoidal distribution is shown below. Note that when $a = 0$, it becomes the uniform, and when $a = 40$, it becomes the triangular.

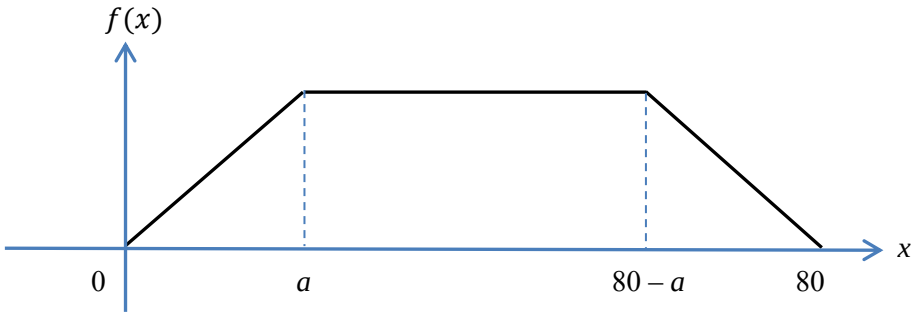


Figure 3. The distribution function of the trapezoidal distribution

We ran simulation with $a = 10, 20, 30, 40$, but to save space, we present the results for a = 20 and 40 only.

Table 2. Simulation results with the Trapezoidal auxiliary variable with $a = 20$

R	Var	Relative Bias (RB)				Relative Root Mean Square Error (RRMSE)			
		U	P	C1	C2	U	P	C1	C2
1	y1	-2.60	0.05	0.08	0.08	2.71	0.67	0.66	0.66
1	y2	-21.24	-0.60	0.01	2.11	22.49	5.93	7.47	7.76
1	y3	-23.01	0.02	0.01	1.78	23.74	3.38	4.89	5.06
1	y4	-2.02	0.00	0.02	-0.72	2.53	1.12	1.66	1.77
2	y1	-3.16	-0.02	0.01	0.01	3.26	0.73	0.71	0.72
2	y2	-26.54	-0.17	-2.53	0.35	27.48	6.32	7.87	7.48
2	y3	-29.13	-0.27	-2.68	-0.14	29.71	3.59	6.01	5.18
2	y4	-1.70	-0.07	1.00	-0.06	2.17	1.08	1.88	1.51
3	y1	-5.27	-0.14	-0.10	-0.10	5.32	0.84	0.79	0.81
3	y2	-44.87	-0.30	-9.28	-0.59	45.33	6.74	12.73	9.04
3	y3	-51.55	-0.20	-13.58	-5.88	51.82	4.51	14.99	8.28
3	y4	-0.77	0.06	5.43	2.14	1.49	1.17	5.78	2.78
4	y1	1.26	0.02	0.03	0.03	1.49	0.77	0.69	0.69
4	y2	-6.49	-1.18	-18.10	-17.52	9.39	5.72	18.97	18.41
4	y3	-2.13	-0.16	-15.49	-14.95	5.94	2.48	15.91	15.39
4	y4	5.97	0.42	6.24	6.05	6.09	1.11	6.36	6.18

Table 3. Simulation results with the Triangular auxiliary variable ($a = 40$)

R	Var	Relative Bias (RB)				Relative Root Mean Square Error (RRMSE)			
		U	P	C1	C2	U	P	C1	C2
1	y1	-2.23	-0.03	-0.01	0.01	2.37	0.77	0.75	0.75
1	y2	-20.16	-0.49	-0.31	1.52	21.58	6.04	7.47	7.58
1	y3	-22.63	-0.07	0.08	1.91	23.58	3.32	5.54	5.68
1	y4	-0.94	0.01	0.04	-0.57	1.64	1.05	1.53	1.56
2	y1	-2.69	-0.01	-0.02	0.00	2.81	0.78	0.76	0.76
2	y2	-25.86	-1.22	-3.22	-0.64	26.90	6.43	8.18	7.51
2	y3	-28.74	0.02	-2.54	0.10	29.41	3.36	6.10	5.30
2	y4	-0.55	0.02	0.88	-0.01	1.38	1.03	1.75	1.44
3	y1	-4.68	-0.01	-0.14	-0.07	4.74	0.81	0.77	0.76
3	y2	-43.75	-0.72	-9.13	-0.90	44.26	6.59	12.48	8.84
3	y3	-52.13	0.00	-13.68	-5.14	52.43	4.02	15.27	8.01
3	y4	0.44	-0.09	4.54	1.59	1.28	1.23	4.93	2.37

R	Var	Relative Bias (RB)				Relative Root Mean Square Error (RRMSE)			
		U	P	C1	C2	U	P	C1	C2
4	y1	0.71	0.03	-0.12	-0.12	0.99	0.68	0.64	0.64
4	y2	-7.02	-1.65	-15.50	-15.18	9.82	5.99	16.54	16.23
4	y3	-5.22	-0.29	-15.42	-15.07	7.86	2.25	15.92	15.57
4	y4	4.64	0.39	4.99	4.90	4.78	1.05	5.13	5.03

From these tables, we can see that as the distribution of the *x*-variable moves from the uniform to the triangular distribution, the bias of all estimators is generally reduced. Exceptions are that the PWC estimator performs slightly worse for *y*2-variable and that the bias for *y*3-variable virtually does not change – the bias for this variable (generated from the logistic model) is not much affected by non-uniform symmetric distribution.

When we used a more realistic normal distribution for the auxiliary variable, the result is very similar to that for the triangular auxiliary variable (compare Tables 3 and 4).

Table 4. Simulation results with the normal auxiliary variable

R	Var	Relative Bias (RB)				Relative Root Mean Square Error (RRMSE)			
		U	P	C1	C2	U	P	C1	C2
1	y1	-1.24	-0.06	-0.04	-0.04	1.44	0.72	0.70	0.70
1	y2	-12.47	0.08	-0.17	1.03	15.91	9.52	10.02	10.06
1	y3	-20.75	-0.03	-0.72	1.43	22.69	6.84	9.36	9.36
1	y4	-0.71	-0.05	0.09	-0.13	1.16	0.86	0.96	0.96
2	y1	-1.41	-0.01	0.01	0.02	1.55	0.67	0.66	0.66
2	y2	-16.48	-1.11	-2.50	-0.87	18.98	9.19	9.94	9.61
2	y3	-26.18	-0.50	-3.45	-0.49	27.73	6.87	10.46	9.81
2	y4	-0.67	-0.02	0.38	0.06	1.11	0.84	1.06	0.97
3	y1	-2.61	-0.05	-0.03	-0.02	2.69	0.75	0.71	0.71
3	y2	-30.88	-0.77	-9.23	-3.82	32.27	11.13	14.35	11.70
3	y3	-49.36	-0.61	-15.85	-6.14	50.13	9.01	19.92	13.64
3	y4	-0.59	0.00	1.91	0.84	1.08	1.00	2.23	1.39
4	y1	0.32	0.07	0.03	0.03	0.77	0.69	0.66	0.66
4	y2	-3.52	-0.26	-7.24	-7.11	10.68	9.52	12.06	11.98
4	y3	-8.81	-0.57	-14.90	-14.64	12.72	5.81	16.73	16.48
4	y4	1.41	0.19	1.39	1.37	1.68	0.90	1.66	1.64

We also used the gamma distribution for the auxiliary variable. The result is shown in Table 5, which is quite surprising. The bias of all variables is reduced substantially by all estimators except for the *y*2-variable, for which the bias has gotten much worse. This shows that the interaction between the model that generates the survey variable and the model that describes the distribution of the auxiliary variable can seriously affect the performance of the one-step procedures, which can be further aggravated by the shape of response mechanism (e.g., R4). But we consider that R3 (logistic type) and R4 (quadratic) are somewhat unusual and not frequently observed in reality.

Table 5. Simulation results with the Gamma auxiliary distribution

R	Var	Relative Bias (RB)				Relative Root Mean Square Error (RRMSE)			
		U	P	C1	C2	U	P	C1	C2
1	y1	-1.09	-0.09	-0.08	-0.08	1.37	0.84	0.82	0.82
1	y2	-18.05	3.79	10.32	15.91	143.46	145.32	145.69	146.22
1	y3	-18.44	0.33	0.33	1.12	20.01	4.20	6.28	6.33

R	Var	Relative Bias (RB)				Relative Root Mean Square Error (RRMSE)			
		U	P	C1	C2	U	P	C1	C2
1	y4	-5.46	0.05	0.09	-0.17	5.83	1.38	1.70	1.69
2	y1	-1.31	-0.07	-0.03	-0.04	1.58	0.87	0.86	0.86
2	y2	-36.46	-10.11	-5.46	3.51	149.85	153.25	153.43	154.93
2	y3	-24.37	0.01	-1.20	0.11	25.59	4.81	6.69	6.45
2	y4	-6.80	0.07	0.46	0.05	7.09	1.38	1.74	1.67
3	y1	-1.46	-0.01	0.10	0.08	1.69	0.89	0.86	0.86
3	y2	-70.85	-27.28	-46.75	-35.13	164.10	165.18	166.87	165.89
3	y3	-35.61	-0.24	-9.54	-7.28	36.31	4.96	11.82	9.99
3	y4	-6.98	0.17	2.26	1.69	7.28	1.45	2.90	2.52
4	y1	1.94	-0.06	0.07	0.04	2.10	0.82	0.79	0.79
4	y2	22.42	-7.13	-70.64	-52.62	142.11	155.12	162.41	155.87
4	y3	33.97	-0.04	-10.16	-8.06	34.80	3.19	11.26	9.29
4	y4	11.94	0.22	3.16	2.29	12.11	1.49	3.51	2.74

3.2 Simulation with PUMS

We used the 2014 ACS Public Use Microdata Sample (PUMS) data to create the population data for simulation. Table 6 provides the PUMS variables we used for the study.

Table 6. PUMS variables used for the study (auxiliary variables are in red)

Variable Name	Variable Label
HINCP	HOUSEHOLD INCOME (PAST 12 MONTH)
D_HHT	HOUSEHOLD/FAMILY TYPE: 1 = MARRIED, 0 = OTHER
D_TEN	TENURE: 1 = OWNED, 0 = NOT OWNED
VALP_CAT	CATEGORIZED PROPERTY VALUE (Int(VALP/100,000))
FINCP	FAMILY INCOME(PAST 12 MONTHS)
VALP	PROPERTY VALUE
GRNTP	GROSS RENT (MONTHLY AMOUNT)
D_HHL	HOUSEHOLD LANGUAGE: 1 = ENGLISH ONLY, 0 = NOT ENGLISH ONLY)
D_NOC	NUMBER OF OWN CHILDREN IN HOUSEHOLD: 1 = YES, 0 = NO
D_WIF	WORKERS IN FAMILY DURING THE PAST 12 MONTHS): 1 = YES, 0 = NO
R18	PRESENCE OF PERSONS UNDER 18 YEARS IN HOUSEHOLD: (0,1)
R65	PRESENCE OF PERSONS 65 YEARS AND OVER IN HOUSEHOLD: (0,1)

The simulation population of size $N = 1{,}000$ was created by selecting a simple random sample from the PUMS data (with 809,302 records). Simulation was run with one auxiliary variable (HINCP), two auxiliary variables (HINCP and D_TEN or VALP_CAT), and three auxiliary variables (HINCP, D_TEN, and D_HHT). As before, we used a census, and for each simulation setup, we generated 500 respondent sets by Poisson sampling with four response mechanisms: (1) inverse linear; (2) exponential; (3) logistic type; (4) quadratic, with an average response rate of 50 percent. The auxiliary vector for the calibration estimators always includes 1 as the first component.

CASE1:*:One auxiliary variable (HINCP)*

We generated 500 Poisson samples of respondents with response probability, ϕ_k using the response indicator $R_k \sim B(1, \phi_k)$ that follows four different response mechanisms as given below with x = HINCP/10000.

(R1) Inverse linear: $\phi_1 = (1.83 + 0.024x)^{-1}$;
(R2) Exponential: $\phi_2 = \exp(-0.58 - 0.015x)$;
(R3) Logistic type: $\phi_3 = 0.17 + 0.35\{1 + \exp(-4 + x/8)\}^{-1}$;
(R4) Quadratic: $\phi_4 = 0.77 + 0.0025x - 0.4(x/40 - 1)^2$;

Four estimators (besides the U-estimator) were computed using the same weighting methods used in Section 3.1: P, C1 (Linear), C2 (Exponential), C3 (Logistic type). The P-estimator was based on this logistic model: $\text{logit}(R) = \beta_0 + \beta_1 x$. The result is shown in Table 7 but the result of D_WIF is omitted to save space and also because it is very similar to that of D_HHL. Likewise, the results for R18 and R65 are omitted because they are similar to that of D_NOC. C2 and C3 performed almost identically with the upper bound (U) we used to avoid the situation, where CALMAR could not find a solution during simulation. So, the result for C3 is omitted in all the result tables (7-11) obtained from the PUMS data.

Table 7. Simulation results with one auxiliary variable (HINCP) from PUMS data

Var	R	Relative Bias (RB)				Relative Root Mean Square Error (RRMSE)			
		U	P	C1	C2	U	P	C1	C2
D_HHL	1	-0.15	0.12	0.12	0.10	1.56	1.61	1.57	1.57
D_HHL	2	-0.34	0.01	0.05	0.01	1.55	1.55	1.52	1.52
D_HHL	3	-0.23	0.01	0.05	0.03	1.58	1.62	1.59	1.59
D_HHL	4	0.73	0.05	0.33	0.25	1.78	1.79	1.76	1.76
D_NOC	1	-0.70	-0.40	-0.22	-0.17	4.16	4.18	4.17	4.17
D_NOC	2	-0.66	-0.30	-0.13	-0.05	4.32	4.44	4.41	4.41
D_NOC	3	-0.62	-0.53	-0.16	-0.12	4.17	4.31	4.21	4.21
D_NOC	4	1.30	0.47	0.01	0.08	3.98	4.18	3.96	3.98
FINCP	1	-7.87	-0.64	0.01	0.02	8.42	1.41	0.25	0.25
FINCP	2	-10.39	-1.12	-0.01	0.00	10.72	1.69	0.27	0.27
FINCP	3	-8.28	-1.47	-0.03	-0.03	8.76	1.97	0.26	0.26
FINCP	4	18.73	-0.19	0.21	-0.05	18.91	0.60	0.53	0.22
GRNTP	1	374.51	-0.24	-0.32	-0.09	374.84	6.79	6.79	6.78
GRNTP	2	372.77	-0.56	-1.10	-0.72	373.07	6.93	7.01	6.95
GRNTP	3	377.02	-0.26	-1.54	-1.28	377.40	7.02	7.25	7.18
GRNTP	4	403.13	0.23	-2.50	-1.84	403.51	7.04	7.29	7.11
VALP	1	24.43	0.25	0.47	0.36	25.12	4.70	4.69	4.71
VALP	2	22.68	0.28	0.89	0.75	23.32	4.53	4.60	4.67
VALP	3	23.31	-0.63	0.38	0.30	23.89	4.43	4.33	4.37
VALP	4	43.82	0.24	1.58	1.00	44.07	3.19	3.70	3.41

Except the unadjusted estimator, all other estimators have near zero bias (absolute RB < 3%). It appears that the double protection is at play in this situation. HINCP and FINCP have a linear relationship with very strong correlation (0.99), and the calibration estimators particularly perform very well for FINCP.

***CASE2**: Two auxiliary variables (HINCP and D_TEN)*
Respondents were generated using the following response mechanisms with x_1 = HINCP/10000 and x_2 = D_TEN:

(R1) Inverse linear: $\phi_1 = (1.88 + 0.01x_1 + 0.2x_2)^{-1}$;
(R2) Exponential: $\phi_2 = \exp(-0.60 - 0.01x_1 - 0.1x_2)$;

(R3) Logistic type: $\phi_3 = 0.17 + 0.35\{1 + \exp(-4 + x_1/8 + x_2)\}^{-1}$;
(R4) Quadratic: $\phi_4 = 0.70 + 0.0015x_1 + 0.005x_2 - 0.4(x_1/40 + x_2 - 1)^2$;

Then we calculated four estimators (besides the U-estimator) using the same weighting methods used in Section 3.1: P, C1(Linear), C2 (Exponential), C3 (Logistic type). The P-estimator is defined using the logistic regression model with two independent variables: $\text{logit}(R) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. The result is summarized in the table below but as in Table 7, the results for D_WIF, R18, R65, and of the C3-estimator are omitted.

Table 8. Simulation results with two auxiliary variables from PUMS data

Var	R	Relative Bias (RB)				Relative Root Mean Square Error (RRMSE)			
		U	P	C1	C2	U	P	C1	C2
D_HHL	1	0.21	-0.05	0.03	0.03	1.63	1.88	1.90	1.62
D_HHL	2	-0.08	-0.27	-0.15	-0.15	1.51	1.83	1.86	1.54
D_HHL	3	-0.04	-0.21	0.04	0.04	1.69	2.00	2.10	1.73
D_HHL	4	-1.33	-0.65	-0.10	-0.10	2.05	2.16	2.15	1.60
D_NOC	1	-1.26	0.11	0.08	0.06	4.64	5.05	5.01	4.42
D_NOC	2	-1.95	-0.54	-0.32	-0.36	4.73	4.82	4.78	4.24
D_NOC	3	-0.93	-0.46	0.29	0.24	4.48	4.88	4.87	4.35
D_NOC	4	7.12	-0.78	0.93	0.66	8.14	4.94	4.95	4.21
FINCP	1	-2.98	-0.24	0.00	0.00	4.18	1.62	1.60	0.26
FINCP	2	-6.59	-0.49	-0.01	0.00	7.14	1.70	1.73	0.27
FINCP	3	-8.30	-1.13	-0.03	-0.02	8.78	2.96	2.75	0.29
FINCP	4	9.11	-0.16	0.06	0.00	9.50	1.42	1.34	0.30
GRNTP	1	377.79	-0.59	-0.02	-0.01	378.20	4.80	5.03	3.70
GRNTP	2	374.37	-0.10	-0.10	-0.12	374.75	5.14	5.79	3.60
GRNTP	3	373.43	-0.93	0.14	0.03	373.77	5.15	5.84	3.42
GRNTP	4	377.27	-1.28	-3.27	-2.68	377.44	3.05	3.42	3.93
VALP	1	27.27	0.15	0.24	0.21	27.76	4.42	4.67	3.91
VALP	2	24.45	-0.35	0.17	0.10	25.01	4.96	5.11	4.36
VALP	3	23.13	-0.76	0.02	-0.02	23.75	5.68	5.74	4.47
VALP	4	43.85	0.17	2.24	1.73	44.19	4.10	4.24	4.19

The bias for all estimators except the unadjusted is still fairly well contained (absolute RB < 4%) although the bias slightly increased for some cases.

***CASE3**: Two auxiliary variables (HINCP and VALP_CAT)*
Respondents were generated using the following mechanisms with an interaction term, where x_1 = HINCP/10000 and x_2 = VALP_CAT.

(R1) Inverse linear: $\phi_1 = (1.88 + 0.01\,x_1 + 0.01x_2 + 0.001\,x_1x_2)^{-1}$;
(R2) Exponential: $\phi_2 = \exp(-0.60 - 0.01x_1 - 0.005x_2 + 0.0003x_1x_2)$;
(R3) Logistic type: $\phi_3 = 0.25 + 0.35\{1 + \exp(-4 + 0.3x_1/8 + 0.4x_2 + 0.1x_1x_2)\}^{-1}$;
(R4) Quadratic: $\phi_4 = 0.67 + 0.001x_1 + 0.001x_2 - 0.2(0.1x_1/40 + 0.03x_2 + 0.0003x_1x_2 - 1)^2$;

Then we calculated five estimators (besides the unadjusted estimator) using the same weighting methods used in Section 3.1: two versions of P, C1 (Linear), C2 (Exponential), C3 (Logistic type). The result is summarized in Table 10. The P-estimator is defined using the logistic regression model with two independent variables with an interaction

term (denoted as P1) or without it (denoted as P2) as given below:

- P1: $\text{logit}(R) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$
- P2: $\text{logit}(R) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

The unadjusted estimator performed similarly as shown in Table 8, and it is not shown in Table 9. Instead both P1 and P2 are shown. P1 and P2 performed very similarly, which means the inclusion of interaction terms does not affect the PWC estimator much. The results for D_WIF, R18, and R65 and of the C3-estimator are omitted to save space.

Table 9. Simulation results with two auxiliary variables from PUMS data

Var	R	Relative Bias (RB)				Relative Root Mean Square Error (RRMSE)			
		P1	P2	C1	C2	P1	P2	C1	C2
D_HHL	1	-0.14	-0.14	0.05	0.03	1.93	1.94	1.63	1.63
D_HHL	2	-0.18	-0.21	-0.08	-0.10	1.77	1.84	1.55	1.55
D_HHL	3	-0.81	-0.71	0.11	-0.03	2.15	2.11	1.52	1.54
D_HHL	4	-0.09	-0.14	0.05	0.05	1.81	1.90	1.62	1.62
D_NOC	1	-0.38	-0.30	-0.33	-0.24	4.63	4.60	4.08	4.07
D_NOC	2	-0.50	-0.27	-0.05	0.00	4.22	4.40	3.90	3.90
D_NOC	3	-1.51	-1.11	-0.97	0.06	5.33	5.28	4.58	4.42
D_NOC	4	-0.37	-0.60	-0.72	-0.59	4.31	4.35	3.99	3.97
FINCP	1	-0.56	-0.62	0.00	0.00	1.87	1.77	0.25	0.25
FINCP	2	-0.45	-0.42	0.00	0.01	1.68	1.71	0.25	0.25
FINCP	3	-1.21	-0.88	0.02	0.02	3.10	2.79	0.26	0.26
FINCP	4	-0.11	-0.11	0.01	0.01	1.10	1.05	0.23	0.23
GRNTP	1	0.68	0.61	-0.61	-0.13	6.60	6.58	6.64	6.53
GRNTP	2	-0.56	0.18	0.10	0.39	7.11	7.46	6.72	6.69
GRNTP	3	5.30	4.14	3.52	8.29	8.29	7.38	9.18	11.06
GRNTP	4	-1.05	-1.04	-2.61	-2.06	6.70	6.81	7.17	6.95
VALP	1	-0.95	-0.62	-0.11	-0.07	3.19	2.89	0.50	0.49
VALP	2	-0.27	-0.03	0.02	0.05	2.88	2.73	0.47	0.46
VALP	3	-2.16	-1.64	0.42	0.85	4.76	4.10	0.87	1.08
VALP	4	0.11	0.05	-0.22	-0.16	1.36	1.25	0.45	0.42

The P1-estimator and P2-estimator similarly performed but the P2-estimator did slightly worse, which is surprising because we expect the other way around. Perhaps, adding the interaction term in the logistic model causes more noise in the estimate of the propensity score when the interaction term in the response model is weak.

***CASE4**: Three auxiliary variables (HINCP10, D_TEN, and D_HHT)*
Respondents were generated using the following response mechanisms, where x_1 = HINCP/10000, x_2 = D_TEN, and x_3 = D_HHT:

(R1) Inverse linear: $\phi_1 = (1.83 + 0.01x_1 + 0.1x_2 + 0.1x_3)^{-1}$;
(R2) Exponential: $\phi_2 = \exp(-0.57 - 0.01x_1 - 0.05x_2 - 0.05x_3)$;
(R3) Logistic type: $\phi_3 = 0.17 + 0.35\{1 + \exp(-4 + x_1/8 + 0.5x_2 + 0.5x_3)\}^{-1}$;
(R4) Quadratic: $\phi_4 = 0.57 + 0.0015x_1 + 0.002x_2 + 0.002x_3 - 0.26(x_1/40 + 0.3x_2 + 0.3x_3 - 1)^2$;

Then we calculated four estimators (besides the unadjusted estimator) using the same weighting methods used in Section 3.1: P, C1(Linear), C2 (Exponential), C3 (Logistic

type). The P-estimator is defined using the logistic regression model with three independent variables: $\text{logit}(R) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. We also used two classification tree algorithms, GUIDE (Loh, 2015) and RPART to form the weighting classes. RPART is an R-version of CART (Classification and Regression Trees, Breiman et al., 1984). For both GUIDE and RPART, the three auxiliary variables were used for classification with default options. The result is summarized in Table 11. We omit the unadjusted estimator and C3, which very similarly performed as C2.

Table 10. Simulation results with three auxiliary variables from PUMS data

Var	R	Relative Bias (RB)					Relative Root Mean Square Error (RRMSE)				
		P	C1	C2	G	T	P	C1	C2	G	T
D_HHL	1	-0.30	-0.07	-0.07	-0.01	-0.15	2.01	1.66	1.66	0.44	1.90
D_HHL	2	-0.18	-0.01	-0.01	-0.05	0.08	1.92	1.63	1.64	1.28	1.94
D_HHL	3	-0.39	-0.01	-0.02	-0.15	-0.12	1.91	1.64	1.64	1.30	1.80
D_HHL	4	-0.28	-0.03	-0.03	-0.04	-0.32	1.93	1.68	1.68	1.23	1.90
D_NOC	1	-0.39	-0.27	-0.29	-0.22	0.39	4.51	4.03	4.03	3.24	4.12
D_NOC	2	-0.22	0.07	0.04	-0.46	-0.13	4.64	4.18	4.17	2.81	4.11
D_NOC	3	-0.82	-0.05	-0.14	-0.24	-0.52	5.09	4.48	4.49	2.67	4.06
D_NOC	4	-0.23	0.07	0.03	0.33	0.30	4.77	4.13	4.13	2.77	4.06
FINCP	1	-0.26	-0.01	0.00	-3.72	-1.04	1.65	0.26	0.26	4.87	3.84
FINCP	2	-0.65	-0.03	-0.02	-8.19	-2.72	1.66	0.25	0.25	8.82	4.68
FINCP	3	-1.15	-0.08	-0.06	-10.63	-5.16	3.02	0.30	0.29	11.02	6.66
FINCP	4	-0.27	-0.08	-0.07	4.30	-1.27	1.43	0.25	0.25	5.25	2.91
GRNTP	1	-0.18	-0.25	-0.25	-2.73	-0.05	5.12	3.53	3.53	7.08	7.26
GRNTP	2	-0.16	0.04	-0.02	-1.52	-0.34	5.33	3.39	3.37	6.58	6.33
GRNTP	3	-0.42	0.45	0.28	-0.63	-0.28	6.03	3.73	3.69	6.36	6.96
GRNTP	4	-0.18	-0.78	-0.58	6.42	3.46	4.91	3.17	3.11	8.92	7.56
VALP	1	-0.12	-0.07	-0.10	-0.91	-0.51	4.53	3.94	3.94	5.36	5.19
VALP	2	-0.39	0.13	0.10	-3.19	-0.90	4.99	4.08	4.09	6.06	5.36
VALP	3	-0.51	0.10	0.06	-4.16	-1.96	5.96	4.58	4.65	6.76	5.92
VALP	4	0.32	1.02	0.91	1.14	-1.05	4.27	3.67	3.63	5.50	5.20

The classification tree algorithms tend to have a larger bias for continuous variables such as FINCP. This suggests that it may be better to use the regression tree option rather than the classification tree. This is a subject of future research.

We also ran simulation under more complex scenarios using PUMS such as that some auxiliary variables involved in the response models were omitted in the estimation model. When this happened, the nonresponse bias increased substantially. We also tried the two-step procedures under non-census sampling experiments. The two-step procedure was slightly less biased and more efficient in our simulation.

4. Discussion and Concluding Remarks

From the simulation results, we provide the following summary:

- The one-step procedure is virtually unbiased under the double protection if either A1 or A2 condition is met.
- However, the one-step procedure is more vulnerable than the two-step procedure in the absence of double protection (that is, neither A1 nor A2 condition is met).
- For natural populations and under more realistic situations, the one-step

procedure is fairly robust although it can be less efficient than the two-step procedure.
- Inclusion of important predictors in the estimation model is important and seems more important than correct model specification.
- Classification tree algorithms are viable options for weighting but more study is needed.

We did not study variance estimation. Variance estimation based on the Taylor method is discussed in Kim and Haziza (2014) and Kott and Liao (2015). It appears that the resampling based methods are not well developed and need more research.

References

Bethlehem, J.G. (2002). Weighting Nonresponse Adjustments Based on Auxiliary Information. *Survey Nonresponse*, R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont.

Flores Cervantes, I., and Brick, J.M. (2008). Empirical evaluation of raking ratio adjustments for nonresponse. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 2218-2225.

Folsom, R. E., & Singh, A. C. (2000). The Generalized Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse, and Poststratifcation. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 598-603.

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

Kim, J.K., and Haziza, D. (2014). Doubly robust inference with missing survey data. *Statistica Sinica*, 24, 375-394.

Kim, J. K., & Park, H. (2006). Imputation Using Response Probability. *Canadian Journal of Statistics*, 34, 1-12.

Kott, P. (2006). Using calibration weighting to adjust for nonresponse and undercoverage. *Survey Methodology*. 32, 133-142.

Kott, P.S., and Liao, D. (2012). Providing double protection for unit nonresponse with nonlinear calibration-weighting routine. *Survey Research Methods*, 6, 105-111.

Kott, P.S., and Liao, D. (2015). On step or two? Calibration weighting from a complete list frame with nonresponse. *Survey Methodology*, 41, 165-181.

Fuller, W. A., Loughin, M. M., & Baker, H. D. (1994). Regression Weighting for the 1987-88 National Food Consumption Survey. *Survey Methodology*, 20, 75-85.

Loh, W.-Y. (2015). GUIDE Classification and Regression Trees: User Manual for Version 19.0. University of Wisconsin – Madison.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Official Statistics*, 15, 305–327.

Oh, H.L., and Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys,* (Eds., W.G. Madow, I. Olkin and D.B. Rubin), New York: Academic Press, 2.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Chichester, UK: Wiley.

Särndal, C.-E., and Lundström, S. (2008). Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator. *Journal of Official Statistics*, 4, 251–260.