Statistics and Applications {**ISSN 2452-7395**}
Volume 16 No. 1, 2018 (New Series), pp 57-70

Adjustment for Unit Non-response in Survey Sampling: A Selective Review

Yogendra P. Chaubey
Department of Mathematics and Statistics, Concordia University, Montreal

Received: November 27, 2017; Reviewed: December 10, 2017; Accepted: December 18, 2017

Abstract

This paper provides an overview of the estimators of finite population mean in the presence of unit non-response. A special emphasis is placed on estimating the response probabilities using non-parametric methods. These estimated probabilities may then be used for the adjustment of Horvitz-Thompson estimator or other estimators considered in the literature. Nonparametric predicted values based on some auxiliary information are used to find the estimate of the finite population mean, while adjusting it by a non-parametric estimate of the response probability.

Key words: Generalized Regression Estimator; Local Polynomial Regression; Nonresponse; Superpopulation.

1. Introduction

In many sample surveys some of the units contacted do not respond to some or all the items on a questionnaire. Such non-response, is common in practice whenever the population consists of units such as individuals, households, or businesses. Non-response is treated as one of the potential sources of missing data that may occur as unit non-response, item non-response or as partial non-response where the sampled unit may not respond to all the survey items, the sampled unit fails to provide responses to a small number of items and the sampled unit fails to provide a large number of survey items, respectively. Hansen and Hurwitz (1946) brought the problem of non-response in surveys to forefront by highlighting the bias incurred by using the naive estimator of mean of the subsample of respondents. A large number of techniques to handle missing data comes under the rubric of weighting adjustments and imputation techniques (see Brick and Kalton (1996) for an extensive review in this category). Weighting adjustments inflate the weights for the responding units with an aim of obtaining (approximately) unbiased estimators whereas the goal of the imputation methods, as put forward in Brick and Kalton (1996), is to "compensate for the missing data in such a manner that the analysis file may be subjected to any form of analysis without the need for further consideration of the missing data." The subject matter treated in this article is the estimation of the population mean under unit non-response; the focus is on the estimators based on estimators of response probabilities.

In order to put things in perspective, we consider a finite population $\mathbb{U} = \{1, 2, ..., N\}$ consisting of N units, with N being known. The characteristic of interest is denoted by y and

Corresponding author: Yogendra P. Chaubey
E-mail: yogen.chaubey@concordia.ca

the corresponding population mean that we are interested in estimating is given by

$$\bar{Y} = \frac{1}{N}\sum_{i\in\mathbb{U}} y_i. \tag{1.1}$$

We will also assume the availability of an auxiliary variable x (or a set of κ auxiliary variables $x_1, ..., x_\kappa$) whose values are known for all the units of the population. We will denote by vector $\mathbf{x}, \mathbf{y}$ the values of x and y for all the units of the population.

Let $\mathbf{s}$ be a sample of fixed size n drawn from $\mathbb{U}$ according to a known sampling design $p(.)$ that is independent of the characteristic of interest, but may depend on the auxiliary characteristic, such that

$$p(\mathbf{s}) \geq 0 \text{ for all } \mathbf{s} \in \mathcal{L} \tag{1.2}$$

and

$$\sum_{\mathbf{s}\in\mathcal{L}} p(\mathbf{s}) = 1 \tag{1.3}$$

where $\mathcal{L}$ is the set of all $\mathbf{s}$ of fixed size n. The inclusion probability of unit i is defined as

$$\pi_i = \sum_{\mathbf{s}\in\mathcal{L}_i} p(\mathbf{s}) \text{ such that } \pi_i > 0 \text{ for all } i \in \mathbb{U} \tag{1.4}$$

where $\mathcal{L}_i = \{\mathbf{s} : i \in \mathbf{s}\}$.

With these notations on hand the well known Horvitz-Thompson (abbreviated as H-T) estimator of $\bar{Y}$ is given by

$$\bar{y}_\pi = \frac{1}{N}\sum_{i\in\mathbf{s}} \omega_i y_i \tag{1.5}$$

where $\omega_i = \pi_i^{-1}$. Defining the random variable I_i as the indicator of the inclusion of the i^{th} unit in the sample, $\mathbf{I} = \{I_1, ..., I_N\}$ is used to denote the vector of sample inclusion indicators.

In order to model the non-response, it is common to use a probabilistic approach, where every unit in the population is assumed to be a (potential) respondent or non-respondent. Let R_i denote the indicator variable of i^{th} unit responding in the survey, $i \in \mathbb{U}$, the sample of respondents will be denoted by $\mathbf{s_R}$, *i.e.* $\mathbf{s_R} = \{i \in \mathbf{s} : R_i = 1\}$. The number of respondents will be denoted by n_R. The distribution of the vector $(R_i : i \in \mathbf{s})$ is called the *response mechanism,* that will be denoted by $q(.)$. In this article we consider the response mechanism generated by Poisson sampling where for the random variable R_i for each $i \in \mathbb{U}$,

$$P(R_i = 1|\mathbf{I, y, x}) = P(R_i = 1|\mathbf{x}) \equiv \phi_i \equiv \phi(x_i) \tag{1.6}$$

for all $i \in \mathbb{U}$, where $\phi(.)$ is a smooth function, but otherwise unspecified.

As an alternative to the H-T estimator, it is also customary to consider the ratio estimator (see Cassel, Särndal and Wretman (1977), §15.6) that is just based on the responding units, *i.e.*

$$\bar{y}_{rat,\pi} = \frac{\sum_{i\in\mathbf{s}} \omega_i y_i}{\sum_{i\in\mathbf{s}} \omega_i}. \tag{1.7}$$

(The reference to 'Cassel, Särndal and Wretman', henceforth will be shortened as 'CSW'.) In what follows we provide a selective review of the modifications of these estimators under the probabilistic set-up of non-response. The main focus of this review article is the case of unknown ϕ_i, however, first we expose the case of known ϕ_i.

2. Adjustment for Unit Non-response with Known Response Probabilities

2.1 Nargundkar-Joshi adjustment

Assuming the knowledge of the response probabilities Nargundkar and Joshi (1975) modified the Horvitz-Thompson estimator as

$$\bar{y}_{\pi\phi} = \frac{1}{N} \sum_{i \in \mathbf{s_R}} \frac{\omega_i y_i}{\phi_i}. \tag{2.1}$$

An estimator M is said to be design unbiased or $p-$ unbiased for the population mean $\bar{Y}$ if

$$E_p(M) = \bar{Y}.$$

The Horvitz-Thompson estimator is $p-$ unbiased, however, since a non-response distribution is introduced into the estimation procedure, unbiasedness must now be defined with respect to the design p and the response mechanism q.

It can be easily shown that the Nargundkar-Joshi estimator is pq unbiased if the true distribution of the respondents is $q(.)$. In practice assumptions must be made about the responding distribution which may be mis-specified and in turn cause pq bias. Nargundkar and Joshi (1975) also derived an unbiased estimator of the variance of their estimator.

The ratio estimator given in (1.7) may be adjusted using the above strategy resulting into the ratio estimator

$$\bar{y}_{rat,\pi\phi} = \frac{\sum_{i \in \mathbf{s_R}} \omega_i \phi_i^{-1} y_i}{\sum_{i \in \mathbf{s_R}} \omega_i \phi_i^{-1}}. \tag{2.2}$$

It may be noted that the above estimator is not affected by the non-response mechanism in case $\phi_i \equiv \phi,\ \forall i$, as it conveniently cancels out from the numerator and denominator. This is the case of ignorable non-response, a case that is not a common occurrence as the non-response probability differs from unit to unit in a practical setting.

2.2 CSW - generalised regression estimator

The basic idea behind the GRE is to write the population total t_y as

$$t_y = \sum_{i \in \mathbb{U}} m_i + \sum_{i \in \mathbb{U}} e_i \tag{2.3}$$

where $e_i = y_i - m_i$, for a given set of $\{m_1, ..., m_N\}$. Now the second term can be estimated by $\hat{t}_{e\pi} = \sum_{i \in \mathbf{s_R}} \omega_i e_i$ for a sample with complete response. And it can be adjusted taking into account the non-response by considering the estimator $\hat{t}_{e\pi\phi} = \sum_{i \in \mathbf{s_R}} \omega_i e_i / \phi_i$. When m_i are obtained by predicted values from a multiple linear regression of y on a set of p regressors, we get the celebrated generalised regression estimator due to Cassel, Särndal and Wretman (1979).

Here we outline the details of this estimator by considering only one auxiliary variable x. The motivation for this approach is to consider the finite population as a realization from an infinite super population ξ, in which

$$y_i = m(x_i) + \epsilon_i, i \in \mathbb{U} \tag{2.4}$$

where ϵ_i are independent random variables with mean 0 and variance $\upsilon(x_i)$, $\mu(x)$ is a smooth function of x and $\upsilon(x)$ is smooth and strictly positive. The function $m(x)$ may be called the mean function and $\upsilon(x)$ the variance function, as given x_i, the super population model ξ ensures that

$$\mathbb{E}_\xi[y_i] = m(x_i) \text{ and } \mathbb{V}_\xi(y_i) = \upsilon(x_i). \tag{2.5}$$

The multiple linear model is the usual model that is used to define $m_i = \hat{y}_i$. (Its form is expressed later in this section). This estimator may be preferred in practice as it is shown to be design consistent and model unbiased. As a result it may provide large gains in efficiency over the usual H-T estimator when the model is correct other wise it could provide an estimator that may not lose much efficiency over the usual H-T estimator. Thus the generalized regression estimator in the presence of non-response is given by

$$\bar{y}_{GRE\pi\phi} = \frac{1}{N}\sum_{i\in\mathbb{U}} \hat{y}_i + \frac{1}{N}\sum_{i\in\mathbf{s_R}} \omega_i\phi_i^{-1}(y_i - \hat{y}_i). \tag{2.6}$$

Another variant of the above estimator, that naturally arises based on the methodology of the ratio estimator, namely

$$\bar{y}_{GRE,rat\pi\phi} = \frac{1}{N}\sum_{i\in\mathbb{U}} \hat{y}_i + \frac{\sum_{i\in\mathbf{s_R}} \omega_i\phi_i^{-1}(y_i - \hat{y}_i)}{\sum_{i\in\mathbf{s_R}} \omega_i\phi_i^{-1}}. \tag{2.7}$$

In the general case of ν predictors, the predicted values used in the so called GRE are motivated by model assisted approach (see Särndal, Swensson and Wretman (1992)), where the population $\mathbb{U}$ as a random sample from a multiple linear regression model given by

$$\mathbb{E}_\xi[y_i] = \mathbf{x}_i'\beta \text{ and } \mathbb{V}_\xi(y_i) = \upsilon(\mathbf{x}_i) = \upsilon_i, \tag{2.8}$$

where $\mathbf{x}_i$ represents the vector of explanatory variables corresponding to the unit $i \in \mathbb{U}$. The variance function $\upsilon(\mathbf{x}_i)$ is assumed to be known except for a multiplicative constant. In this set-up $\hat{y}_i$ is given by

$$\hat{y}_i = [1\ x_{1i}\ ...\ x_{\nu i}]\underline{\hat{\beta}}$$

with $\underline{\hat{\beta}}$ given by

$$\underline{\hat{\beta}} = \left(\mathbf{X}'_{\mathbf{s_R}} V^{-1}_{\mathbf{s_R}} \Pi^{-1}_{\mathbf{s_R}} \Phi^{-1}_{\mathbf{s_R}} \mathbf{X}_{\mathbf{s_R}}\right)^{-1} \mathbf{X}'_{\mathbf{s_R}} V^{-1}_{\mathbf{s_R}} \Pi^{-1}_{\mathbf{s_R}} \Phi^{-1}_{\mathbf{s_R}} \mathbf{y}_{\mathbf{s_R}}, \tag{2.9}$$

where $V_{\mathbf{s_R}}$, $\Pi_{\mathbf{s_R}}$, and $\Phi_{\mathbf{s_R}}$ are $(n_R \times n_R)$ diagonal matrices with diagonal elements respectively of υ_i, π_i and ϕ_i, $(i \in \mathbf{s_R})$, and $\mathbf{X}_{\mathbf{s_R}}$ is a $((\nu+1) \times n_R)$ matrix with first column of 1's and the next columns being populated by the values of ν predictors corresponding to the respondents, and $\mathbf{y}_{\mathbf{s_R}}$ is the vector of $y-$ observations for the sample of respondents.

2.3 Breidt-Opsomer - local polynomial regression (LPR) estimator

Breidt and Opsomer (2000) considered non-parametric prediction of $m(x)$ by local polynomial (of degree ν) kernel-regression, that is outlined below. For a symmetric kernel K and band-width h define $n \times (\nu+1)$ matrix

$$\mathbf{X}_{si} = [1\ x_j - x_i, ..., (x_j - x_i)^\nu]_{j\in\mathbf{s_R}} \tag{2.10}$$

and define $n \times n$ matrix

$$\mathbf{W}_{si} = \text{diag}\left\{ \frac{1}{h\pi_j\phi_j} K\left(\frac{x_j - x_i}{h}\right)\right\}_{j\in \mathbf{s_R}}. \tag{2.11}$$

Then the local polynomial regression estimator of $m(x_i)$ is given by

$$\hat{m}_{i\pi\phi} = \mathbf{e_1'}(\mathbf{X'_{si}W_{si}X_{si}})^{-1}\mathbf{X'_{si}W_{si}y_s}, \tag{2.12}$$

where $\mathbf{e_1}$ is a column vector of appropriate dimension with first element 1 and rest of the elements being zero.

Thus the LPR estimator under non-response is given by

$$\bar{y}_{LPR,\pi\phi} = \frac{1}{N}\sum_{i\in\mathbb{U}} \hat{m}_{i\pi\phi} + \frac{1}{N}\sum_{i\in\mathbf{s_R}} \frac{\omega_i(y_i - \hat{m}_{i\pi\phi})}{\phi_i}. \tag{2.13}$$

This may be again modified with ratio estimator in the background as

$$\bar{y}_{LPRrat,\pi\phi} = \frac{1}{N}\sum_{i\in\mathbb{U}} \hat{m}_{i\pi\phi} + \frac{\sum_{i\in\mathbf{s_R}} \omega_i\phi_i^{-1}(y_i - \hat{m}_{i\pi\phi})}{\sum_{i\in\mathbf{s_R}} \omega_i\phi_i^{-1}}. \tag{2.14}$$

2.4 Chaubey-Crisalli - generalised smoothing estimator

Chaubey and Crisalli (2002) proposed the **generalized smoothing estimator** (GSE) of $\bar{y}_N$ by replacing $\hat{y}_i$ in GRE by adapting non-parametric Nadaraya-Watson regression estimator (Nadaraya (1964), Watson (1964)) of $m(x_i)$. As is well-known the Nadaraya-Watson regression estimator is a special case of the local polynomial kernel regression with $\nu = 0$, Chaubey and Crisalli estimator may be considered as a special case of the estimator considered above. However, there is a shuttle difference between the two estimators. Noting that for the matrices involved in finding $\hat{m}_{\pi\phi}$, their elements represent estimation of various population totals, which are known for the auxiliary variable. Chaubey and Crisalli estimator uses their population values instead of their estimators. We can explicitly write the corresponding expression as

$$\tilde{m}_{\pi\phi}(x_i) = \frac{\sum_{j\in\mathbf{s_R}} \omega_j K_h\left(x_i - x_j\right)\phi_j^{-1}y_j}{\sum_{j\in\mathbf{U}} K_b\left(x_i - x_j\right)} \equiv \tilde{m}_{i\pi\phi}, \tag{2.15}$$

where $K_h(.) = h^{-1}K(./h)$. The motivation of the above estimator comes from the population based nonparametric kernel estimator of $m(x)$ given by

$$m_o\left(x\right) = \sum\nolimits_{j\in U} w_j(x)y_j \tag{2.16}$$

where

$$w_j(x) = \frac{K_h\left(x - x_j\right)}{\sum\limits_{i\in U} K_h\left(x - x_i\right)}.$$

The difference between the nonparametric smoothing based estimators proposed in Chaubey and Crisalli (2002) and Breidt and Opsomer (2002), now becomes clear. Where as Breidt and Opsomer (2002) consider the sample based estimator of the denominator in (2.15), Chaubey

and Crisalli (2000) use the population based value. The estimator parallel to (2.13) and (2.14) using the sample estimate of the numerator of the Nadaraya-Watson estimator are

$$\bar{y}_{GSE,\pi\phi} = \frac{1}{N}\sum_{i\in\mathbb{U}} \tilde{m}_{i\pi\phi} + \frac{1}{N}\sum_{i\in\mathbf{s_R}} \frac{\omega_i(y_i - \tilde{m}_{i\pi\phi})}{\phi_i}. \tag{2.17}$$

and

$$\bar{y}_{GSErat,\pi\phi} = \frac{1}{N}\sum_{i\in\mathbb{U}} \tilde{m}_{i\pi\phi} + \frac{\sum_{i\in\mathbf{s_R}} \omega_i\phi_i^{-1}(y_i - \tilde{m}_{i\pi\phi})}{\sum_{i\in\mathbf{s_R}} \omega_i\phi_i^{-1}}. \tag{2.18}$$

respectively.

Our emphasis in the present paper is centered around the non-parametric regression based predicted values in considering the so called generalized difference estimator (see Särndal, Swensson and Wretman 1992, p.22), however, other modern estimators as discussed in a recent review article by Breidt and Opsomer (2017) may also be used.

Remark 2.1: The ratio estimator $\bar{y}_{rat,\pi}$ is to be preferred over the H-T in a variety of situations, even though, it may not be design unbiased. The reader is referred to Särndal, Swensson and Wretman (1992) §5.7 for a variety of examples in favor of the ratio estimator. One may extrapolate these advantages in favor of the estimator (2.2) over (2.1) or that of the estimator (2.14) over (2.13) and that of (2.14) over (2.13).

Remark 2.2: One may be tempted to derive a similar conclusion while contrasting the nonparametric regression estimators (2.12) and (2.15) also. In this situation, however, the advantage of $\hat{m}$ over $\tilde{m}$ is not clear. As for as I am aware of, there is no definitive answer in the literature, especially in the context of non-response problem.

3. Adjustment for Unit Non-response with Unknown Response Probabilities

In practice the response probabilities are not known and they must be estimated. All the estimators presented in the previous section may be adapted by plug-in method. For example, for a given set of estimated response probabilities $\hat{\phi}_i, i \in \mathbf{s_R}$, the estimator $\bar{y}_{\pi\phi}$ becomes

$$\bar{y}_{\pi\hat{\phi}} = \frac{1}{N}\sum_{i\in\mathbf{s_R}} \frac{\omega_i y_i}{\hat{\phi}_i}. \tag{3.1}$$

Särndal and Hui (1981) reformulated Nargundkar and Joshi (1975) procedure by embedding the response probabilities in their model assisted approach. In doing this, they assumed that the individual response probabilities ϕ_k are dependent on the known vector of auxiliary variables $\mathbf{x}_i, i = 1, 2, ..., q$, i.e.,

$$\phi_k = f(\mathbf{x}_k'; \underline{\theta}) \tag{3.2}$$

where $\underline{\theta}$ is an unknown vector of coefficients that can be estimated from the available data. The function f may be thought of as coming from a *link function* through generalized linear model set up. For example, using the logistic regression

$$f(\mathbf{x}_k'; \underline{\theta}) = \frac{\exp(\mathbf{x_k'}\theta)}{1 + \exp(\mathbf{x_k'}\theta)}.$$

The unknown $\underline{\theta}$ is estimated by minimizing the likelihood function

$$L\left(\underline{\theta}\right) = \prod_{i \in \mathbf{s_R}} \phi_i \prod_{i \in \mathbf{s_R^c}} \left(1 - \phi_i\right). \tag{3.3}$$

The resulting estimated parameter $\hat{\theta}$ is used to estimate the individual response probabilities

$$\hat{\phi}_i = f(\mathbf{x}_i'; \hat{\underline{\theta}}). \tag{3.4}$$

Särndal and Hui (1981) investigated the properties of this method by means of Monte Carlo experiments. They concluded that if the regression model is representative of the population point scatter, then the estimator $\bar{y}_{GRE,\pi\hat{\phi}}$ is design unbiased even if the response probabilities are wrongly estimated using the response model. On the other hand if regression model is not representative of the population point scatter then $\hat{m}_{GRE,\pi\hat{\phi}}$ can be response unbiased if the response mechanism is correctly modeled but then the variance of $\bar{y}_{GRE,\pi\hat{\phi}}$ increases.

In order not to be over influenced by the mis-specification of the response modeling, non-parametric approach of estimation of the response probabilities have been advocated by several authors in recent years, some of which are featured next.

3.1 Nonparametric estimators of response probabilities

Giommi (1985, 1987) considered non-parametric kernel regression based estimator of response probabilities using uniform kernel and Gaussian kernels respectively. This approach was further investigated further investigated by Niyosenga (1994). However, these authors did not consider the inclusion probabilities in their nonparametric estimators. Crisalli (1999) proposed estimation of response probabilities, in his doctoral thesis, by using a general kernel K^* while incorporating the inclusion probabilities

$$\hat{\phi}_i = \frac{\sum_{j \in \mathbf{s}} \omega_j K_h^*(x_j - x_i) R_j}{\sum_{j \in \mathbf{s}} \omega_j K_h^*(x_j - x_i)}. \tag{3.5}$$

The notation K^* here indicates that the kernel chosen here may be different from that employed in estimating $m(x)$.

Da Silva and Opsomer (2006) investigated the asymptotic properties of $\bar{y}_{\pi\hat{\phi}}$ and those of $\bar{y}_{rat,\pi\hat{\phi}}$ estimator parallel to Hajék's set-up under a super-population model under non-response. Further, Da Silva and Opsomer (2009) extended the results in the above paper using local polynomial kernel regression

Remark 3.1: Da Silva and Opsomer (2006) pointed out one possible difficulty in using the kernel estimator of the response probabilities when there are no respondents for $x \in (x_i - h, x_i + h)$ in which case $\hat{\phi}_i$ will be zero and $bary_{\hat{\phi}}$ is not well defined. For this reason this estimator was modified as

$$\hat{\phi}_i = \frac{\max(\sum_{j \in \mathbf{s}} \omega_j K_h(x_j - x_i) R_j, \delta(Nn)^{-1})}{\sum_{j \in \mathbf{s}} \omega_j K_h(x_j - x_k)} \tag{3.6}$$

that is bounded away from zero. Hence forth, $\hat{\phi}_k$ will refer to these modified values.

Remark 3.2: Through a number of simulation studies, Crisalli (1999, Chap. 5) affirmed that the generalised smoothing estimator

$$\bar{y}_{GSE} = \frac{1}{N}\sum\nolimits_{i\in U} \tilde{m}_{i\pi\hat{\phi}} + \frac{1}{N}\sum\nolimits_{i\in \mathbf{s_R}} \omega_i e_i. \tag{3.7}$$

that does not incorporate the estimated response probabilities, may still be viable alternative to $\bar{y}_{\pi\hat{\phi}}$. Consequently, it affirms the dictum pronounced in Särndal and Hui (1981) that if the population values are estimated accurately, prescription of the response mechanism may not be as important (see Crisalli (1999), Chapter 5).

Under a series of technical assumptions, Da Silva and Opsomer (2006) [see their Theorem 1, Eqs. (10) and (11)]

$$\mathbb{E}_{pq}\left[\frac{1}{N}\sum_{i\in\mathbf{s}}\frac{\omega_k y_k R_k}{\hat{\phi}_k} - \frac{1}{N}\sum_{k\in\mathbb{U}} y_k\right] = O(h_n^{3/2}) + O\left(\frac{1}{nh_n}\right),\ a.s.\ \mathbb{P}_X. \tag{3.8}$$

Using this result and applying it to $y_j K_b(x_i - x_j)$ in place of y_j we claim that

$$\mathbb{E}\left[\hat{m}_{\hat{\phi}}(x_i) - \frac{\frac{1}{N}\sum_{j\in\mathbf{U}} y_j K_b^*(x_i - x_j)}{\frac{1}{N}\sum_{j\in\mathbb{U}} K_b^*(x_i - x_j)}\right] = O(h_n^{3/2}) + O\left(\frac{1}{nh_n}\right),\ a.s.\ \mathbb{P}_X. \tag{3.9}$$

Similarly the second term of $\bar{y}_{GSE_{\hat{\phi}}}$ can be roughly approximated by

$$\frac{1}{N}\sum\nolimits_{i\in\mathbf{s}} \frac{\omega_i(y_i - \hat{m}_{\hat{\phi}}(x_i))R_i}{\hat{\phi}_i} \approx \frac{1}{N}\sum\nolimits_{i\in\mathbf{s}} \frac{\omega_i(y_i - \hat{m}_o(x_i))R_i}{\hat{\phi}_i}. \tag{3.10}$$

Now, using Theorem 1 of da Silva and Opsomer (2006) again on the right hand side of the above equation we get

$$\mathbb{E}\left[\bar{y}_{GSE_{\hat{\phi}}} - \bar{y}_N\right] \to 0, \text{a.s. } \mathbb{P}_\xi. \tag{3.11}$$

This provides a rough sketch of the asymptotic unbiasedness and model consistency of the model assisted estimator $\bar{Y}_{GSE_{\hat{\phi}}}$.

In the interest of the readability of the material, we have avoided discussions of technical materials. For these and other related discussion, the reader may be interested in the recent article by Breidt and Opsomer (2017). In order to gauge the performance of the estimators discussed here we report on a simulation study in the following section. In this study, we use the predicted values y_k for executing the GSE based on the Nadaraya-Watson estimator as well as the spline regression. We have also considered a modification of the estimator of $\hat{\phi}_k$ using the kernel regression and spline regression that are summarized in the next subsection.

3.2 Binary kernel regression for non-response probabilities

Note that the kernel based non-parametric regression estimator, such as that of ϕ_k proposed in Eq. (3.5), generally assumes homoscedasticity of the response variables. However, in the present case the response variables $R_1..., R_n$ are heteroscedastic, since

$$\text{Var}\{R_j\} = \phi_j(1 - \phi_j). \tag{3.12}$$

Thus, a weighted nonparametric regression model will be better in providing more efficient estimates. Using weights

$$\upsilon_k = \frac{1}{\phi_k(1-\phi_k)},$$

will be the proper way to execute the nonparametric regression, however, since these values are unknown, we propose an iterative procedure.

Step 1. Fit the regression model by nonparametric regression

$$\hat{\phi}_k^{[t]} = \sum_{i\in\mathbf{s}} w_{ki} R_i,\ t=0$$

where

$$w_{ki} = \frac{\omega_i K_b(x_k - x_i)}{\sum_{i\in\mathbf{s}} \omega_i K_b(x_k - x_i)}.$$

Step 2. Estimate the weights υ_k using the results of Step 1,

$$\upsilon_k = \frac{1}{\hat{\phi}_k^{[t]}(1-\phi_k^{[t]})}.$$

Step 3. The estimated weights are then used to transform the variables x_k, R_k as

$$x_k^* = x_k/\sqrt{\upsilon_k}, R_k^* = R_k/\sqrt{\upsilon_k}.$$

Step 4. The new estimate of ϕ_k is now given by

$$\hat{\phi}_k^{[t+1]} = \sum_{k\in\mathbf{s}} w_{ki} R_i^*$$

with

$$w_{ki}^* = \frac{\omega_i K_b(x_k^* - x_i^*)}{\sum_{i\in\mathbf{s}} \omega_i K_b(x_k^* - x_i^*)}.$$

Step 5. Steps 2-4 are repeated until

$$\|\hat{\Phi}^{[t+1]}\| - \|\hat{\Phi}^{[t]}\| \leq \epsilon$$

for some prescribed threshold ϵ; $\hat{\Phi}^{[t]}$ denotes the vector of the estimated response probabilities at the t^{th} iteration.

3.3 Binary spline regression for non-response probabilities

Using the same setup as before we shall now describe the binary spline smoother. The goal of this procedure is to minimize the penalized residual sum of squares which is

$$\sum_{k\in\mathbf{s}} (R_k - \phi(x_k))^2 + \eta \int (\phi''(x))^2 dx \tag{3.13}$$

over all functions $\phi(.)$ with continuous first and integrable second derivatives. As before the parameter η represents the rate of exchange between the residual error and the roughness of the curve $\phi(.)$ and therefore is a smoothing parameter which has the same function as the bandwidth, in kernel regression. It was shown in Schoenberg (1964) (see also Wahba (1990)) that the unique solution for the problem is a cubic spline, that has the following properties:

a. A cubic polynomial fits the data between two successive sampled values.

b. At the sampled values x_k of x, $\hat{\phi}(x)$ and its two first derivatives are continuous.

c. At the boundary points $x_{(1)}$ and $x_{(n)}$ of x, the second derivatives of $\hat{\phi}(x)$ exists. integrable.

Therefore we estimate the response probability under this setup as

$$\phi_k = \hat{\phi}(x_k), \text{ for all } k \in . \tag{3.14}$$

The parameter η is chosen by minimizing the cross-validating the sum of squares criterion

$$CV(\eta) = \frac{1}{n}\sum_{k\in \mathbf{s}}(R_k - \hat{\phi}_{-k}(x_k))^2 \tag{3.15}$$

where $\hat{\phi}_{-k}(x_k)$ is obtained using the spline smoothing leaving out the x_k observation. We again deal with the heteroscedasticity of $R_k, k = 1, ..., n$ in the same manner as in the case of binary kernel regression.

Step 1. Fit the regression model by spline smoothing on the data $\{x_1, ..., x_n\}$. Denote the spline smoother by $\tilde{\phi}(.)$. Set

$$\tilde{\phi}_k^{[t]} = \tilde{\phi}(x_k),\ t = 0.$$

Step 2. Estimate the weights υ_k using the results of Step 1,

$$\upsilon_k = \frac{1}{\hat{\phi}_k^{[t]}(1 - \phi_k^{[t]})}.$$

Step 3. The estimated weights are then used to transform the variables x_k, R_k as

$$x_k^* = x_k/\sqrt{\upsilon_k}, R_k^* = R_k/\sqrt{\upsilon_k}.$$

Step 4. The new estimate of ϕ_k is now given by

$$\hat{\phi}_k^{[t+1]} = \hat{\phi}(x_k^*)$$

where R_k is replaced by R_k^*.

Step 5. Steps 2-4 are repeated until

$$\|\hat{\Phi}^{[t+1]}\| - \|\hat{\Phi}^{[t]}\| \leq \epsilon$$

for some prescribed threshold ϵ; $\hat{\Phi}^{[t]}$ denotes the vector of the estimated response probabilities at the t^{th} iteration.

3.4 Locally weighted likelihood estimation for response probabilities

An alternative for estimating response probabilities to the above methods may be sought in locally weighted likelihood procedure (Crisally, 1999, Chap. 7). In this approach a parametric form such as logit for the response probabilities is assumed. Thus for the logit model with one predictor variable,

$$\phi_k = \frac{\exp(\theta_0 + \theta_1 x_k)}{1 + \exp(\theta_0 + \theta_1 x_k)}, \tag{3.16}$$

θ_0 and θ_1 are estimated by maximizing the local log-likelihood

$$\ell(\theta_0, \theta_1) = \sum_{j \in \mathbf{s_R}} \ell_j(\theta_0, \theta_1) K_h^*(x_j - x_k) \tag{3.17}$$

where

$$\ell_j(\theta_0, \theta_1) = R_j \log(\frac{\phi_j}{1 - \phi_j}) + \log(1 - \phi_j) \tag{3.18}$$

is the contribution to the likelihood for the j^{th} observation. Maximizing $\ell(\theta_0, \theta_1)$ provides locally weighted likelihood estimates $\hat{\theta}_0, \hat{\theta}_1$ that are then used in finding the locally weighted likelihood estimate of ϕ_k given by

$$\hat{\phi}_k = \frac{\exp(\hat{\theta}_0 + \hat{\theta}_1 x_k)}{1 + \exp(\hat{\theta}_0 + \hat{\theta}_1 x_k)}. \tag{3.19}$$

4. Estimation of Variances of the Estimators

The variance of the estimator $\bar{y}_{\pi\phi}$ has been derived in Nargundkar and Joshi (1975) that consists of two parts, one due to the design p and the other due to the response mechanism q. This results in an unbiased estimator of $\mathbb{V}(\bar{y}_{\pi\phi})$ (see Nargundkar and Joshi (1975)). However, for our discussions this expression is not of much use, since as shown by Kim and Kim (2007), the variance of $\bar{y}_{\pi\phi}$ overestimates the variance of $\bar{y}_{\pi\hat{\phi}}$. Da Salva and Opsomer (2006, 2009) outlined the adaptation of numerical method of variance estimation of a linear estimator proposed in Fay (1991) (and used by Shao and Steel (1999) and Fuller and Kim (2005)) in the context of non-response variance estimation. They start with a linear estimator of the form in the absence of non-response

$$\hat{\theta} = \frac{1}{N} \sum_{i \in \mathbf{s}} w_i y_i, \tag{4.1}$$

and consider the replication based estimators

$$\hat{\theta}^{(\ell)} = \frac{1}{N} \sum_{i \in \mathbf{s}} w_i^{(\ell)} y_i, \ \ell = 1, ..., L, \tag{4.2}$$

where $w_i^{(\ell)}$ is replication modified weight for the $i - \text{th}$ unit based on replication ℓ. The replication based estimator of the variance of $\hat{\theta}$ is then given by

$$\mathbb{V}(\hat{\theta}) = \sum_{\ell=1}^{L} c_\ell (\hat{\theta}_L - \hat{\theta})^2. \tag{4.3}$$

For example, using the jackknife replication method (see Rao (1988), Rao, Wu and Yue (1992)) $w_i^{(\ell)} = (n-1)^{-1} I_{i \neq \ell}$ and $c_\ell = (1 - nN^{-1})(n-1)n^{-1}$. Other methods in this category are

grouped jackknife (see Rao, Wu and Yue (1992) and Shao and Wu (1989)) and balanced repeated replication (BRR) (Shao (1993), Shao and Tu (1995)). Berger and Skinner (2005) proposed a jackknife based method for variance estimation under unequal probability sampling. The above estimator is computationally cumbersome and may produce negative estimate. An alternative jackknife method has been investigated in the context of $\bar{y}_{rat,\pi}$ by Berger (2007), that could be potentially used in the non-response setup. Rust and Rao (1996) discuss a variety of replication methods for variance estimation in complex survey sampling. Lin *et al.* (2013) discuss replication variance estimation using BRR and bootstrap methods in the context of unequal probability sampling, that could also be adapted in our context.

Most of these methods have been studied in context to stratified sampling and/or with imputation and calibration techniques. There is still a lot of scope for variance estimation in the context of estimators using estimated response probabilities. In order to use these methods for estimation of variance of $\bar{y}_{GRE,\pi\hat{\phi}}$ and $\bar{y}_{GSE,\pi\hat{\phi}}$, the variation is basically attributed to errors $y_k - \hat{y}_k$, (see Chaubey and Crisalli (2002)). Hence, the replication method can be applied to $(1/N)\sum_{i\in\mathbf{s}}\omega_i\hat{\phi}^{-1}e_i$ to handle $\bar{y}_{GRE,\pi\hat{\phi}}$ and $\bar{y}_{GSE,\pi\hat{\phi}}$, where e_i represent residuals $y_i - \hat{y}_i$ from the regression model or from the nonparametric regression model. The methods described in Berger (2007) and Rust et al. (2013) will be applicable for estimating the variance of $\bar{y}_{GRErat,\pi\hat{\phi}}$ and $\bar{y}_{GSErat,\pi\hat{\phi}}$.

5. Discussion

In this paper I provide a selective review of the topic of estimation of mean under unit non-response, when the response probabilities are estimated using some parametric models or using some nonparametric methods. The estimation of variance of the resulting estimators is an important topic for further research as the plug-in estimator (the variance estimator with known response probabilities substituted with their estimates) is not consistent. In recent literature, replication variance estimators for these have been suggested but their investigation is limited to jackknife method only that too with simple random sampling. Non-parametric smoothing based estimator of mean as an alternative to model based generalised regression estimator may be strong competitor, especially in face of model mis-specification and needs further investigation for complex surveys. The modern replication techniques for variance estimation developed for design based estimator may be easily adapted to model/smoothing assisted estimators simply by replacing the observations by the prediction errors. This opens a large scope of the modern smoothing methods for finite population sampling and needs further investigation.

Acknowledgments

This research is partially supported by a the Discovery Research Grant from NSERC of Canada awarded to the author that is gratefully acknowledged.

References

Berger, Y. G. (2007). A Jackknife Variance Estimator for Unistage Stratified Samples with Unequal Probabilities. *Biometrika*, **94,** 953-964.

Berger, Y. G. and Skinner, C. J. (2005). A Jackknife Variance Estimator for Unequal Probability Sampling. *Journal of the Royal Statistical Society,* Series B, **67**, 79-89.

Breidt, F. J. and Opsomer, J. D. (2000). Local Polynomial Regression Estimators in Survey Sampling. *The Annals of Statistics,* **28,** 1026-1053.

Breidt, F. J. and Opsomer, J. D. (2017). Model-Assisted Survey Estimation with Modern Prediction Techniques. *Statistical Science*, **32**, 190-205

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling.* John Wiley and Sons, New York, USA (ISBN 0-471-02563-1).

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1979). Prediction Theory for Finite Populations when Model-Based and Design-Based Principles are Combined. *Scandinavian Journal of Statistics,* **6**, 97-106.

Chaubey, Y. P. and Crisalli, A. N. (2002). The Generalized Smoothing Estimator. *Journal of Statistical Research,* **36,** 111-129.

Crisalli, A. N. (1999). *Nonparametric Prediction in Survey Sampling and its Application to the Nonresponse Problem.* Ph. D. disseration, Department of Mathematics and Statistics, Concordia University.

Da Silva, D. N. and Opsomer, J. D. (2006). A Kernel Smoothing Method to Adjust for Unit Nonresponse in Sample Surveys. *Canadian Journal of Statistics,* **34,** 563-579.

Da Silva, D. N. and Opsomer, J. D. (2009). Nonparametric Propensity Weighting for Survey Nonresponse Through Local Polynomial Regression. *Survey Methodology*, **35**, 165-176.

Fay, R. E. (1991). A Design-Based Perspective on Missing Data Variance. In *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 429-440.

Fuller, W. A. and Kim, J.-K. (2005). Hot Deck Imputation for the Response Model. *Survey Methodology,* **31,** 139-149.

Giommi, A. (1985). On Estimation in Nonresponse Situations. *Statistica*, **1**, 57-63.

Giommi, A. (1987). Nonparametric Methods for Estimating Individual Response Probabilities. *Survey Methodology*, **13**, 127-133.

Hansen, M. H., and Hurwitz, W. N. (1946). The Problem of Non-Response in Sample Surveys. *Journal of American Statistical Association,* **41**, 517-529.

Horvitz, D. G., and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of American Statistical Association,* **47**, 663-685.

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

Lin, C. D., Lu, W. W., Rust, K. and Sitter, R. R. (2013). Replication variance estimation in unequal probability sampling without replacement: One-stage and two-stage. *The Canadian Journal of Statistics*, **41,** 696-716.

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability and Applications,* **10,** 189-190.

Nargundkar, M. S., and Joshi, G. B. (1975). Nonresponse in Sample Surveys. 40th Session of the International Statistical Institute, Warsaw, *Contributed papers,* 626-628.

Niyonsenga, T. (1994). Nonparametric Estimation of Response Probabilities in Sampling Theory. *Survey Methodology,* **20,** 177-184.

Rao, J. N. K. (1988): Variance Estimation in Sample Surveys. In *Handbook of Statistics,* Volume 6, (Eds.: P. R. Krishnaiah and C. R. Rao), Elsevier Science, Amsterdam, 427-447.

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

Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992): Some recent work on resampling methods for complex surveys. *Survey Methodology*, **18**, 209-217.

Rust, K. F. and Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, **5**, 283-310.

Särndal, C. E. and Hui, T. K. (1981). Estimation for Nonresponse Situations: To What Extent Must We Rely on Models? In *Current Topics in Survey Sampling,* Academic Press, New York,USA, 227-246.

Särndal, C. E., Swensson, B. and Wretman J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York, USA (ISBN 0-387-40620-4) .

Schoenberg, I. J. (1964). Spline Functions and the Problem of Graduation. *Proceedings of the National Academy of Sciences of the United States of America,* **52(4)**, 947-950.

Shao, J. (1993). Balanced Repeated Replication. In *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 544-549.

Shao, J. and Steel, P. (1999). Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions. *Journal of the American Statistical Association*, **94**, 254-265.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap.* Springer-Verlag, New York, USA (ISBN 978-1-4612-0795-5).

Shao, J. and Wu, C. F. J. (1989): A General Theory for Jackknife Variance Estimation. *Annals of Statistics*, **17**, 1176-1197.

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series, SIAM, Philadelphia.

Watson, G. S. (1964). Smooth Regression Analysis. *Sankhya: The Indian Journal of Statistics*, Series A, **26(4)**, 359-372.