*STATISTICS IN TRANSITION new series, March 2023*
*Vol. 24, No. 2, pp. 93–111, DOI 10.59170/stattrans-2023-022*
Received – 28.04.2020; accepted – 09.12.2022

# On representativeness, informative sampling, nonignorable nonresponse, semiparametric prediction and calibration

**Abdulhakeem Eideh**[1]

## Abstract

Informative sampling refers to a sampling design for which the sample selection probabilities depend on the values of the model outcome variable. In such cases the model holding for the sample data is different from the model holding for the population data. Similarly, nonignorable nonresponse refers to a nonresponse mechanism in which the response probability depends on the value of a missing outcome variable. For such a nonresponse mechanism the model holding for the response data is different from the model holding for the population data. In this paper, we study, within a modelling framework, the semi-parametric prediction of a finite population total by specifying the probability distribution of the response units under informative sampling and nonignorable nonresponse. This is the most general situation in surveys and other combinations of sampling informativeness and response mechanisms can be considered as special cases. Furthermore, based on the relationship between response distribution and population distribution, we introduce a new measure of the representativeness of a response set and a new test of nonignorable nonresponse and informative sampling, jointly. Finally, a calibration estimator is obtained when the sampling design is informative and the nonresponse mechanism is nonignorable.

**Key words:** calibration, representative measure, response distribution, nonignorable nonresponse, informative sampling esign.

## 1. Introduction

Informative sampling refers to sampling design for which the sample selection probabilities depend on the values of the model outcome variable (or the model outcome variable is correlated with design variables not included in working model). In such cases the model holding for the sample data (after sampling) is different from the model holding for the population data (before sampling); see Pfeffermann et al. (1998). In the same way, nonignorable nonresponse refers to nonresponse mechanism in which the response probability depends on the value of a missing outcome variable;

[1] Department of Mathematics, Al-Quds University, Abu-Dees Campus, Al-Quds, Palestine.
E-mail: msabdul@staff.alquds.edu. ORCID: https://orcid.org/0000-0002-9077-5795.

see Little (1982). For such nonresponse mechanism the model holding for the response data (after responding) is different from the model holding for the population data; Eideh (2007, 2012). From the literature review on survey sampling, it is clear that ignoring informative sampling or nonignorable nonresponse yield biased descriptive and analytics inferences about finite population parameters; see, for example, Chambers and Skinner (2003) and Eideh (2009). In recent articles, Eideh (2016, 2020) considers parametric prediction of finite population total under informative sampling design and nonignorable nonresponse. The author proved that, the failure to account informative sampling and nonignorable nonresponse in the analysis of survey data leads to biased inferences about the population of interest. In this paper, we study, within a modeling framework, the semi-parametric prediction of finite population total, by specifying the probability distribution of the observed measurements under informative sampling and nonignorable nonresponse. This is the most general situation in surveys and other combinations of sampling informativeness and response mechanisms can be considered as special cases. Furthermore, based on the relationship between response distribution and population distribution, we introduced a new measure of representativeness of a response set, called generalized R-indicator, and a new test of nonignorable nonresponse and informative sampling, jointly.

The paper is structured as follows. Section 2 is devoted to notations. In Section 3 we review the definition of sample, sample-complement, response, and nonresponse distributions, and relationships between their mathematical expectations. Section 4 describes estimation of response probabilities under nonignorable nonresponse. In Section 5 a new test of nonignorable nonresponse and informative sampling was developed. In Section 6 we discuss different ways to generalize a measure of representativeness. Section 7 is devoted to the basic idea of prediction. In Section 8 we present the methodology of the semiparametric prediction of finite population total under informative sampling and nonignorable nonresponse. Finally, Section 9 provides the conclusions.

## 2. Notations

Let $U = \{1, \dots, N\}$ denote a finite population consisting of $N$ units. Let $y$ be the study or outcome variable of interest and let $y_i$ be the value of $y$ for the $i$-th population unit. A probability sample $s$ is drawn from $U$ according to a specified sampling design. The sample size is denoted by $n$. Let $\boldsymbol{x}_i = \left(x_{i1}, \dots, x_{ip}\right)'$, $i \in U$ be the values of a vector of auxiliary variables, $x_1, \dots, x_p$, and $\boldsymbol{z} = \{z_1, \dots, z_N\}$ be the values of known design variables, used for the sample selection process not included in the model under consideration. In what follows, we consider a sampling design with selection probabilities$\pi_i = Pr(i \in s) > 0$, and sampling weight $w_i = 1/\pi_i$; $i = 1, \dots, N$.

In practice, the $\pi_i$'s may depend on the population values $(\boldsymbol{x}, y, \boldsymbol{z})$. We express this dependence by writing: $\pi_i = Pr( i \in s|\boldsymbol{x}, y, \boldsymbol{z})$ for all units $i \in U$. Denote by $\boldsymbol{I} = (I_1, \dots, I_N)'$ the $N$ by 1 sample indicator (vector) variable, such that $I_i = 1$ if unit $i \in U$ is selected to the sample and $I_i = 0$ if otherwise. Therefore, $s = \{i|i \in U, I_i = 1\}$ and its complement is $\bar{s} = c = \{i|i \in U, I_i = 0\}$. We consider the population values $y_1, \dots, y_N$ as random variables, which are independent realizations from a distribution with probability density functions (pdf) $f_p(y_i|\boldsymbol{x}_i; \theta)$, indexed by a vector of parameters $\theta$.

In addition to the effect of complex sample design, one of the major problems in the analysis of survey data is that of missing values. Denote by $R = (R_1, \dots, R_N)'$ the $N$ by 1 response indicator (vector) variable such that $R_i = 1$ if unit $i \in s$ is observed and $R_i = 0$ if otherwise. We assume that these random variables are independent of one another and of the sample selection mechanism. The response set is defined accordingly as $r = \{i \in s|R_i = 1\}$ and the nonresponse set by $\bar{r} = \{i \in s|R_i = 0\}$. We assume probability sampling so that $\pi_i = Pr( i \in s) > 0$ for all units $i \in U$. Let $\psi_i = Pr( i \in r|\boldsymbol{x}, y, \boldsymbol{z}) > 0$ and $\varphi_i = 1/\psi_i$ be the response probability and response weights for all units $i \in s$. Let $O = \{(\boldsymbol{x}_i, I_i), i \in U\}, \{\pi_i, R_i, i \in s\} \cup \{(y_i, \boldsymbol{x}_i), i \in r\}$ and $N, n$, and $m$, be the available information from the sample and response sets. Furthermore, the following notations are frequently used in the paper: let $f_p$ , $E_p(\cdot)$ ;$f_s$ ,$E_s(\cdot)$;$f_{\bar{s}}$, $E_{\bar{s}}(\cdot)$; $f_r$,$E_r(\cdot)$; and $f_{\bar{r}}$, $E_{\bar{r}}(\cdot)$denote the probability density functions and mathematical expectations of the population, sample, and sample-complement, response and nonresponse distributions, respectively.

## 3. Key equations

This section is based on Eideh (2020) and the references therein. The methodology in this paper is based on the following equations:

$$E_p(y_i|x_i) = E_r(\varphi_i w_i y_i|x_i)/E_r(\varphi_i w_i|x_i), \tag{1}$$

$$E_s(y_i|x_i) = \frac{E_r(\varphi_i y_i|x_i)}{E_r(\varphi_i|x_i)}, \tag{2}$$

$$E_{\bar{s}}(y_i|x_i) = E_r\{\varphi_i(w_i - 1)y_i|x_i\}/E_r\{\varphi_i(w_i - 1)|x_i\}, \tag{3}$$

$$E_{\bar{r}}(y_i|x_i) = E_r\{(\varphi_i - 1)y_i|x_i\}/E_r\{(\varphi_i - 1)|x_i\}, \tag{4}$$

$$f_r(y_i|x_i) = \frac{E_r(\varphi_i w_i|x_i)}{E_r(\varphi_i w_i|x_i, y_i)} f_p(y_i|x_i). \tag{5}$$

Consequently, $E_p(y_i|x_i)$, $E_s(y_i|x_i)$, $E_{\bar{s}}(y_i|x_i)$, $E_r(y_i|x_i)$ and $E_{\bar{r}}(y_i|x_i)$ can be estimated based on $\left\{x_i, y_i, \hat{\phi}_i, w_i; i \in r\right\}$. For estimation of $\varphi_i$, see Section 4.

## 4. Estimation of response probabilities under nonignorable nonresponse

Under nonignorable nonresponse, the values of $y_i$ for $i \in r$ are available, but for $i \notin r$are not available, so we cannot fit the following nonresponse model:

$$\psi_i = Pr(R_i = 1|i \in s, x_i, y_i) = \frac{exp(\gamma_0+\gamma_1 x_i+\gamma_2 y_i)}{1+exp(\gamma_0+\gamma_1 x_i+\gamma_2 y_i)} \tag{6}$$

directly using the maximum likelihood method. A recent approach of estimation $\psi_i$ under nonignorable nonresponse is discussed by Sverchkov (2008) using missing information principle. Assume$R_i$: $Bernoulli(\psi_i(x_i, y_i, \gamma))$, then

$$f(r_i|x_i, y_i) = (\psi_i(x_i, y_i, \gamma))^{r_i}(1-\psi_i(x_i, y_i, \gamma))^{1-r_i}. \tag{7}$$

The maximum likelihood estimator of $\gamma$ satisfies:

$$\frac{\partial l(\gamma)}{\partial \gamma} = \sum_{i \in r} \frac{\partial \, log(\psi_i(x_i, y_i, \gamma))}{\partial \gamma} + \sum_{i \in \bar{r}} \frac{\partial \, log(1-\psi_i(x_i, y_i, \gamma))}{\partial \gamma} = 0. \tag{8}$$

Using (4), the observed log-likelihood equation is:

$$\sum_{i \in r} \frac{\partial \, log(\psi_i(x_i, y_i, \gamma))}{\partial \gamma} + \sum_{i \in \bar{r}} \frac{\left\{E_r\left(\frac{[\varphi_i(x_i, y_i, \gamma)-1]\partial \, log(1-\psi_i(x_i, y_i, \gamma))}{\partial \gamma}\right)\Big| x_i\right\}}{E_r[(\varphi_i(x_i, y_i, \gamma)-1)|x_i]}$$

$$= \sum_{i \in r} \frac{\partial \, log(\psi_i(x_i, y_i, \gamma))}{\partial \gamma} + \sum_{i \in \bar{r}} \frac{\int \frac{[\varphi_i(x_i, y_i, \gamma)-1]\partial \, log(1-\psi_i(x_i, y_i, \gamma))}{\partial \gamma} f_r(y_i|x_i) dy_i}{\int (\varphi_i(x_i, y_i, \gamma)-1) f_r(y_i|x_i) dy_i} = 0. \tag{9}$$

Hence, $\hat{\psi}_i = \psi_i(\hat{\gamma}) = \psi_i(x_i, y_i, \hat{\gamma}) = Pr(i \in r|x_i, y_i, \hat{\gamma})$ and then $\hat{\varphi}_i = 1/\hat{\psi}_i$. From now on, to simplify notation, we will use $\psi_i$ to denote $\psi_i$or $\hat{\psi}_i$.

## 5. New test of nonignorable nonresponse and informative sampling, jointly

According to (5), the response distribution $f_r(y_i|x_i)$ of $y_i, i \in r$, is different from the population distribution, $f_p(y_i|x_i)$, unless $E_r(\varphi_i w_i|x_i, y_i) = E_r(\varphi_i w_i|x_i)$ for all units $i \in U$, that is when the sampling design is noninformative and nonresponse mechanism is ignorable. In such cases, the model holding for the response data (after sampling) is the same as the model holding for the population data (before sampling). The main target of inference is estimation$E_p(y_i|\boldsymbol{x}_i)$. According to Eideh (2020), we have:

$$E_r(y_i|\boldsymbol{x}_i) = E_p\left\{\frac{E_s(\psi_i|x_i,y_i,\gamma)}{E_s(\psi_i|x_i,\theta,\eta,\gamma)}\frac{E_p(\pi_i|x_i,y_i,\gamma)}{E_p(\pi_i|x_i,\theta,\gamma)}y_i\middle|\boldsymbol{x}_i\right\} \neq E_p(y_i|\boldsymbol{x}_i). \tag{10}$$

This relationship illustrates that the failure to account nonignorable nonresponse and informative sampling design can bias the inference. So, testing the ignorable of nonresponse and the informativeness of sampling design is necessary, which is the aim of this section.

Recall that if $E_r(\varphi_i w_i|x_i, y_i) = E_r(\varphi_i w_i|x_i)$, for all units $i \in U$, then$f_r(y_i|x_i) = f_p(y_i|x_i)$. Consequently, in the spirit of equation (5), we introduce the following new test of ignorable nonresponse and informative sampling, jointly, by testing

$$H_0: E_r(\varphi_i w_i|x_i, y_i) = E_r(\varphi_i w_i|x_i) \text{ versus } H_1: E_r(\varphi_i w_i|x_i, y_i) \neq E_r(\varphi_i w_i|x_i) \tag{11}$$

at $\alpha$ level of significance.

In addition to that, testing of noninformative sampling design and nonresponse mechanism is missing completely at random, and can be approached by testing:

$$H_0: E_r(\varphi_i w_i|x_i, y_i) = \text{constant versus } H_1: E_r(\varphi_i w_i|x_i, y_i) \neq \text{constant} \tag{12}$$

at $\alpha$ level of significance.

**Particular cases:**

(a) If the sampling design is noninformative, that is, the sample selection process can be ignored, then the test of nonignorable nonresponse is determined by testing:

$$H_0: E_r(\varphi_i|x_i, y_i) = E_r(\varphi_i|x_i) \text{ versus } H_1: E_r(\varphi_i|x_i, y_i) \neq E_r(\varphi_i|x_i) \tag{13}$$

at $\alpha$ level of significance.

(b) If nonresponse mechanism is ignorable, then the test of informativeness can be conducted by testing:

$$H_0: E_r(w_i|x_i, y_i) = E_r(w_i|x_i) \text{ versus } H_1: E_r(w_i|x_i, y_i) \neq E_r(w_i|x_i) \tag{14}$$

at $\alpha$ level of significance.

The above hypotheses can be tested by using the general regression test approach, by specifying the full model. For example, in (11) and (12), assume the full model is given by:

$$E_r(\varphi_i w_i|x_i, y_i) = \beta_0 + \beta_1 x_i + \beta_2 y_i. \tag{15}$$

Then, (11) becomes$H_0$: $\beta_2 = 0$, and (12) says $H_0$: $\beta_1 = \beta_2 = 0$.

## 6. Generalized Measure of Representativeness

Schouten et al. (2009) proposed an indicator which we call an R-indicator ('R' for representativeness), for the similarity between the response to a survey and the sample or the population under investigation. This similarity can be referred to as "representative response". The R-indicator that they proposed employs estimated response probabilities.

**Definition 1 (strong):** A response subset is representative with respect to the sample if the response propensities $\rho_i$ are the same for all units in the population. That is,

$$\rho_i = Pr(R_i = 1|I_i = 1) = \rho \text{ for all units } i \in U \tag{16}$$

and if the response of a unit is independent of the response of all other units.

Under the assumption that the individual response propensities $\rho_i$ are known, Schouten et al. (2009) defined the R-indicator as:

$$R(\rho) = 1 - 2S(\rho), \tag{17}$$

where

$$S(\rho) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \bar{\rho})^2}, \bar{\rho} = \frac{1}{N}\sum_{i=1}^{N}\rho_i. \tag{18}$$

One may view R as a lack of the association measure. When $R(\rho) = 1$ there is no relation between any survey item and the missing-data mechanism. The R-indicator takes values on the interval [0, 1] with the value 1 being strong representativeness and the value 0 being the maximum deviation from strong representativeness.

Schouten et al. (2009) pointed that "in practice these propensities are unknown. Furthermore, in a survey, we only have information about the response behaviour of sample units. We, therefore, have to find alternatives to the indicators R. Let $\hat{\rho}_i$ denote an estimator for $\rho_i$ which uses all or a subset of the available auxiliary variables.

Methods that support such estimation are, for instance, logistic or probit regression models". The authors replace R by the estimators$\hat{R}$:

$$\hat{R}(\rho) = 1 - 2\hat{S}(\rho), \hat{S}(\rho) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{I_i}{\pi_i}(\hat{\rho}_i - \hat{\bar{\rho}})^2}, \hat{\bar{\rho}} = \frac{1}{N}\sum_{i=1}^{N}\frac{I_i}{\pi_i}\rho_i. \quad (19)$$

The R-indicator introduced by Schouten et al. (2009) assumed that the sampling design is noninformative and nonresponse mechanism is ignorable. In this section we develop a new indicator of representative response of the survey and the population when the sampling design is informative and the nonresponse mechanism nonignorable. It should be pointed here that, under nonignorable nonresponse, we cannot compute the propensity scores for all units in the sample, see Section 4, consequently, the formulas defined in equation (19) are not applicable in such cases.

For simplicity, assume no auxiliary variables are available. In the essence of equation (5), let us consider the following four cases.

**Case 1:** Sampling design is informative and nonresponse mechanism is nonignorable (in), then

$$f_r(y_i) = \frac{E_r(\varphi_i w_i)}{E_r(\varphi_i w_i | y_i)} f_p(y_i). \quad (20)$$

In this case, the response distribution represents the population distribution if $E_r(\phi_i w_i) = E_r(\varphi_i w_i | y_i)$. That is, $E_r(\varphi_i w_i | y_i) = \text{constant}$. In this case, we introduce the following definition.

**Definition 2:** A response set is representative with respect to the population if the product of the response weights and sampling weights $\varphi_i w_i = d_i$ are the same for all units in the population and if the response of a unit is independent of the response of all other units. That is, $\varphi_i w_i = d_i = d$ for all units $i \in U$.

Thus, we define a generalized R-indicator as follows:

$$\hat{R}(d) = 1 - 2\hat{S}(d), \quad (21)$$

where

$$\hat{S}(d) = \sqrt{\frac{1}{\sum_{i=1}^{m}\hat{d}_i}\sum_{i=1}^{m}\hat{d}_i\left(\hat{d}_i - \hat{\bar{d}}\right)^2} \quad (22)$$

and

$$\hat{\bar{d}} = \frac{1}{\sum_{i=1}^{m}\hat{d}_i}\sum_{i=1}^{m}\hat{d}_i(\hat{d}_i), \quad \hat{\varphi}_i w_i = \hat{d}_i. \quad (23)$$

**Case 2:** Sampling design is noninformative and nonresponse mechanism is nonignorable (nn), then

$$f_r(y_i) = \frac{E_r(\varphi_i)}{E_r(\varphi_i|y_i)} f_p(y_i). \quad (24)$$

In this case, the response distribution represents the population distribution if $E_r(\phi_i) = E_r(\varphi_i|y_i)$. That is, $E_r(\varphi_i|y_i)$ = constant. In this constant, we introduce the following definition.

**Definition 3:** A response set is representative with respect to the population if the response weights $\varphi_i$ are the same for all units in the population and if the response of a unit is independent of the response of all other units. That is, $\varphi_i = \varphi$ for all units $i \in U$.

Thus, we define a generalized R-indicator as follows:

$$\hat{R}_{nn}(\varphi) = 1 - 2\hat{S}_{nn}(\varphi), \quad (25)$$

where

$$\hat{S}_{nn}(\varphi) = \sqrt{\frac{1}{\sum_{i=1}^{m} \hat{\varphi}_i} \sum_{i=1}^{m} \hat{\varphi}_i (\hat{\varphi}_i - \hat{\hat{\varphi}}_{nn})^2} \quad (26)$$

and

$$\hat{\hat{\varphi}}_{nn} = \frac{1}{\sum_{i=1}^{m} \hat{\varphi}_i} \sum_{i=1}^{m} \hat{\varphi}_i^2. \quad (27)$$

**Case 3:** Sampling design is informative and nonresponse mechanism is ignorable (ii), then

$$f_r(y_i) = \frac{E_r(w_i)}{E_r(w_i|y_i)} f_p(y_i). \quad (28)$$

In this case, the response distribution represents the population distribution if $E_r(w_i) = E_r(w_i|y_i)$. That is, $E_r(w_i|y_i)$ = constant. In this case, we introduce the following definition.

**Definition 4:** A response set is representative with respect to the population if the sampling weighs $w_i$ are the same for all units in the population and if the response of a unit is independent of the response of all other units. That is, $w_i = w$ for all units $i \in U$.

Thus, we define a generalized R-indicator as follows:

$$\hat{R}_{ii}(w) = 1 - 2\hat{S}_{ii}(w), \quad (29)$$

where

$$\hat{S}_{ii}(w) = \sqrt{\frac{1}{\sum_{i=1}^{m} w_i} \sum_{i=1}^{m} w_i (w_i - \hat{\hat{w}})^2} \quad (30)$$

and

$$\widehat{w} = \frac{1}{\sum_{i=1}^{m} w_i} \sum_{i=1}^{m} w_i^2. \quad (31)$$

**Case 4:** Sampling design is noninformative and nonresponse mechanism is ignorable (ii), then

$$f_r(y_i) = f_p(y_i). \quad (32)$$

In this case, the response distribution represents the population distribution, and no need a measure of a representative subset.

Research in this highly interested generalized R-indicator is in progress.

## 7. Prediction of Finite Population Total

This section is devoted to the basics of the prediction of finite population total, taking into account informative sampling design and nonignorable nonresponse mechanism. Assume single-stage population model. Let

$$T = \sum_{i=1}^{N} y_i = \sum_{i \in s} y_i + \sum_{i \in \bar{s}} y_i = \sum_{i \in r} y_i + \sum_{i \in \bar{r}} y_i + \sum_{i \in \bar{s}} y_i \quad (33)$$

be the finite population total that we want to predict using the data from the response set and possibly values of auxiliary variables. Let $\hat{T} = \hat{T}(O)$ define the predictor of $T$ based on the available information, from the sample and response set $O = \{(I_i), i \in U\}, \{\pi_i, R_i, i \in s\} \cup \{(y_i), i \in r\}$and $N, n,$ and $m$. The mean square error (MSE) of $\hat{T}$ given $O$ with respect to the population pdf is defined by:

$$MSE_p(\hat{T}) = E_p\left\{\left(\hat{T} - T\right)^2 | O\right\} = \left\{\hat{T} - E_p(T|O)\right\}^2 + Var_p(T|O). \quad (34)$$

It obvious that (16) is minimized when$\hat{T} = E(T|O)$. Hence, the minimum mean squared error best linear unbiased predictor (BLUP) of $T = \sum_{i=1}^{N} y_i$ is given by:

$$T^* = E_p(T|O) = \sum_{i \in r} y_i + \sum_{i \in \bar{r}} E_{\bar{r}}(y_i|O) + \sum_{i \in \bar{s}} E_{\bar{s}}(y_i|O). \quad (35)$$

For more information about the parametric prediction of finite population total under informative sampling and nonignorable nonresponse; see Eideh (2020).

In the next section we consider the semiparametric prediction of finite population total under informative sampling.

## 8. Semiparametric Prediction of Finite Population Total under Informative Sampling and Nonignorable Nonresponse

Sverchkov and Pfeffermann (2004) studied the semiparametric prediction of finite population totals under informative sampling. In this section we develop the semi-parametric prediction of finite population total under informative sampling and nonignorable nonresponse. According to (35), the prediction of finite population total requires predication of $\sum_{i\in\bar{r}} y_i$, and $\sum_{i\in\bar{s}} y_i$. That is, to predict $T$we need to predict values for$\{y_i, i \in \bar{r}\}$ and $\{y_i, i \in \bar{s}\}$, based on the prediction of nonsampled and nonresponse models. Fuller (2009, p. 282) pointed that "The analysis of data with unplanned nonresponse requires the specification of a model for the nonresponse. Models for nonresponse address two characteristics: the probability of obtaining a response and the distribution of the characteristic. In one model it is assumed that the probability of response can be expressed as a function of auxiliary data. The assumption of a second important model is that the expected value of the unobserved variable is related to observable auxiliary data. In some situations models constructed under the two models lead to the same estimator. Similarly, specifications containing models for both components can be developed."

### 8.1. General Theory

Assume that

(a) the sample-complement model (or nonsampled model or imputation model for non-sampled units) takes the form:

$$y_i = S_{\boldsymbol{\beta}}(\boldsymbol{x}_i) + \varepsilon_i, \text{ for all } i \in \bar{s}, \tag{36}$$

$E_{\bar{s}}(\varepsilon_i|\boldsymbol{x}_i) = 0$ , $E_{\bar{s}}(\varepsilon_i^2|\boldsymbol{x}_i) = Var_{\bar{s}}(\varepsilon_i|\boldsymbol{x}_i) = \sigma_\varepsilon^2 v(\boldsymbol{x}_i)$ , and $E_{\bar{s}}(\varepsilon_j\varepsilon_k|\boldsymbol{x}_i) = Cov(\varepsilon_j, \varepsilon_k|\boldsymbol{x}_i) = 0, j \neq k$.

(b) and the response-complement model (or nonresponse model or missing data model or imputation model for nonrespondent units) is:

$$y_i = Z_{\boldsymbol{\alpha}}(\boldsymbol{x}_i) + \tau_i, \text{ for all } i \in \bar{r}, \tag{37}$$

$E_{\bar{r}}(\tau_i|\boldsymbol{x}_i) = 0$, $E_{\bar{r}}(\tau_i^2|\boldsymbol{x}_i) = Var_{\bar{r}}(\tau_i|\boldsymbol{x}_i) = \sigma_\tau^2 u(\boldsymbol{x}_i)$, and $E_{\bar{r}}(\tau_j\tau_k|\boldsymbol{x}_i) = Cov(\tau_j, \tau_k|\boldsymbol{x}_i) = 0,\ j \neq k$, where $S_{\boldsymbol{\beta}}(\boldsymbol{x}_i)$ and $Z_{\boldsymbol{\alpha}}(\boldsymbol{x}_i)$ are known functions of $\boldsymbol{x}_i$ that depend on unknown vector parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ , respectively. The variances $\sigma_\varepsilon^2 v(\boldsymbol{x}_i)$ and $\sigma_\tau^2 u(\boldsymbol{x}_i)$ are assumed known except for $\sigma_\varepsilon^2$ and $\sigma_\tau^2$.

For the prediction process, we need the estimation of $S_{\boldsymbol{\beta}}(\boldsymbol{x}_i)$ and $Z_{\boldsymbol{\alpha}}(\boldsymbol{x}_i)$.

**(i) Estimation of $S_{\boldsymbol{\beta}}(\boldsymbol{x}_i)$:**

**Method 1:** using (3), we have

$$S_{\boldsymbol{\beta}}(\boldsymbol{x}_i) = arg \min_{S_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{x}_i)} E_{\bar{s}} \left\{ \frac{\left(y_i - S_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{x}_i)\right)^2}{v(\boldsymbol{x}_i)} \middle| \boldsymbol{x}_i \right\} = arg \min_{S_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{x}_i)} E_r \left\{ c_i \frac{\left(y_i - S_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{x}_i)\right)^2}{v(\boldsymbol{x}_i)} \middle| \boldsymbol{x}_i \right\}, \quad (38)$$

where $c_i = \{\varphi_i(w_i - 1)/E_r(\varphi_i(w_i - 1)|\boldsymbol{x}_i)\}$.

Hence, the vector $\boldsymbol{\beta}$ can be estimated by:

$$\widehat{\boldsymbol{\beta}}_1 = arg \min_{\widetilde{\beta}} \sum_{i \in r} \left( \hat{c}_i \frac{\left(y_i - S_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{x}_i)\right)^2}{v(\boldsymbol{x}_i)} \right), \quad (39)$$

where $\hat{c}_i = \left\{\hat{\varphi}_i(w_i - 1)/\hat{E}_r(\hat{\varphi}_i(w_i - 1)|\boldsymbol{x}_i)\right\}$.

**Method 2:** using (3) and (36), and assume that $E_r(\varphi_i(w_i - 1)|\boldsymbol{x}_i) \approx E_r\left(\varphi_i(w_i - 1)\right)$, we have:

$$E_{\bar{s}} \left\{ \frac{\left(y_i - S_{\boldsymbol{\beta}}(\boldsymbol{x}_i)\right)^2}{v(\boldsymbol{x}_i)} \middle| \boldsymbol{x}_i \right\} = E_r \left\{ \frac{\varphi_i(w_i - 1)}{E_r\left(\varphi_i(w_i - 1)\right)} \frac{\left(y_i - S_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{x}_i)\right)^2}{v(\boldsymbol{x}_i)} \right\}. \quad (40)$$

Hence,

$$\widehat{\boldsymbol{\beta}}_2 = arg \min_{\widetilde{\beta}} \sum_{i \in r} \left( \hat{\varphi}_i(w_i - 1) \frac{\left(y_i - S_{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{x}_i)\right)^2}{v(\boldsymbol{x}_i)} \right), \quad (41)$$

since $E_r\left(\varphi_i(w_i - 1)\right)$ is constant.

**(ii) Estimation of $Z_{\boldsymbol{\alpha}}(\boldsymbol{x}_i)$**

**Method 1:** using (4), we have:

$$Z_{\boldsymbol{\alpha}}(\boldsymbol{x}_i) = arg \min_{Z_{\widetilde{\boldsymbol{\alpha}}}(\boldsymbol{x}_i)} E_r \left\{ k_i \frac{\left(y_i - Z_{\widetilde{\boldsymbol{\alpha}}}(\boldsymbol{x}_i)\right)^2}{u(\boldsymbol{x}_i)} \middle| \boldsymbol{x}_i \right\}, \quad (42)$$

where $k_i = \left\{(\varphi_i - 1)/E_r\left((\varphi_i - 1)|\boldsymbol{x}_i\right)\right\}$.

Hence, the vector $\widehat{\boldsymbol{\alpha}}$ can be estimated by

$$\widehat{\boldsymbol{\alpha}}_1 = arg \min_{\widetilde{\beta}} \sum_{i \in r} \left( \hat{k}_i \frac{\left(y_i - Z_{\widetilde{\boldsymbol{\alpha}}}(\boldsymbol{x}_i)\right)^2}{u(\boldsymbol{x}_i)} \right), \quad (43)$$

where $\hat{k}_i = \left\{(\hat{\varphi}_i - 1)/\hat{E}_r\left((\hat{\varphi}_i - 1)|\boldsymbol{x}_i\right)\right\}$.

**Method 2:** using (3) and (37), and assume that $E_r\big((\varphi_i - 1)|\boldsymbol{x}_i\big) \approx E_r(\varphi_i - 1)$, we have:

$$E_{\bar{r}}\left\{\frac{\left(y_i - Z_{\alpha}(x_i)\right)^2}{u(x_i)}\middle| \boldsymbol{x}_i\right\} = E_r\left\{\frac{(\varphi_i - 1)}{E_r(\varphi_i - 1)}\frac{\left(y_i - Z_{\alpha}(x_i)\right)^2}{u(x_i)}\right\}. \tag{44}$$

Thus,

$$\widehat{\boldsymbol{\alpha}}_2 = arg\min_{\widetilde{\alpha}} \sum_{i\in r}\left((\widehat{\varphi}_i - 1)\frac{\left(y_i - Z_{\widetilde{\alpha}}(x_i)\right)^2}{u(x_i)}\right). \tag{45}$$

Hence,

$$\widehat{T}_{in,1} = \widehat{E}_p(T|O) = \sum_{i\in r} y_i + \sum_{i\in\bar{r}} Z_{\widehat{\boldsymbol{\alpha}}_1}(\boldsymbol{x}_i) + \sum_{i\in\bar{s}} S_{\widehat{\boldsymbol{\beta}}_1}(\boldsymbol{x}_i) \tag{46}$$

and

$$\widehat{T}_{in,2} = \widehat{E}_p(T|O) = \sum_{i\in r} y_i + \sum_{i\in\bar{r}} Z_{\widehat{\boldsymbol{\alpha}}_2}(\boldsymbol{x}_i) + \sum_{i\in\bar{s}} S_{\widehat{\boldsymbol{\beta}}_2}(\boldsymbol{x}_i). \tag{47}$$

The benefits of using the predictor $\widehat{T}_{in,2}$ over using the predictor $\widehat{T}_{in,1}$ is that $\widehat{T}_{in,2}$ does not require the identification and estimation of $\varphi(\boldsymbol{x}_i) = E_r\big((\varphi_i - 1)|\boldsymbol{x}_i\big)$. On the other hand, in situations where this expectation can be estimated properly, the predictor $\widehat{T}_{in,1}$is likely to be more accurate since the weights $k_i = \big\{(\varphi_i - 1)/E_r\big((\varphi_i - 1)|\boldsymbol{x}_i\big)\big\}$ will often be less variable than the weights $(\varphi_i - 1)$. This is because the weights $k_i = \big\{(\varphi_i - 1)/E_r\big((\varphi_i - 1)|\boldsymbol{x}_i\big)\big\}$ only account for the net effect of the response process on the target conditional distribution $f_{\bar{r}}(y_i|\boldsymbol{x}_i, \theta, \eta, \gamma)$whereas the weights $(\varphi_i - 1)$account for the effect of the response process on the joint distribution $f_{\bar{r}}(y_i, \boldsymbol{x}_i; \theta, \eta, \gamma)$.

### 8.1.1. Particular cases

For illustration we use method 2 only under different famous models in survey sampling.

**Case 1: Common Mean Model:**

**Sample-complement model:** $E_{\bar{s}}(y_i) = \mu_{\bar{s}}$ and $Var_{\bar{s}}(y_i) = \sigma_{\varepsilon}^2$.

**Response-complement model:** $E_{\bar{r}}(y_i|x_i) = \mu_{\bar{r}}$ and $Var_{\bar{r}}(y_i) = \sigma_{\tau}^2$.

After some algebra, the weights under the 4 different combinations of sampling design (informative (i), noninformative (n)) and nonresponse mechanism (ignorable (i), nonignorable (n)) are summarized in Table 1.

**Table 1:** $w_{ir}$ **- Homogeneous Model ,** $\hat{T}_{in,2} = \sum_{i \in r} w_{ir} y_i$

| SD-NM | $w_{ir}$ |
|---|---|
| ii | $1 + \frac{(n-m)}{m} + (N-n)\frac{w_i - 1}{\sum_{i \in r}(w_i - 1)}$ |
| in | $1 + (n-m)\frac{(\hat{\varphi}_i - 1)}{\sum_{i \in r}(\hat{\varphi}_i - 1)} + (N-n)\frac{\hat{\varphi}_i(w_i - 1)}{\sum_{i \in r}\hat{\varphi}_i(w_i - 1)}$ |
| ni | $\frac{N}{m}$ |
| nn | $1 + (n-m)\frac{(\hat{\varphi}_i - 1)}{\sum_{i \in r}(\hat{\varphi}_i - 1)} + (N-n)\frac{\hat{\varphi}_i}{\sum_{i \in r}\hat{\varphi}_i}$ |

Note that $\sum_{i \in r} w_{ir} = \sum_{i \in U} 1 = N$.

**Case 2: Simple linear regression model:**
**Sample-complement model:** $E_{\bar{s}}(y_i|x_i) = \beta_0 + \beta_1 x_i$ and $Var_{\bar{s}}(y_i) = \sigma_\varepsilon^2$.
**Response-complement model:** $E_{\bar{r}}(y_i|x_i) = \alpha_0 + \alpha_1 x_i$ and $Var_{\bar{r}}(y_i) = \sigma_\tau^2$.

After some algebra, the weights under the 4 different combinations of sampling designs (informative (i), noninformative (n)) and nonresponse mechanism (ignorable (i), nonignorable (n)) are summarized in Table 2.

**Table 2:** $w_{ir}$ **- Simple linear regression model,** $\hat{T}_{in,2} = \sum_{i \in r} w_{ir} y_i$

| SD-NM | $w_{ir}$ | $\bar{x}_{\varphi^*}$ | $\bar{x}_{w^*}$ |
|---|---|---|---|
| ii | $1 + \frac{(n-m)}{m} + (n-m)(\bar{x}_{\bar{r}} - \bar{x}_{\varphi^*})\frac{(x_i - \bar{x}_{\varphi^*})}{\sum_{i \in r}(x_i - \bar{x}_{\varphi^*})^2}$ <br> $(N-n)\frac{(w_i - 1)}{\sum_{i \in r} i(w_i - 1)} +$ <br> $(N-n)(\bar{x}_{\bar{s}} - \bar{x}_{w^*})\frac{(w_i - 1)(x_i - \bar{x}_{w^*})}{\sum_{i \in r}(w_i - 1)(x_i - \bar{x}_{w^*})^2}$ | $\frac{\sum_{i \in r} x_i}{m}$ | $\frac{\sum_{i \in r}(w_i - 1)x_i}{\sum_{i \in r}(w_i - 1)}$ |
| in | $1 + (n-m)\frac{(\hat{\varphi}_i - 1)}{\sum_{i \in r}(\hat{\varphi}_i - 1)} +$ <br> $(n-m)(\bar{x}_{\bar{r}} - \bar{x}_{\varphi^*})\frac{(\hat{\varphi}_i - 1)(x_i - \bar{x}_{\varphi^*})}{\sum_{i \in r}(\hat{\varphi}_i - 1)(x_i - \bar{x}_{\varphi^*})^2}$ <br> $(N-n)\frac{\hat{\varphi}_i(w_i - 1)}{\sum_{i \in r}\hat{\varphi}_i(w_i - 1)} +$ <br> $(N-n)(\bar{x}_{\bar{s}} - \bar{x}_{w^*})\frac{\hat{\varphi}_i(w_i - 1)(x_i - \bar{x}_{w^*})}{\sum_{i \in r}\hat{\varphi}_i(w_i - 1)(x_i - \bar{x}_{w^*})^2}$ | $\frac{\sum_{i \in r}(\hat{\varphi}_i - 1)x_i}{\sum_{i \in r}(\hat{\varphi}_i - 1)}$ | $\frac{\sum_{i \in r}\hat{\varphi}_i(w_i - 1)x_i}{\sum_{i \in r}\hat{\varphi}_i(w_i - 1)}$ |
| ni | $1 + \frac{(n-m)}{m} + (n-m)(\bar{x}_{\bar{r}} - \bar{x}_{\varphi^*})\frac{(x_i - \bar{x}_{\varphi^*})}{\sum_{i \in r}(x_i - \bar{x}_{\varphi^*})^2}$ <br> $\frac{(N-n)}{m} + (N-n)(\bar{x}_{\bar{s}} - \bar{x}_{w^*})\frac{(x_i - \bar{x}_{w^*})}{\sum_{i \in r}(x_i - \bar{x}_{w^*})^2}$ | $\frac{\sum_{i \in r} x_i}{m}$ | $\frac{\sum_{i \in r} x_i}{m}$ |

**Table 2:** $w_{ir}$ **- Simple linear regression model,** $\hat{T}_{in,2} = \sum_{i\in r} w_{ir} y_i$ (cont.)

| SD-NM | $w_{ir}$ | $\bar{x}_{\varphi^*}$ | $\bar{x}_{w^*}$ |
|---|---|---|---|
| **nn** | $1+(n-m)\frac{(\hat{\varphi}_i-1)}{\sum_{i\in r}(\hat{\varphi}_i-1)}+(n-m)(\bar{x}_{\bar{r}}-\bar{x}_{\varphi^*})\frac{(\hat{\varphi}_i-1)(x_i-\bar{x}_{\varphi^*})}{\sum_{i\in r}(\hat{\varphi}_i-1)(x_i-\bar{x}_{\varphi^*})^2}$ $(N-n)\frac{\hat{\varphi}_i}{\sum_{i\in r}\hat{\varphi}_i}+(N-n)(\bar{x}_{\bar{s}}-\bar{x}_{w^*})\frac{\hat{\varphi}_i(x_i-\bar{x}_{w^*})}{\sum_{i\in r}\hat{\varphi}_i(x_i-\bar{x}_{w^*})^2}$ | $\frac{\sum_{i\in r}(\hat{\varphi}_i-1)x_i}{\sum_{i\in r}(\hat{\varphi}_i-1)}$ | $\frac{\sum_{i\in r}\hat{\varphi}_i x_i}{\sum_{i\in r}\hat{\varphi}_i}$ |

Note that $\sum_{i\in r} w_{ir} x_i = \sum_{i\in U} x_i$.

**Case 3: Simple ratio model:**

**Sample-complement model:** $E_{\bar{s}}(y_i|x_i) = \beta x_i$ and $Var_{\bar{s}}(y_i) = \sigma_\varepsilon^2 x_i$.

**Response-complement model:** $E_{\bar{r}}(y_i|x_i) = \alpha\, x_i$; and $Var_{\bar{r}}(y_i) = \sigma 2_\tau\, x_i$.

After some algebra, the weights under the 4 different combinations of sampling designs (informative (i), noninformative (n)) and nonresponse mechanism (ignorable (i), nonignorable (n)) are summarized in Table 3.

**Table 3:** $w_{ir}$ **- Simple ratio (or proportional) model,** $\hat{T}_{in,2} = \sum_{i\in r} w_{ir} y_i$

| SD-NM | $w_{ir}$ | $\bar{x}_{\varphi^*}$ | $\bar{x}_{w^*}$ |
|---|---|---|---|
| **ii** | $1+\frac{(n-m}{m})\frac{\bar{x}_{\bar{r}}}{\bar{x}_{\varphi^*}}+(N-n)\frac{(w_i-1)}{\sum_{i\in r}(w_i-1)}\frac{\bar{x}_{\bar{s}}}{\bar{x}_{w^*}}$ | $\frac{\sum_{i\in r} x_i}{m}$ | $\frac{\sum_{i\in r}(w_i-1)x_i}{\sum_{i\in r}(w_i-1)}$ |
| **in** | $1+(n-m)\frac{(\hat{\varphi}_i-1)}{\sum_{i\in r}(\hat{\varphi}_i-1)}\frac{\bar{x}_{\bar{r}}}{\bar{x}_{\varphi^*}}+(N-n)\frac{\hat{\varphi}_i(w_i-1)}{\sum_{i\in r}\hat{\varphi}_i(w_i-1)}\frac{\bar{x}_{\bar{s}}}{\bar{x}_{w^*}}$ | $\frac{\sum_{i\in r}(\hat{\varphi}_i-1)x_i}{\sum_{i\in r}(\hat{\varphi}_i-1)}$ | $\bar{x}_{w^*} = \frac{\sum_{i\in r}\hat{\varphi}_i(w_i-1)x_i}{\sum_{i\in r}\hat{\varphi}_i(w_i-1)}$ |
| **ni** | $1+\frac{(n-m)}{m}\frac{\bar{x}_{\bar{r}}}{\bar{x}_{\varphi^*}}+\frac{(N-n)}{m}\frac{\bar{x}_{\bar{s}}}{\bar{x}_{w^*}}$ | $\frac{\sum_{i\in r} x_i}{m}$ | $\frac{\sum_{i\in r} x_i}{m}$ |
| **nn** | $1+(n-m)\frac{(\hat{\varphi}_i-1)}{\sum_{i\in r}(\hat{\varphi}_i-1)}\frac{\bar{x}_{\bar{r}}}{\bar{x}_{\varphi^*}}+(N-n)\frac{\hat{\varphi}_i}{\sum_{i\in r}\hat{\varphi}_i}\frac{\bar{x}_{\bar{s}}}{\bar{x}_{w^*}}$ | $\frac{\sum_{i\in r}(\hat{\varphi}_i-1)x_i}{\sum_{i\in r}(\hat{\varphi}_i-1)}$ | $\frac{\sum_{i\in r}\hat{\varphi}_i x_i}{\sum_{i\in r}\hat{\varphi}_i}$ |

Note that $\sum_{i\in r} w_{ir} x_i = \sum_{i\in U} x_i$.

### 8.2. Bias correction method- multiple regression

According to (35), with auxiliary variables, the prediction of finite population total requires computation of $\sum_{i\in\bar{s}} E_{\bar{s}}(y_i|\boldsymbol{x}_i)$ and $\sum_{i\in\bar{r}} E_{\bar{r}}(y_i|\boldsymbol{x}_i)$. Here, we use the "bias correction method" proposed by (Chambers 2003), see Chambers and Clark (2012, page 114).

**Computation of** $\sum_{i\in\bar{s}} E_{\bar{s}}(y_i|\boldsymbol{x}_i)$:

$$\sum_{i\in\bar{s}} E_{\bar{s}}(y_i|\boldsymbol{x}_i) = \sum_{i\in\bar{s}}\{E_{\bar{s}}(y_i|\boldsymbol{x}_i) + E_s(y_i|\boldsymbol{x}_i) - E_s(y_i|\boldsymbol{x}_i)\}$$
$$= \sum_{i\in\bar{s}} E_s(y_i|\boldsymbol{x}_i) + \sum_{i\in\bar{s}}\{E_{\bar{s}}(y_i|\boldsymbol{x}_i) - E_s(y_i|\boldsymbol{x}_i)\}$$
$$\cong \sum_{i\in\bar{s}} E_s(y_i|\boldsymbol{x}_i) + N - n\frac{1}{N-n}\sum_{i\in\bar{s}} E_{\bar{s}}\big(y_i - E_s(y_i|\boldsymbol{x}_i)\big). \quad (48)$$

Now, using (3), $E_{\bar{s}}\big(y_i - E_s(y_i|\boldsymbol{x}_i)\big)$ can be estimated by

$$\hat{E}_{\bar{s}}\left(y_i - \hat{E}_s(y_i|\boldsymbol{x}_i)\right) = \frac{1}{\sum_{i\in r}\hat{\varphi}_i(w_i-1)}\sum_{i\in r}\hat{\varphi}_i(w_i-1)\left(y_i - \hat{E}_r\left(\frac{\hat{\varphi}_i}{\hat{E}_r(\hat{\varphi}_i|\boldsymbol{x}_i)}y_i\right)\right). \quad (49)$$

Also, using (2), $E_s(y_i|\boldsymbol{x}_i)$ can be estimated by

$$\hat{E}_s(y_i|\boldsymbol{x}_i) = \hat{E}_r\left(\frac{\hat{\varphi}_i}{\hat{E}_r(\hat{\varphi}_i|\boldsymbol{x}_i)}y_i\right). \quad (50)$$

**Computation of** $\sum_{i\in\bar{r}} E_{\bar{r}}(y_i|\boldsymbol{x}_i)$:

Similarly,

$$\sum_{i\in\bar{r}} E_{\bar{r}}(y_i|\boldsymbol{x}_i) = \sum_{i\in\bar{r}}\{E_{\bar{r}}(y_i|\boldsymbol{x}_i) + E_r(y_i|\boldsymbol{x}_i) - E_r(y_i|\boldsymbol{x}_i)\}$$
$$= \sum_{i\in\bar{r}} E_r(y_i|\boldsymbol{x}_i) + \sum_{i\in\bar{r}}\{E_{\bar{r}}(y_i|\boldsymbol{x}_i) - E_r(y_i|\boldsymbol{x}_i)\}$$
$$\cong \sum_{i\in\bar{r}} E_r(y_i|\boldsymbol{x}_i) + n - m\frac{1}{n-m}\sum_{i\in\bar{r}} E_{\bar{r}}\big(y_i - E_r(y_i|\boldsymbol{x}_i)\big). \quad (51)$$

But, using (4), $E_{\bar{r}}\big(y_i - E_r(y_i|\boldsymbol{x}_i)\big)$ can be estimated by

$$\hat{E}_{\bar{r}}\left(y_i - \hat{E}_r(y_i|\boldsymbol{x}_i)\right) = \frac{1}{\sum_{i\in r}(\hat{\varphi}_i-1)}\sum_{i\in r}(\hat{\varphi}_i-1)\left(y_i - \hat{E}_r(y_i|\boldsymbol{x}_i)\right). \quad (52)$$

Hence,

$$\hat{T}_{3,in} = \sum_{i\in r} y_i + \sum_{i\in\bar{r}} \hat{E}_r(y_i|\boldsymbol{x}_i) + (n-m)\frac{1}{\sum_{i\in r}(\hat{\varphi}_i-1)}\sum_{i\in r}(\hat{\varphi}_i-1)\left(y_i - \hat{E}_r(y_i|\boldsymbol{x}_i)\right)$$
$$+ \sum_{i\in\bar{s}} \hat{E}_r\left(\frac{\hat{\varphi}_i}{\hat{E}_r(\hat{\varphi}_i|\boldsymbol{x}_i)}y_i\right) + (N-n)\frac{1}{\sum_{i\in r}\hat{\varphi}_i(w_i-1)}\sum_{i\in r}\hat{\varphi}_i(w_i-1)\left(y_i - \hat{E}_r\left(\frac{\hat{\varphi}_i}{\hat{E}_r(\hat{\varphi}_i|\boldsymbol{x}_i)}y_i\right)\right). \quad (53)$$

### 8.3. Generalized regression estimator (GREG) under informative sampling and nonignorable nonresponse

Assume that

$$E_p(y_i) = \boldsymbol{x}_i'\boldsymbol{\beta} = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p, \tag{54}$$

then the GREG estimator is:

$$\hat{T} = \sum_{i=1}^{N} \hat{E}_p(y_i) = \sum_{i=1}^{N} \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\varphi w} \quad , \widehat{\boldsymbol{\beta}}_{\varphi w} = \left(\sum_{i\in r} \varphi_i w_i \boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1}\left(\sum_{i\in r} \varphi_i w_i \boldsymbol{x}_i' y_i\right). \tag{55}$$

**Justification:** under (54) and using (1), we can show that

$$\widehat{\boldsymbol{\beta}} = arg\min_{\tilde{\beta}} E_p(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 = arg\min_{\tilde{\beta}} \sum_{i\in r} \varphi_i w_i (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2. \tag{56}$$

Therefore,

$$\widehat{\boldsymbol{\beta}}_{\varphi w} = \left(\sum_{i\in r} \varphi_i w_i \boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1}\left(\sum_{i\in r} \varphi_i w_i \boldsymbol{x}_i' y_i\right) = (\boldsymbol{x}'(\Phi W)\boldsymbol{x})^{-1}\boldsymbol{x}'(\Phi W)\boldsymbol{y}, \tag{57}$$

where $= diag(\varphi_1 w_1, \dots, \varphi_r w_r)$ $\boldsymbol{x} = (\boldsymbol{x}_1', \dots, \boldsymbol{x}_r')'$. Then

$$\hat{T} = \sum_{i=1}^{N} \hat{E}_p(y_i) = \sum_{i=1}^{N} \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\varphi w}. \tag{58}$$

Now, if $\boldsymbol{x}_i = \left(x_{i1}, \dots, x_{ip}\right)' = (1, \tilde{\boldsymbol{x}}_i')'$, then

$$\hat{T} = \sum_{i=1}^{N} \hat{E}_p(y_i) = \left(\sum_{i=1}^{N} [1, \tilde{\boldsymbol{x}}']\right)\begin{bmatrix} \hat{\beta}_{1\varphi w} \\ \widehat{\tilde{\boldsymbol{\beta}}}_{\varphi w}' \end{bmatrix} = \sum_{i\in r} \varphi_i w_i g_i y_i. \tag{59}$$

where $\bar{\tilde{\boldsymbol{x}}}_U = \left(\bar{x}_{2U}, \dots, \bar{x}_{pU}\right)'$, $\bar{x}_{jU} = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$, $\hat{\beta}_{1\varphi w} = \bar{y}_{\varphi w} - \widehat{\tilde{\boldsymbol{\beta}}}_{\varphi w}'\bar{\tilde{\boldsymbol{x}}}_{\varphi w}$,

$\bar{x}_{j\phi w} = \dfrac{\sum_{i\in r} \phi_i w_i x_{ij}}{\sum_{i\in r} \phi_i w_i}$ $\bar{y}_{\varphi w} = \frac{\sum_{i\in r} \varphi_i w_i y_i}{\sum_{i\in r} \varphi_i w_i}$, and

$$g_i = N\left\{\frac{1}{\sum_{i\in r} \varphi_i w_i} + \left(\bar{\tilde{\boldsymbol{x}}}_U - \bar{\tilde{\boldsymbol{x}}}_{\varphi w}\right)'\left(\sum_{i\in r} \varphi_i w_i \boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1}\boldsymbol{x}_i\right\}. \tag{60}$$

Not that if $\bar{\tilde{\boldsymbol{x}}}_U - \bar{\tilde{\boldsymbol{x}}}_{\varphi w} = 0$, that is, $\bar{x}_{jU} = \bar{x}_{j\varphi w}$, or

$$\frac{1}{N}\sum_{i=1}^{N} x_{ij} = \frac{\sum_{i\in r} \varphi_i w_i x_{ij}}{\sum_{i\in r} \varphi_i w_i}, \tag{61}$$

then

$$\hat{T} = \left(\frac{N}{\sum_{i\in r} \varphi_i w_i}\right)\sum_{i\in r} \varphi_i w_i y_i. \tag{62}$$

Furthermore, $\hat{T} = \sum_{i\in r} \varphi_i w_i g_i y_i$ (59) belongs to the class of calibration estimators since $\sum_{i\in r} \varphi_i w_i g_i \boldsymbol{x_i} = \boldsymbol{x}_U$, see Deville and Särndal (1992). According to this, we can derive a calibration estimator of the finite population total $T = \sum_{i=1}^{N} y_i$ when sampling design is informative and nonresponse mechanism is nonignorable as follows.

It should be noted here that equation (61) can be considered as calibration constraint when sampling design is informative and nonresponse mechanism is nonignorable. That is, the calibration estimator of $T = \sum_{i=1}^{N} y_i$can be obtained by minimizing

$$\sum_{i\in r} \frac{\left(w_i^{cal} - \varphi_i w_i\right)^2}{\varphi_i w_i} \tag{63}$$

with respect to $w_i^{cal}$, subject to the constraint

$$\sum_{i\in r} w_i^{cal} \boldsymbol{x}_i = \boldsymbol{X}_U. \tag{64}$$

Let $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)'$ be the Lagrange multiplier, so the Lagrange function is:

$$\psi\left(w_1^{cal}, \dots, w_n^{cal}; \boldsymbol{\lambda}\right) = \sum_{i\in r} \frac{\left(w_i^{cal} - \varphi_i w_i\right)^2}{\varphi_i w_i} - 2\left(\sum_{i\in r} w_i^{cal} \boldsymbol{x}_i - \boldsymbol{X}_U\right)' \boldsymbol{\lambda}. \tag{65}$$

Differentiating (65) with respect to $w_i^{cal}$ and $\boldsymbol{\lambda}$, and then equating the derivatives to zero, we get the calibration weights:

$$w_i^{cal} = \varphi_i w_i (1 + \boldsymbol{x}_i' \boldsymbol{\lambda}). \tag{66}$$

where $\boldsymbol{\lambda}$ is determined by the constraint $\sum_{i\in r} w_i^{cal} \boldsymbol{x}_i = \boldsymbol{X}_U$, which is equal to

$$\boldsymbol{\lambda} = \left(\sum_{i\in r} \varphi_i w_i \boldsymbol{x}_i' \boldsymbol{x}_i\right)^{-1} \left(\sum_{i\in r} w_i^{cal} \boldsymbol{x}_i - \boldsymbol{X}_U\right). \tag{67}$$

If $(\sum_{i\in r} \varphi_i w_i \boldsymbol{x}_i' \boldsymbol{x}_i)$ is invertible, then the calibration estimator of $T = \sum_{i=1}^{N} y_i$ is

$$\hat{T}_{Cal} = \sum_{i\in r} \varphi_i w_i g_i^{cal} y_i = \sum_{i\in r} w_i^{cal} y_i \tag{68}$$

where $w_i^{cal} = \varphi_i w_i g_i^{cal}$ and $g_i^{cal}$ is given by:

$$g_i^{cal} = 1 + \left(\tilde{\boldsymbol{x}}_U - \tilde{\boldsymbol{x}}_{\varphi w}\right)' \left(\sum_{i\in r} \varphi_i w_i \boldsymbol{x}_i' \boldsymbol{x}_i\right)^{-1} \boldsymbol{x}_i. \tag{69}$$

Variance of $\hat{T}_{Cal} = \sum_{i\in r} \varphi_i w_i g_i^{cal} y_i = \sum_{i\in r} w_i^{cal} y_i$.

Following Deville and Särndal (1992), the estimated variance of $\hat{T}_{Cal}$ (equation 68) is given by:

$$\hat{V}\left(\hat{T}_{Cal}\right) = \sum_{i\in r} \sum_{j\in r} \left(1 - \frac{(\psi_i \pi_i)(\psi_j \pi_j)}{(\psi_{ij} \pi_{ij})}\right) \left(w_i^{cal} e_i\right)\left(w_j^{cal} e_j\right), \tag{70}$$

where $\psi_{ij} = Pr(i, j \in r)$, $\pi_{ij} = Pr(i, j \in s)$ and $e_i = y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_{\varphi w}$.

## 9. Conclusions

In this paper, we study, within a modeling framework, the semi-parametric prediction of finite population total, by specifying the probability distribution of the observed measurements under informative sampling and nonignorable nonresponse. This is the most general situation in surveys and other combinations of sampling informativeness and response mechanisms can be considered as special cases. Furthermore, based on the relationship between response distribution and population distribution, we introduced a new measure of representativeness of a response set and a new test of nonignorable nonresponse and informative sampling, jointly. In addition to that, generalized regression (GREG) and calibration estimators under informative sampling and nonignorable nonresponse are derived.

The paper is purely mathematical and focuses on the role of informativeness of sampling design and informativeness of nonresponse in adjusting various predictors for bias reduction. Further experimentation (simulation and real data problem) with this kind of semiparametric predictors, generalized measures of representativeness, tests of nonignorable nonresponse, informativeness of sampling design, and calibration estimators are therefore highly recommended. The author hopes that the new mathematical results obtained will encourage further theoretical, empirical and practical research in these directions.

## Acknowledgements

I would like to thank the editor-in-chief and the referees for their valuable comments. The research was partially supported by a grant from DAAD (Deutscher Akademischer Austauschdienst German Academic Exchange Service) – Research Stays for University Academics and Scientists, 2018. Also, the author would like to thank Professor Timo Schmid for hosting me during my visit to Free University Berlin.

## References

Chambers, R., Skinner, C., (2003). *Analysis of Survey Data*. New York: John Wiley.

Deville, J. C., Särndal, C. E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87, pp. 376–382.

Eideh, A. H., (2020). Parametric Prediction of Finite Population Total under Informative Sampling and Nonignorable Nonresponse. *Statistics In Transition new series*, March 2020, vol. 21, no.1, pp. 13–35.

Eideh, A. H., (2016). Estimation of Finite Population Mean and Superpopulation Parameters when the Sampling Design is Informative and Nonresponse Mechanism is Nonignorable. Pakistan Journal of Statistics and Operation Research (PJSOR). *Pak.j.stat.oper.res*, vol. XII, no. 3, 2016, pp. 467–489.

Eideh A. H., (2012). Estimation and Prediction under Nonignorable Nonresponse via Response and Nonresponse Distributions. *Journal of Indian Society of Agriculture Statistics*, 66(3), pp. 359–380.

Eideh A. H., (2009). On the use of the Sample Distribution and Sample Likelihood for Inference under Informative Probability Sampling. *DIRASAT (Natural Science),* vol. 36 (2009), no.1, pp. 18–29.

Eideh A. H., (2007). Method of Moments Estimators of Finite Population Parameters in Case of Full Response and Nonresponse. *Contributed Paper for the 56th Biennial Session of the International Statistical Institute*, August 22–29, Lisboa, Portugal, ISI 2007 Book of Abstracts, p. 430.

Fuller W. A., (2009). Sampling Statistics. Wiley.

Little, R. J. A., (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, **77**, pp. 237–250.

Pfeffermann, D., Krieger, A. M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, pp. 1087–1114.

Schouten, B., Cobben, F. and Bethlehem, J., (2009). Indicators for the representativeness of survey Response. *Survey Methodology*, 35, pp. 101–113.

Sverchkov, M., (2008). A New Approach to Estimation of Response Probabilities When Missing Data Are Not Missing at Random. *Proceedings of the Survey Research Methods Section*, pp. 867–874.

Sverchkov, M., Pfeffermann, D., (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology*, 30, pp. 79–92.