Prediction of Finite Population Total under Nonignorable Nonresponse via Response and Nonresponse Distributions

Eideh, Abdulhakeem
Al-Quds University, Department of Mathematics
Abu-Dies Campus, Palestine
P.O. Box 20002, Jerusalem
E-mail: msabdul@science.alquds.edu

Abstract
This paper defined and studies the use of the response and nonresponse distributions for the prediction of finite population totals under single-stage sampling, when the sampling design is noninformative and missing value mechanism is nonignorable. The proposed predictors employ the response set values of the target study variable, and the weights of the response set units. The prediction problem is treated by estimating the expectations of the study values for unobserved units in sample – nonresponse set, and for non-sample units. The main features of the present predictors are their behaviours in terms of the nonignorable nonresponse parameters. Also the use of the best linear unbiased predictors and estimators that ignore the nonignorable nonresponse yield biased predictors and bias estimators.

Keywords: Best Linear Unbiased Predictor, Nonignorable Nonresponse, Probability-Weighted Estimator, Response distribution.

1. Introduction

Survey data may be viewed as the outcome of two processes: the process that generates the values of units in the finite population, often referred as the superpopulation model, and the process of selecting the sample units from the finite population, known as the sample selection mechanism. Analytic inference from sample survey data refers to the superpopulation model. When the sample selection probabilities depend on the values of the model response, even after conditioning on the auxiliary variables, the sampling mechanism becomes informative (outcome variable is correlated with design variables not included in the model) and the selection effects need to be accounted for in the inference process. In addition to the effect of complex sample design, one of the major problems in the analysis of survey data is the inability to obtain useful data on all questionnaire items from all members of the sample. We call this problem nonresponse or missing value problem. In short, by nonresponse (or missing value) is meant that the desired data are not obtained for the entire sample. For more discussion on nonresponse; see foe example, Särndal, C.E. and Lundstorm, S. (2005), and Little and Rubin (2002).

The plan of this paper is as follows. In Section 2 we discuss response and nonresponse distribution. Section 3 justifies unified probability weighted estimators via method of moments in case of nonignorable nonresponse. Section 4, is devoted to response likelihood and estimation. Section 5 discussed the prediction of finite population total under nonignorable nonresponse. We conclude with brief conclusions in Section 6.

2. Response and Nonresponse Distributions

Let $U = \{1,...,N\}$ denote a finite population consisting of N units. Let y be the study variable of interest and let y_i be the value of y for the ith population unit. A predictor is needed for the population total of y, $T = \sum_{i \in U} y_i$. A probability sample s is drawn from U according to a specified sampling design. The sample size is denoted by n. The sampling design induces inclusion probabilities for the different units of U. Let $\pi_i = \Pr(i \in s)$ be the first order inclusion probability

of the ith population unit. Let $\mathbf{x}_i = \left(x_{i1},...,x_{ip}\right)'$, $i \in U$ be the values of a vector of auxiliary variables, $x_1,...,x_p$, and $\mathbf{z} = \{z_1,...,z_N\}$ 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 $\pi_i = \Pr(i \in s)$, and sampling weight $w_i = 1/\pi_i$; $i = 1,...,N$. In practice, the π_i's may depend on the population values $(\mathbf{x},\mathbf{y},\mathbf{z})$. We express this dependence by writing: $\pi_i = \Pr(i \in s \mid \mathbf{x},\mathbf{y},\mathbf{z})$ for all units $i \in U$. The sample s consists of the subset of U selected at random by the sampling scheme with inclusion probabilities $\pi_1,...,\pi_N$.

Denote by $\mathbf{I} = \left(I_1,...,I_N\right)'$ 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. The sample s is defined accordingly as $s = \{i \mid i \in U, I_i = 1\}$ and its complement by $c = \bar{s} = \{i \mid i \in U, I_i = 0\}$. We assume probability sampling, so that $\pi_i = \Pr(i \in s) > 0$ for all units $i \in U$.

Denote by $R = \left(R_1,...,R_N\right)'$ 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 unit $i \in s$ is not observed. The response set is defined accordingly as $r = \{i \mid i \in s, R_i = 1\}$ and the nonresponse set by $\bar{r} = \{i \mid i \in s, R_i = 0\}$. We assume probability sampling, so that $\pi_i = \Pr(i \in s) > 0$ for all units $i \in U$. Let the response probability (or propensity score) $\psi_i = \Pr(i \in r \mid i \in s, \mathbf{x},\mathbf{y},\mathbf{z}) = \Pr(i \in r \mid \mathbf{x},\mathbf{y},\mathbf{z})$ for all units $i \in s$ and $\phi_i = 1/\psi_i$ be the response weight for $i \in r$.

A key issue that must be confronted when dealing with missing data or nonresponse is the relationship between the response indicator (vector) variable, the sample selection indicator membership, the study variable, and the auxiliary population variable. Little and Rubin (2002) consider three types of nonresponse mechanism or missing data mechanism:

(a) Missing completely at random (MCAR): If the response probability does not depend on the study variable, or the auxiliary population variable, the missing data are MCAR.

(b) Missing at random (MAR) given auxiliary population variable: if the response probability depends on the auxiliary population variable but not on the study variable, the missing data are MAR.

(c) Not missing at random (NMAR): if the response probability depends on the value of a missing study variable, the missing data are NMAR.

In this paper we make distinction between ignorable and nonignorable response mechanism. (a) The response mechanism can be ignored conditional on $\mathbf{x}_i$ (or ignorable nonresponse) if:

$$\Pr(i \in r \mid i \in s, \mathbf{x}_i, y_i) = \Pr(i \in r \mid i \in s, \mathbf{x}_i) \quad \text{for all possible values } y_i$$

(b) The response mechanism cannot be ignored (nonignorable nonresponse) if:

$$\Pr(i \in r \mid i \in s, \mathbf{x}_i, y_i) \neq \Pr(i \in r \mid i \in s, \mathbf{x}_i) \quad \text{for all possible values } y_i$$

Before defining the response and nonresponse distribution mathematically, let us introduce the following notations: f_p and $E_p(\cdot)$ denote the probability density function (pdf) and the mathematical expectation of the population distribution, respectively, f_s and $E_s(\cdot)$ denote the pdf and the mathematical expectation of the sample distribution, respectively, f_r and $E_r(\cdot)$ denote the pdf and the mathematical expectation of the response distribution, respectively, and $f_{\bar{r}}$ and $E_{\bar{r}}(\cdot)$ denote the pdf and the mathematical expectation of the nonresponse distribution, respectively.

Eideh (2009) defined and studies the properties of response and nonresponse distributions when the sampling design is informative and missing value mechanism is nonignorable. In this paper, from now on, we assume that the sampling design is noninformative, that is, $f_s(y_i \mid \mathbf{x}_i, \theta, \gamma) = f_p(y_i \mid \mathbf{x}_i, \theta)$. Using the results derived in Eideh (2009), we have:

(a) The (marginal) response pdf of y_i is:

$$f_r(y_i \mid \mathbf{x}_i) = \frac{E_p(\psi_i \mid \mathbf{x}_i, y_i) f_p(y_i \mid \mathbf{x}_i)}{E_p(\psi_i \mid \mathbf{x}_i)} \tag{1}$$

(c) The (marginal) nonresponse pdf of y_i is defined as:

$$f_{\bar{r}}(y_i \mid \mathbf{x}_i) = \frac{\{1 - E_p(\psi_i \mid \mathbf{x}_i, y_i)\} f_p(y_i \mid \mathbf{x}_i)}{\{1 - E_p(\psi_i \mid \mathbf{x}_i)\}} \tag{2}$$

(d) For vector of random variables $(y_i, \mathbf{x}_i)$, the following relationships hold:

$$E_r(\phi_i \mid y_i) = \{E_p(\psi_i \mid y_i)\}^{-1} \tag{3a}$$

$$E_p(y_i) = \{E_r(\phi_i)\}^{-1} E_r(\phi_i y_i) \tag{3b}$$

$$E_{\bar{r}}(y_i \mid \mathbf{x}_i) = \frac{E_p\{(1-\psi_i) y_i \mid \mathbf{x}_i\}}{E_p\{(1-\psi_i) \mid \mathbf{x}_i\}} = \frac{E_r\{(\phi_i - 1) y_i \mid \mathbf{x}_i\}}{E_r\{(\phi_i - 1) \mid \mathbf{x}_i\}} \tag{3c}$$

According to equation (1), for a given population distribution, the response distribution is completely determined by the specification of the conditional expectations of response probabilities, $E_p(\psi_i \mid y_i, \mathbf{x}_i)$. So in order to obtain the response pdf of y_i, we need to model these population conditional expectations. In this paper we consider only the following model for this population conditional expectation:

$$E_p(\psi_i \mid y_i, \mathbf{x}_i) = \Pr(i \in r \mid y_i, \mathbf{x}_i . i \in s) = \exp(a_0 + a_1 y_i + h_1(\mathbf{x}_i)) \tag{4}$$

for some function $h_1(\mathbf{x})$, where $\{a_j, j = 0,1\}$ are unknown parameters to be estimated from the respondent set.

3. Unified Probability Weighted Estimators under Nonignorable Nonresponse

In this section, we derive known results in probability sampling theory from the relationships given in Section 2. Also we prove that probability weighted estimator, in case of nonresponse, is just the method of moments estimator based the response and nonresponse distributions. So we have a new justification of the use of probability weighted estimators.

Let $y_1, ..., y_N$ be N independent and identically distributed random variable from a population with finite mean $E_p(y_i) = \mu$ and finite variance $Var_p(y_i) = \sigma^2$. The method of moments estimates of μ and σ^2 are the solutions of the method of moments equations: $E_p(y_i) = \mu = \bar{Y}_U = N^{-1} \sum_{i \in U} y_i$ and $E_p(y_i^2) = \sigma^2 + \mu^2 = N^{-1} \sum_{i \in U} y_i^2$ which are: $\tilde{\mu} = \bar{Y}_U$ and $\tilde{\sigma}^2 = N^{-1} \sum_{i \in U} y_i^2 - \bar{Y}_U^2$. But $\sum_{i \in U} y_i$ and $\sum_{i \in U} y_i^2$ are unknown finite population parameters that need estimation. Now using (3b), we can show that the method of moments estimators of $\bar{Y}_U$ and $S_U^2 = N^{-1} \sum_{i \in U} (y_i - \bar{Y})^2$ under nonignorable nonresponse are:

$$\tilde{\bar{Y}}_U = \bar{y}_w = \frac{\sum_{\in r} \phi_i y_i}{\sum_{\in r} \phi_i} = \bar{y}_\phi$$

and

$$\tilde{S}_U^2 = \frac{\sum_{\in r} \phi_i (y_i - \bar{y}_\phi)^2}{\sum_{\in r} \phi_i}$$

which are the well known probability weighted estimator.

If the nonresponse mechanism is ignorable: in this case, $E_p(y_i) = E_r(y_i)$, so that the method of moments estimators of $\bar{Y}_U = N^{-1} \sum_{i \in U} y_i$ is given by $\hat{\bar{Y}}_{UR} = \sum_{i \in r} y_i / \sum_{i \in r} 1$.

4. Response Likelihood and Estimation

Having derived the response distribution, and if the response measurements are independent, then the logarithm of the response likelihood for θ and η is:

$$l_r(\theta,\gamma)=\sum_{i=1}^{m}\log f_r(y_i\mid \mathbf{x}_i,\theta,\eta)=l_{ign}(\theta)+\sum_{i=1}^{m}\log E_p(\psi_i\mid \mathbf{x}_i,y_i,\eta)-\sum_{i=1}^{m}\log E_p(\psi_i\mid \mathbf{x}_i,\theta,\eta) \quad (5)$$

where $l_{ign}(\theta)=\sum_{i\in r}\log\left(f_p(y_i\mid \mathbf{x}_i,\theta)\right)$ is the classical log-likelihood obtained under ignorable nonresponse.

Based on the response data $\{y_i,\mathbf{x}_i,\phi_i;\ i\in r\}$ we can estimate the parameters of the population model in two steps:

Step-one: Estimate the nonignorable nonresponse parameters η using the following relationship in (3a).

Thus the nonignorable nonresponse parameters can be estimated using regression analysis. Denoting the resulting estimate of η by $\tilde{\eta}$.

Step-two: Substitute $\tilde{\eta}$ in (5), and then maximize the resulting response log-likelihood function with respect to the population parameters, θ:

$$l_r(\theta,\tilde{\eta})=l_{ign}(\theta)+\sum_{i=1}^{m}\log E_p(\psi_i\mid \mathbf{x}_i,y_i,\tilde{\eta})-\sum_{i=1}^{m}\log E_p(\psi_i\mid \mathbf{x}_i,\theta,\tilde{\eta}) \quad (6)$$

where $l_r(\theta,\tilde{\eta})$ is the response log-likelihood after substituting $\tilde{\eta}$ in the response log-likelihood function, (5).

5. Prediction of Finite Population Total under Nonignorable Nonresponse

Sverchkov and Pfeffermann (2004) use sample and sample complement distributions for the prediction of finite population totals under informative sampling for single-stage sampling designs. Later Eideh and Nathan (2009) extend the theory to general linear functions of the population values and to two-stage informative cluster sampling. In this section we use the response and nonresponse distributions to predict the finite population total under noninformative sampling design and under nonignorable nonresponse. We consider the prediction for single-stage sampling and under simple ratio population model (R).

Assume single-stage population model. Let

$$T=\sum_{i\in U}y_i=\sum_{i\in s}y_i+\sum_{i\in \bar{s}}y_i=\sum_{i\in r}y_i+\sum_{i\in \bar{r}}y_i+\sum_{i\in \bar{s}}y_i \quad (7)$$

be the finite population total that we want to predict using the data from the response set and possibly values of auxiliary variables that may include some or all of the design variables. Notice that T can be decomposed into three components: the first component represents the total for observed units in the sample – response set, $\sum_{i\in r}y_i$, the second component represents the total for unobserved units in sample – nonresponse set, $\sum_{i\in \bar{r}}y_i$, and the third component represents the total for non-sample units, $\sum_{i\in \bar{s}}y_i$.

For the prediction process we have the following available information:

(a). The information that comes from the sampling design denoted by:
$O_s=\left[\{(x_i,I_i),i\in U\},\{\pi_i,i\in s\}\right]$, where $I_i=1$ for $i\in s$ and $I_i=0$ for $i\notin s$.

(b). Information that comes from the response set denoted by: $O_r=\left[\{(y_i,\psi_i),i\in r\}\middle| i\in s\right]$, N, n, and m.

Thus the available information, from the sample and response set, for the prediction process is $O=O_s\bigcup O_r$.

Let $\hat{T}=\hat{T}(O)$ define the predictor of T based on O. We can show that the mean square error (MSE) of $\hat{T}$ given O: $MSE_p(\hat{T})=E_p\left\{(\hat{T}-T)^2\mid O\right\}=\left\{\hat{T}-E_p(T\mid O)\right\}^2+Var_p(T\mid O)$ is minimized when

$\hat{T} = E(T \mid O) = \sum_{i\in r} y_i + \sum_{i\in \bar{r}} E_{\bar{r}}\left(y_i|O\right) + \sum_{i\in \bar{s}} E_p\left(y_i|O\right)$. We know the values $\{y_i's, i \in r\}$, so the sum $\sum_{i\in r} y_i$ is known. Thus to estimate for our response set, we need to predict the total for unobserved units in the sample – nonresponse set, $\sum_{i\in \bar{r}} y_i$, and the total for non-sample units, $\sum_{i\in \bar{s}} y_i$. That is, to predict T we need to predict values for the $\{y_i's, i \in \bar{r}\}$ and values for the $\{y_i's, i \in \bar{s}\}$. According to (3b and 3c), we can show that:

$$\begin{aligned} \hat{T}_{nign} &= \hat{T}_{ign} - \sum_{i\in \bar{r}} \frac{Cov_p\left[(\psi_i, y_i)|O\right]}{1 - E_p(\psi_i)} \\ &= \hat{T}_{ign} - \sum_{i\in \bar{r}} \frac{Cov_r(\phi_i, y_i)}{E_r(\phi_i)E_r\{(\phi_i - 1)\}} \end{aligned} \tag{8}$$

where

$$\hat{T}_{ign} = \sum_{i\in r} y_i + \sum_{i\in \bar{r}} E_p\left(y_i|O\right) + \sum_{i\in \bar{s}} E_p\left(y_i|O\right)$$

is the best linear unbiased predictor (BLUP) of $T = \sum_{i\in U} y_i$ under ignorable nonresponse.

We can show that the nonresponse bias of $\hat{T}_{nign}$ is given by:

$$B\left(\hat{T}_{nign}\right) = E_p\left(\hat{T}_{nign} - T\right) = -\sum_{i\in \bar{r}} \frac{Cov_p(\psi_i, y_i)}{E_p(1 - \psi_i)} \tag{9}$$

Hence, the predictor $\hat{T}_{nign}$ is unbiased if there is no correlation between the study variable and the response probabilities ψ_i. The stronger the relationship between the study variable and the response probability, the larger the bias. Similar result was obtained by Bethlehem (1988).

As an illustration, we apply the results under simple ratio population model (R) stating that: $y_i|x_i \underset{p}{\sim} N\left(\beta x_i, \sigma^2 x_i\right)$ are independent normal random variable. If $E_p(\psi_i \mid y_i) = \exp(\eta y_i)$, then we can show that, $y_i \mid x_i \underset{r}{\sim} N\left(\left(\eta\sigma^2 + \beta\right)x_i, \sigma^2 x_i\right), i \in r$. Thus,

$$\hat{T}_{nign,r} = \hat{T}_{R,ign} - \sum_{i\in \bar{r}} \left\{ \left(\eta\sigma^2 x_i\right) \frac{\exp\left(\eta(\beta x_i) + \frac{\eta^2\sigma^2 x_i}{2}\right)}{1 - \exp\left(\eta(\beta x_i) + \frac{\eta^2\sigma^2 x_i}{2}\right)} \right\} \tag{10}$$

where

$$\hat{T}_{R,ign} = \sum_{i\in r} y_i + \sum_{i\in \bar{r}} \beta x_i + \sum_{i\in \bar{s}} \beta x_i$$

Under the response distribution: $y_i \mid x_i \underset{r}{\sim} N\left(\left(\eta\sigma^2 + \beta\right)x_i, \sigma^2 x_i\right), i \in r$, we can show that the maximum likelihood estimators of the parameters of simple ratio population model are given by:

$$\hat{\beta}_{R,nign} = \frac{\bar{y}_r}{\bar{x}_r} - \tilde{\eta}\frac{1}{r}\sum_{i\in r}\frac{1}{x_i}\left(y_i - \left(\frac{\bar{y}_r}{\bar{x}_r}\right)x_i\right)^2 \tag{11a}$$

and

$$\hat{\sigma}^2_{R,nign} = \frac{1}{r}\sum_{i\in r}\frac{1}{x_i}\left(y_i - \left(\frac{\bar{y}_r}{\bar{x}_r}\right)x_i\right)^2 \tag{11b}$$

where $\tilde{\eta}$ is the least square estimator obtained via the relationship in (3a).

Now, if the missing value mechanism is ignorable, that is, $\eta = 0$, then

$$\hat{\beta}_{R,ign} = \frac{\bar{y}_r}{\bar{x}_r} \quad \text{and} \quad \hat{\sigma}^2_{R,ign} = \frac{1}{r}\sum_{i\in r}\frac{1}{x_i}\left(y_i - \left(\frac{\bar{y}_r}{\bar{x}_r}\right)x_i\right)^2 \tag{12}$$

Therefore,

$$\hat{T}_{R,ign} = \frac{\bar{y}_r}{\bar{x}_r} N\bar{X} \tag{13}$$

which is the classical ratio estimator under nonignorable nonresponse and under noninformative sampling design.

Finally, the bias of $\hat{T}_{R,nign}$ is given by:

$$B\left(\hat{T}_{nign}\right) = E_p\left(\hat{T}_{nign} - T\right) = -\sum_{i\in\bar{r}}\left\{\left(\eta\sigma^2 x_i\right)\frac{\exp\left(\eta\left(\beta x_i\right) + \frac{\eta^2\sigma^2 x_i}{2}\right)}{1-\exp\left(\eta\left(\beta x_i\right) + \frac{\eta^2\sigma^2 x_i}{2}\right)}\right\} \tag{14}$$

6. Conclusions

In this paper we consider a new method of estimating the parameters of the superpopulation model for single-stage sampling from a finite population when the sampling design is noninformative and the response mechanism is nonignorable. We provide new justification for the broad use of probability-weighted estimators in estimating finite population parameters in case of ignorable nonresponse. Furthermore we fit the simple ratio population model under noninformative sampling design and under nonignorable nonresponse. In addition to the estimation problem we introduce new predictors of the finite population total for the simple ratio population model. These new predictors take into account the nonignorable nonresponse. Thus, also provides new justification for the broad use of best linear unbiased predictors (model-based school) in predicting finite population parameters in case of ignorable nonresponse. The main features of the present predictors and estimators are their behaviours in terms of the nonignorable nonresponse parameters. Also the use of the best linear unbiased predictors and estimators that ignore the nonignorable nonresponse yield biased predictors and bias estimators. I hope that the new mathematical results obtained will encourage further theoretical, empirical and practical research in these directions.

REFERENCES

Bethlehem, J.G. (1988). Reduction of Nonresponse Bias through Regression Estimation. Journal of Official Statistics, 4, pp 251-260.

Eideh A.H. (2009). On the use of the Sample Distribution and Sample Likelihood for Inference under Informative Probability Sampling. DIRASAT (Natural Science), Volume 36 (2009), Number 1, pp18-29.

Eideh, A. H. and Nathan, G. (2009). Two-Stage Informative Cluster Sampling with application in Small Area Estimation. Journal of Statistical Planning and Inference.139, pp 3088-3101.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. New York: Wiley.

Särndal, C.E. and Lundstorm, S. (2005). Estimation in Surveys with Nonresponse. New York: John Wiley.

Sverchkov, M., and Pfeffermann, D. (2004). Prediction of Finite Population Totals based on the Sample Distribution. Survey Methodology, 30, pp 79-92.

RÉSUMÉ

Dans cet article, nous étudions l'utilisation de la distribution de réponse et de non réponse pour la prédiction d'un total en population finie, lorsque que le plan d'échantillonnage n'est pas informatif et que les valeurs manquantes sont non ignorable. Le prédicateur proposé utilise les valeurs des répondants pour les variables d'intérêt, et les pondérations des répondants. Les problèmes d'estimation est traité en estimant les espérances des valeurs de la variables d'intérêt pour les unités non observées dans l'échantillon. Nous illustrons les résultats obtenus en dérivant un nouveau prédicateur pour le total d'une population finie en termes de paramètres non ignorable de non réponse. Nous utilisons aussi le meilleur prédicateur non biaisé. Les estimateurs qui ignorent la non réponse non ignorable produisent les prédicateurs et estimateurs biaisés.