Analysis of Survey Data under Informative Sampling Design and Nonignorable Nonresponse Mechanism

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

Abstract

In this paper, we study, within a modeling framework, the joint treatment of not missing at random response mechanism and informative sampling for survey data, by specifying the probability distribution of the observed measurements when the sampling design is informative. This is the most general situation in surveys and other combinations of sampling informativeness and response mechanisms can be considered as special cases. The sample distribution of the observed measurements model is extracted from the population distribution model, such as the normal distribution. The sample distribution is derived first by identifying and estimating the conditional expectations of first order sample inclusion probabilities, given the study variable, based on a variety of models, such as linear, exponential, logit and probit. Next, we consider a logistic model, probit and other models for the not missing at random response mechanism. The proposed method combines two methodologies used in the analysis of sample surveys for the treatment of informative sampling and not missing at random response mechanism. One incorporates the dependence of the first order inclusion probabilities on the study variable, while the other incorporates the dependence of the probability of nonresponse on unobserved or missing observations. The main purpose here is to consider how to account for the joint effects of informative sampling designs and of not missing at random response mechanism in statistical models for complex survey data.

Keywords: response distribution, response likelihood,

1. Introduction

Data collected by sample surveys are used extensively to make inferences on assumed population models. Often, survey design features (clustering, stratification, unequal probability selection, etc.) are ignored and the sample data are then analyzed using classical methods based on simple random sampling. This approach can, however, lead to erroneous inference because of sample selection bias implied by informative sampling - 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 the model). For example, if the sample design is clustered, with PSU's selected with probabilities proportional to size (e.g., size of locality) and the dependent variable (e.g., income) is related to the size of the locality, ignoring the effects of this dependence can cause bias in the estimation of regressions coefficients. In theory, the effect of the sample selection can be controlled for by including among the model all the design variables used for the sample selection. However, this possibility is often not operational because there may be too many of them or because they are not of substantive interest. To overcome the difficulties associated with the use of classical inference procedures for cross sectional survey data, Pfeffermann, Krieger and Rinott (1998) proposed the use of the sample distribution induced by assumed population models, under informative sampling, and developed expressions for its calculation. Similarly, Eideh and Nathan (2006) fitted time series models for longitudinal survey data under informative sampling. 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. 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.

So, the cross-classification of sampling design and response mechanism is summarized in the following table:

Table 1

Sampling Design	**Response Mechanism**		
	MCAR	MAR	NMAR
Informative-INF	INFMCAR	INFMAR	INFNMAR
Noninformative-NINF	NINFMCAR	NINFMAR	NINFNMAR

For inference problem, Little (1982) classify the nonresponse mechanism as ignorable (MAR and MCAR) and nonignorable (NMAR). Foe this sense, the cross classification of sampling design and nonresponse mechanism is:

Table 2

Sampling Design	**Nonresponse Mechanism**	
	Ignorable	Nonignorable
Informative	**ii**	**in**
Noninformative	**ni**	**nn**

Eideh (2012) consider estimation of superpopulation parameters and prediction of finite population parameters (census paparemters) under nonignorable nonresponse via response and nonresponse distributions when the sampling design in noninformative. None of the above studies consider simultaneously the problem of informative sampling and the problem of nonignorable when analyzing survey data.

In this paper, we study, within a modeling framework, the joint treatment of NMAR response mechanism and informative sampling for survey data, by specifying the probability distribution of the observed measurements when the sampling design is informative. This is the most general situation in surveys and other combinations of sampling informativeness and response mechanisms can be considered as special cases.

2. Preliminaries

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 probability sample s is drawn from U according to a specified sampling design. The sample size is denoted by n. Let $\mathbf{x}_i = (x_{i1},..., x_{ip})'$, $i \in U$ 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 $\pi_i = \Pr(i \in s) > 0$, and sampling weight $w_i = 1/\pi_i$; $i = 1,..., N$. Denote by $R = (R_1,..., R_N)'$ 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. The response set is defined accordingly as

$r = \{i \mid i \in s, R_i = 1\}$ and the nonresponse set by $\bar{r} = \{i \mid i \in s, R_i = 0\}$. Let the response probability $\psi_i = \Pr(i \in r \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$ for all units $i \in s$ and $\phi_i = 1/\psi_i$ be the response weight for $i \in s$.

Eideh (2009) defined the (marginal) response pdf of y_i as:

$$\begin{aligned} f_r(y_i \mid \mathbf{x}_i, \theta, \eta, \gamma) &= \frac{E_s(\psi_i \mid \mathbf{x}_i, y_i, \gamma) f_s(y_i \mid \mathbf{x}_i, \theta, \eta)}{E_s(\psi_i \mid \mathbf{x}_i, \theta, \eta, \gamma)} \\ &= \frac{E_s(\psi_i \mid \mathbf{x}_i, y_i, \gamma) E_p(\pi_i \mid \mathbf{x}_i, y_i, \eta) f_p(y_i \mid \mathbf{x}_i, \theta)}{E_s(\psi_i \mid \mathbf{x}_i, \theta, \eta, \gamma) E_p(\pi_i \mid \mathbf{x}_i, \theta, \eta)} \end{aligned} \tag{1}$$

Similarly, the (marginal) nonresponse pdf of y_i is defined as:

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

Furthermore, for vector of random variables $(y_i, \mathbf{x}_i)$, Eideh (2009), proved the following relationship:

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

Aslo, we proved the following new relationships:

$$E_p(y_i \mid \mathbf{x}_i) = \frac{E_r(\phi_i w_i y_i \mid \mathbf{x}_i)}{E_r(\phi_i w_i \mid \mathbf{x}_i)} \tag{4}$$

$$E_{\bar{s}}(y_i \mid \mathbf{x}_i) = \frac{E_r\{\phi_i (w_i - 1) y_i \mid \mathbf{x}_i\}}{E_r\{\phi_i (w_i - 1) \mid \mathbf{x}_i\}} \tag{5}$$

$$E_{\bar{r}}(y_i) = E_s(y_i) - \frac{Cov_s(\psi_i, y_i)}{1 - E_s(\psi_i)} \tag{6}$$

4. Estimation under Informative Sampling and NMAR Nonresponse Mechanism

Having derived the response distribution when the sampling design is informative and the nonresponse mechanism in nonignorable (NMAR) and if the response measurements are independent, then logarithm of the response likelihood for θ (the parameter indexing the superpopulation model), η (the parameter indexing the sampling design) and γ (the parameter indexing the nonresponse mechanism), is given by:

$$\begin{aligned} l_{r.in}(\theta, \eta, \gamma) = &\sum_{i=1}^{m} \log f_p(y_i \mid \mathbf{x}_i, \theta) + \sum_{i=1}^{m} \log E_p(\pi_i \mid \mathbf{x}_i, y_i, \eta) - \sum_{i=1}^{m} \log E_p(\pi_i \mid \mathbf{x}_i, \theta, \eta) \\ &+ \sum_{i=1}^{m} \log E_s(\psi_i \mid \mathbf{x}_i, y_i, \gamma) - \sum_{i=1}^{m} \log E_s(\psi_i \mid \mathbf{x}_i, \theta, \eta, \gamma) \end{aligned} \tag{7}$$

The response likelihood function, $L_{r,in}(\theta,\eta,\gamma)=\exp(l_{r.in}(\theta,\eta,\gamma))$, can be interpreted as a weighted likelihood, where the weight is the product of the two ratios, the first one is $E_p(\pi_i \mid \mathbf{x}_i, y_i, \eta)/E_p(\pi_i \mid \mathbf{x}_i, \theta, \eta)$, which characterize the sampling design, and the second ratio is $E_s(\psi_i \mid \mathbf{x}_i, y_i, \gamma)/E_s(\psi_i \mid \mathbf{x}_i, \theta, \eta, \gamma)$, that characterise the missing data mechanism.

Now, assuming $f_p(y_i \mid \mathbf{x}_i, \theta)$, $E_p(\pi_i \mid \mathbf{x}_i, y_i, \eta)$ and $E_s(\psi_i \mid \mathbf{x}_i, y_i, \gamma)$ are completely specified, then the maximum likelihood (ML) estimator of (θ,η,γ) can be obtained by maximizing the log likelihood function given in (7) with respect to (θ,η,γ) simultaneously, or in four-step method. For modeling of $E_p(\pi_i \mid \mathbf{x}_i, y_i, \eta)$, Pfeffermann et al. (1998) introduced exponential and polynomial function of $(\mathbf{x}_i, y_i)$, later Eideh (2003) considered logit and probit functions. Furthermore, Eideh (2012) adopted the exponential, linear, logit and probit functions for modeling $E_s(\psi_i \mid \mathbf{x}_i, y_i, \gamma)$.In practice the response probabilities are theoretical quantities and they are unknown. For estimation of ψ_i, see Eideh (2012).

Estimation of (θ,η,γ) - Four steps method

Step 1: Estimation of ψ_i. See Eideh (2012). Denote the estimate by $\hat{\psi}_i$, so that $\hat{\phi}_i = 1/\hat{\psi}_i$. We refer to $\hat{\psi}_i$ as the response propensity.

Step 2: Estimation of the effect of nonresponse mechanism. Estimate the parameter γ using the following relationship: $E_s(\psi_i \mid \mathbf{x}_i, y_i, \gamma) = 1/E_r(\phi_i \mid \mathbf{x}_i, y_i, \gamma)$. Thus the parameter γ can be estimated by regressing $\hat{\phi}_i$ on $(\mathbf{x}_i, y_i)$ using the data set $\{\hat{\phi}_i, y_i, \mathbf{x}_i,\ i \in r\}$. Denoting the resulting estimate of γ by $\tilde{\gamma}$.

Step 3: Estimation of the effect of sampling design. Estimate the parameter η using the relationship given in (4).

Step 4: Estimation of the superpopulation model parameter. Substitute $\tilde{\gamma}$ and $\tilde{\eta}$ in the response log-likelihood function, (7), and since $E_p(\pi_i \mid \mathbf{x}_i, y_i, \tilde{\eta})$ and $E_s(\psi_i \mid \mathbf{x}_i, y_i, \tilde{\gamma})$ do not contain θ, then the ML estimator of θ is obtained by maximizing the resulting response log-likelihood function with respect to the population parameter θ, namely:

$$\tilde{l}_{r.in}(\theta) = \sum_{i=1}^{m} \log f_p(y_i \mid \mathbf{x}_i, \theta) - \sum_{i=1}^{m} \log E_p(\pi_i \mid \mathbf{x}_i, \theta, \tilde{\eta}) - \sum_{i=1}^{m} \log E_s(\psi_i \mid \mathbf{x}_i, \theta, \tilde{\eta}, \tilde{\gamma}) \qquad (8)$$

5. Prediction of Finite Population Parameter under Informative Sampling and NMAR Nonresponse Mechanism

Assume single-stage population model. Let

$$T = \sum_{i=1}^{N} y_i = \sum_{i \in s} y_i + \sum_{i \in \bar{s}} y_i = \sum_{i \in r} y_i + \sum_{i \in \bar{r}} y_i + \sum_{i \in \bar{s}} y_i \qquad (9)$$

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.

For the prediction process we have the following available information:

$$O = \left[\{(y_i, \psi_i, \mathbf{x}_i), i \in r\} \mid i \in s\right] \cup \left[\{(\mathbf{x}_i, I_i), i \in U\}, \{\pi_i, R_i, i \in s\}\right]; N, n, \text{and } m,$$

where $I_i = 1$ for $i \in s$ and $I_i = 0$ for $i \notin s$.

Let $\hat{T} = \hat{T}(O)$ define the predictor of T based on O. It obvious that:

$$MSE_p\left(\hat{T}\right)=E_p\left\{\left(\hat{T}-T\right)^2 \mid O\right\}=\left\{\hat{T}-E_p\left(T \mid O\right)\right\}^2+Var_p\left(T \mid O\right) \tag{10}$$

is minimized when $\hat{T}=E\left(T \mid O\right)$. Hence the minimum mean squared error predictor of $T=\sum_{i=1}^{N} y_i$ is given by:

$$T^{*}=E_p\left(T \mid O\right)=\sum_{i\in r} y_i+\sum_{i\in \bar{r}} E_{\bar{r}}\left(y_i|O\right)+\sum_{i\in \bar{s}} E_{\bar{s}}\left(y_i|O\right) \tag{11}$$

According to (5) and (6), we have

$$T_{in}^{*}=\sum_{i\in r} y_i+\sum_{i\in \bar{r}} E_s\left(y_i|O\right)+\sum_{i\in \bar{s}} E_p\left(y_i|O\right)-\left\{\sum_{i\in \bar{r}} \frac{Cov_s\left[\left(\psi_i, y_i\right)|O\right]}{1-E_s\left(\psi_i|O\right)}+\sum_{i\in \bar{s}} \frac{Cov_p\left[\left(\pi_i, y_i\right)|O\right]}{1-E_p\left(\pi_i|O\right)}\right\} \tag{12}$$

Particular cases:

Case 1: Sampling design is noninformative and nonresponse process is nonignorable, therefore,

$$T_{nn}^{*}=\sum_{i\in r} y_i+\sum_{i\in \bar{r}} E_p\left(y_i|O\right)+\sum_{i\in \bar{s}} E_p\left(y_i|O\right)-\sum_{i\in \bar{r}} \frac{Cov_p\left[\left(\psi_i, y_i\right)|O\right]}{1-E_p\left(\psi_i|O\right)} \tag{13}$$

Case 2: Sampling design is noninformative and nonresponse process is ignorable, therefore,

$$T_{ni}^{*}=\sum_{i\in r} y_i+\sum_{i\in \bar{r}} E_p\left(y_i|O\right)+\sum_{i\in \bar{s}} E_p\left(y_i|O\right) \tag{14}$$

Case 3: Sampling design is informative and nonresponse process is ignorable, therefore,

$$T_{ii}^{*}=\sum_{i\in r} y_i+\sum_{i\in \bar{r}} E_s\left(y_i|O\right)+\sum_{i\in \bar{s}} E_p\left(y_i|O\right)-\sum_{i\in \bar{s}} \frac{Cov_p\left[\left(\pi_i, y_i\right)|O\right]}{1-E_p\left(\pi_i|O\right)} \tag{15}$$

Using (5) and (6), we can show that the nonresponse bias of T_{in} is:

$$\begin{aligned} B\left(T_{in}^{*}\right)&=E_p\left(T_{in}-T\right)=-\left\{\sum_{i\in \bar{r}}\left[\left(E_p\left(y_i\right)-E_s\left(y_i\right)\right)+\frac{Cov_s\left(\psi_i, y_i\right)}{E_s\left(1-\psi_i\right)}\right]+\sum_{i\in \bar{s}} \frac{Cov_p\left(\pi_i, y_i\right)}{1-E_p\left(\pi_i\right)}\right\} \\ &=-\left\{\begin{array}{l}\sum_{i\in \bar{r}}\left\{-\frac{Cov_r\left(\phi_i w_i, y_i\right)}{E_r\left(\phi_i w_i\right)E_r\left\{\left(\phi_i-1\right)\right\}}+\frac{E_r\left(\phi_i w_i y_i\right)E_r\left(\phi_i\right)-E_r\left(\phi_i y_i\right)E_r\left(\phi_i w_i\right)}{E_r\left(\phi_i w_i\right)E_r\left\{\left(\phi_i-1\right)\right\}}\right\} \\ -\sum_{i\in \bar{s}} \frac{E_r\left(\phi_i\right)E_r\left(\phi_i w_i y_i\right)-E_r\left(\phi_i y_i\right)E_r\left(\phi_i w_i\right)}{E_r\left(\phi_i w_i\right)\left[E_r\left(\phi_i w_i\right)-E_r\left(\phi_i\right)\right]}\end{array}\right\} \end{aligned} \tag{16}$$

Hence, the predictor T_{in}^{*} is unbiased T if:

(a) $Cov_s\left(\psi_i, y_i\right)=0$, that is, there is no correlation between the study variable and the response probabilities ψ_i, and

(b) $Cov_p\left(\pi_i, y_i\right)=0$, that is, there is no correlation between the study variable and the first order inclusion probabilities π_i .

Note that, the stronger the relationship between the study variable and the response probability, and the study variable and first order inclusion probabilities, the larger the bias.

References

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

Eideh, A. H. and Nathan, G. (2006). Fitting time series models for longitudinal survey data under informative sampling. *Journal of Statistical Planning and Inference* **136, 9**, pp 3052-306. [Corrigendum, 137 (2007), p 628].

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

Little, R.J.A. and Rubin, D.B. (2002). *Statistical analysis with missing data*. New York: Wiley.

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.

Pfeffermann, D. and Sverchkov, M.(1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya*, **61**, **B**, pp 166-186.