Estimation and Imputation under Nonignorable Nonresponse with Missing Covariate Information

Danny Pfeffermann[1] and Anna Sikov[2]
[1]Department of Statistics, Hebrew University of Jerusalem, Jerusalem, 91905, Israel, and
Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK.
[2] Department of Statistics, Hebrew University of Jerusalem, Jerusalem, 91905, Israel.

Abstract
In this research we develop and apply new techniques for handling nonignorable nonresponse. We assume a model for the outcome variable under complete response and a model for the response probability, which is allowed to depend on the outcome and auxiliary variables. The two models define the model holding for the outcomes observed for the responding units. The unknown parameters governing this model are estimated by maximization of the corresponding likelihood, and we develop alternative maximization algorithms that utilize the information on the population means of some or all the auxiliary variables. We estimate the distribution of the missing covariates and use it for imputing the missing values for the nonresponding units and for estimating the population total of the outcome. We illustrate our approach using a real data set.

Key Words: Calibration, H-T estimator, NMAR, Sample distribution, Sample-complement distribution

1. Introduction

Most of the methods dealing with nonresponse assume that it is missing at random, and that the auxiliary (explanatory) variables are observed for both the respondents and the nonrespondents. These assumptions, however, are not always met. In this research we consider the often practical situation where the probability to respond may depend also on the outcome value after conditioning on the explanatory variables. For example, the probability to obtain information on income may depend on the income level as well as socio-demographic variables. For this kind of response mechanism, the missing outcome values are not missing at random (NMAR), since for the non-responding units the probability of not responding depends on the missing outcomes. We also consider the case of 'unit nonresponse', where the auxiliary information for the nonrespondents is likewise unobserved, except, perhaps, for the population means of some or all of the auxiliary variables. These means are often available from administrative or census records.

We assume a model for the outcome variable under complete response (the *population model*) and a model for the response probabilities, and then maximize the likelihood for the model holding for the observed outcomes (the *sample model*). In order to utilize the additional information provided by the known population totals of the covariates, we iterate between maximization of the likelihood for estimating the population model parameters that feature in the sample model, and solving calibration constraints for estimating the parameters governing the model assumed for the response probabilities. The calibration constraints match pseudo probability weighted estimates of the totals of the covariates and a pseudo probability weighted estimate of the population mean of the population model residuals, with their known population values. See Section 3 for details.

Having estimated the parameters of the response model, we predict the population mean of the outcome values by use of Horvitz-Thompson type estimators. When the covariates are observed for all the sampled units, we estimate the conditional distribution of the outcome values for the nonrespondents given their respective covariates, using relationships developed in Sverchkov and Pfeffermann (2004). We use this distribution for imputing the missing sample data, thus providing analysts with a 'complete' data set. Combining the observed and imputed values enables the computation of another predictor of the outcome population mean. In the case of missing covariate information, this is done by first imputing the missing covariate values, again using the results in Sverchkov and Pfeffermann (2004). See Section 4. We illustrate the proposed procedure and compare it with other methods proposed in the literature in Section 5, using a real data set. Section 6 contains some concluding remarks.

2. Existing Approaches

Let Y_i denote the value of an outcome variable Y associated with unit i belonging to a sample $S=\{1,...,n\}$. We assume that the sample is drawn from a finite population $U=\{1,...,N\}$ by probability sampling with known first order inclusion probabilities $\pi_i=\Pr(i\in S)$. Let $X_i=(X_{1i},...,X_{Ki})$ denote the corresponding values of K auxiliary variables (covariates). In what follows we assume that the population outcomes are independent realizations from distributions with probability density functions (*pdf*), $f_U(Y_i\mid X_i)$, governed by an unknown vector parameter θ. Let $R=\{1,...,n_r\}$ define the subsample of respondents (the subsample with observed covariates and outcome values), and $R^c=\{n_r+1,...,n\}$ define the subsample of nonrespondents, for which at least the outcomes are unobserved. The response process is assumed to be independent between units. The observed sample of respondents can be viewed therefore as the result of a two-phase sampling process, where in the first phase the sample S is selected from U with known inclusion probabilities, and in the second phase the sample R is 'self selected' with unknown response probabilities $q_i=\Pr(i\in R\mid i\in S)$; Särndal and Swensson (1987).

In what follows we assume that the sampling process is noninformative such that under complete response, $f_S(Y_i\mid X_i)=f(Y_i\mid X_i,i\in S)=f_U(Y_i\mid X_i)$, where $f_S(Y_i\mid X_i)$ is the model holding for sampled unit i under complete response. Most of the approaches proposed in the literature to deal with nonresponse assume (sometimes implicitly) that the missing data are 'missing at random' (MAR, see Rubin, 1976 and Little, 1982). This type of nonresponse requires that the probability to respond does not depend on the unobserved data, after conditioning on the observed data. Under this condition, and if the parameters governing the distribution under full response are distinct from the parameters governing the response process, the nonresponse can be ignored for likelihood and Bayesian based inference. Notice that in this case,

$$f_R(Y_i\mid X_i)=f(Y_i\mid X_i,i\in R)=f_S(Y_i\mid X_i) \tag{1}$$

where $f_R(Y_i\mid X_i)$ defines the *marginal pdf* for responding unit i and $f_S(Y_i\mid X_i)$ is the corresponding sample *pdf* defined above. There are many approaches for handling MAR nonresponse, see the books by Schafer (1997) and Little and Rubin (2002) and the recent article by Qin *et al*. (2008) for comprehensive accounts.

In this research we focus on situations where the probability to respond may depend on the outcome value even after conditioning on the covariates. Suppose first that all the covariates are known for every sampled unit. Define by R_i the response indicator such that $R_i=1(0)$ if sampled unit i responds on the outcome (does not respond). A possible way to deal with the nonresponse in such situations is by postulating a parametric model for the joint distribution of Y_i and R_i, given X_i. Little and Rubin (2002) distinguish between two ways of formulating the likelihood in this case. Suppressing for convenience the parameters from the notation,

Selection Models specify,

$$f(Y_i,R_i\mid X_i)=\Pr(R_i\mid Y_i,X_i)f_S(Y_i\mid X_i)\ , \tag{2}$$

where $\Pr(R_i\mid Y_i,X_i)$ models the response process. The missing sample values can be imputed in this case by the expectations, $E_{R^c}(Y_i\mid X_i)=E(Y_i\mid X_i,R_i=0)$ or by drawing at random from the *pdf* $f_{R^c}(Y_i\mid X_i)$ $=f(Y_i\mid X_i,R_i=0)$, thus accounting for the variability of the outcomes around their expectations. In practice, the probabilities and densities are replaced by their estimates, obtained by substituting the unknown parameters by their sample estimates. An example of the use of selection models is considered by Greenlees *et al*. (1982). The authors assume that the sample model is normal and the probability to respond is logistic.

Selection models allow estimating all the unknown model parameters, but as noted by Little (1994), they are based inevitably on strong distributional assumptions. Beaumont (2000) proposes to robustify the model considered by Greenlees *et al*. (1982) by dropping the normality assumption for the regression residuals. A drawback of this method is that the probabilities $P(R_i=0\mid X_i)$ appearing in the full likelihood for the responding and nonresponding units cannot actually be calculated, since the sample *pdf* of $Y_i\mid x_i$ is not specified. (For the nonresponding units the only known information is $R_i=0$). The author deals with this problem by expanding $P(R_i=1\mid Y_i,X_i)$ around the mean $E_S(Y_i\mid X_i)$, where $E_S(\cdot)$ is the mean under the sample model, but this amounts to assuming a MAR response. Note also that without further assumptions, the missing outcomes have to be imputed under this approach by the estimated expectations $\hat{E}_S(Y_i\mid x_i)$, instead of the expectations $\hat{E}_{R^c}(Y_i\mid X_i)$.

Pattern-mixture models specify,

$$f(Y_i, R_i \mid X_i) = f(Y_i \mid X_i, R_i)\Pr(R_i \mid X_i), \quad (3)$$

where $f(Y_i \mid X_i, R_i)$ defines the *pdf* under the different patterns of the missing data, $(R_i = 0, R_i = 1)$, and $\Pr(R_i \mid X_i)$ models the response probability given the covariates. A major drawback of pattern-mixture models is that the model holding for the nonrespondents, $f(Y_i \mid X_i, R_i = 0)$, cannot be extracted from the models $f(Y_i \mid X_i, R_i = 1)$ and $\Pr(R_i \mid X_i)$ fitted under this approach.

Tang *et al*. (2003) propose a 'pseudo-likelihood' approach that uses the conditional *pdf*, $f_S(X_i \mid Y_i)$ for the respondents. Application of this approach requires specification of the sample *pdf* $f_S(Y_i \mid X_i)$ and the marginal *pdf* $g_S(X_i)$. The method does not require a parametric model for the response probability but it assumes that it depends only on the outcome. The use of this approach does not enable imputing the missing outcomes from the distribution $f_{R^c}(Y_i \mid X_i) = f(Y_i \mid X_i, R_i = 0)$.

So far we considered methods applicable for the case where the covariates are observed for all the sampled units. Qin *et al*. (2002) propose a method that can be applied when the covariates are only known for the respondents. The method assumes a parametric model for $\Pr(R_i = 1 \mid X_i, Y_i)$ and known population means of the covariates. The authors use an empirical likelihood, addressing the problem of missing covariate information by using the unconditional response probability, $\lambda = \Pr(R_i = 1)$ in the likelihood, instead of the conditional probabilities $\Pr(R_i = 1 \mid X_i)$. The method accounts for the known population means of the covariates by adding constraints to the likelihood. However, our experience so far shows that the good performance of this procedure depends on having sufficient accurate initial values for the response model parameters and the Lagrange multipliers used for the constrained maximization procedure.

Chang and Kott (2008) propose an approach for estimating the response probabilities based on the known totals of calibration variables. The authors assume a parametric response model that can depend on the outcome value, and estimate the unknown parameters governing this model by regressing the Horvitz-Thompson (H-T, 1952) estimators of the totals of the calibration variables, with the response probabilities defined by their values under the model, against the corresponding known totals. See Remark 5 in Section 3 below. Having estimated the response probabilities, the use of this approach allows estimating the population totals of the target variables of interest, but it does not allow imputation of the missing outcomes, since no model is assumed for the outcome values.

3. The Respondents Distribution and Parameters Estimation

3.1 The respondents distribution and its relationship to the sample distribution

The *marginal pdf* of the outcome for a responding unit is obtained, similarly to Pfeffermann *et al*. (1998) as,

$$f_R(Y_i \mid X_i) = f(Y_i \mid X_i, i \in S, R_i = 1) = \frac{\Pr(R_i = 1 \mid Y_i, X_i, i \in S)}{\Pr(R_i = 1 \mid X_i, i \in S)} f_S(Y_i \mid X_i), \quad (4)$$

where $\Pr(R_i = 1 \mid X_i, i \in S) = \int \Pr(R_i = 1 \mid Y_i, X_i, i \in S) f_S(Y_i \mid X_i) dY_i$ and $f_S(Y_i \mid X_i)$ is the sample *pdf* under complete response. (As noted before, in this research we assume that the sample *pdf* and the population *pdf* are the same.) Denote $\pi(Y_i, X_i) = \Pr(R_i = 1 \mid Y_i, X_i, i \in S)$ and $\pi(X_i) = \Pr(R_i = 1 \mid X_i, i \in S)$.

Remark 1. It follows from (4) that the marginal *pdf* for a responding unit and the corresponding marginal sample *pdf* are the same if $\pi(Y_i, X_i) = \pi(X_i)$, that is, if the response probability does not depend on the outcome value given the covariates.

Remark 2. As with selection models, the use of the respondents' model requires modeling the sample *pdf*, $f_S(Y_i \mid X_i)$ and the response probability, $\Pr(R_i = 1 \mid Y_i, X_i, i \in S)$. Notice, however, that the resulting respondents' model can be tested since it relates to the data observed for the responding units.

By (4), if the sample outcomes and the response are independent between the units, and the covariates are only known for the respondents, one can estimate the parameters θ governing the sample model and the parameters γ governing the model for the response probabilities by maximizing the respondents' likelihood,

$$L_{\text{Resp}} = \prod_{i=1}^{r} f(Y_i \mid X_i, R_i = 1, i \in S; \theta, \gamma) = \prod_{i=1}^{r} \frac{\Pr(R_i = 1 \mid Y_i, X_i, i \in S; \gamma) f_S(Y_i \mid X_i; \theta)}{\Pr(R_i = 1 \mid X_i, i \in S; \theta, \gamma)}. \quad (5)$$

The notable property of the likelihood (5) is that it does not require knowledge of the covariates for nonresponding units, or modeling the distribution of the sampled covariates.

3.2 Calibration constraints

The additional information contained in the population size and the population totals of some or all of the covariates is not part of the likelihood in (5). We utilize this information by imposing the following constraints. Suppose that the population totals, $X^{pop} = (X_1{}^{pop},...,X_K{}^{pop})$ of all the covariates $X = (X_1,...,X_K)$ are known. The calibration constraints are,

$$\sum_{i=1}^{r} w_i \frac{X_{ki}}{\pi(Y_i, X_i;\gamma)} = X_k^{pop}, k = 1,...,K;\ \sum_{i=1}^{r} w_i \frac{1}{\pi(Y_i, X_i;\gamma)} = N\ , \tag{6a}$$

where $\{w_i = (1/\pi_i) = 1/\Pr(i \in S)\}$ are the base sampling weights.

When the response model has an intercept, we use the additional constraint,

$$\sum_{i=1}^{r} w_i \frac{(Y_i - E_S(Y_i \mid X_i;\theta))}{\pi(Y_i, X_i;\gamma)} = 0\,. \tag{6b}$$

Our experience so far shows that often the response model does not contain all the covariates included in the sample model, in which case the estimation of the response model parameters is enhanced by replacing the constraint (6b) by the following constraint. Let $X_i = (X_i^{(1)}, X_i^{(2)})$, where $X_i^{(1)} = (X_{i1},...,X_{im})$, $X_i^{(2)} = (X_{i,m+1},...,X_{iK})$. Suppose that $E_S(Y_i \mid X_i;\theta) = \sum_{k=0}^{p} \theta_k X_{ki} = \theta^{(1)\prime} X_i^{(1)} + \theta^{(2)\prime} X_i^{(2)}$ (e.g., the sample model is normal), but $\pi(Y_i, X_i;\gamma) = \pi(Y_i, X_i^{(1)};\gamma)$. Let $X^{pop} = (X^{pop,(1)}, X^{pop,(2)})$. The alternative constraint (together with (6a) but using only the variables in $X^{(1)}$) is,

$$\sum_{i=1}^{r} w_i \frac{\theta^{(2)\prime} X_i^{(2)}}{\pi(Y_i, X_i^{(1)};\gamma)} = \theta^{(2)\prime} X^{pop,(2)}\ . \tag{6c}$$

Notice that the constraints (6a) and (6c) imply, $\sum_{i=1}^{r} \frac{E_S(Y_i \mid X_i^{(1)}, X_i^{(2)})}{\pi(Y_i, X_i^{(1)};\gamma)} = \sum_{i=1}^{n} E_S(Y_i \mid X_i^{(1)}, X_i^{(2)})$.

Remark 3. The left hand sides of (6a) and (6c) are the familiar H-T estimators of the corresponding totals under the following two-phase sampling process: in the first phase a sample S of size n is sampled with inclusion probabilities $\Pr(i \in S) = \pi_i = 1/w_i$; in the second phase the sampled units respond with probabilities $\pi(Y_i, X_i) = \Pr(R_i = 1 \mid Y_i, X_i, i \in S)$. We assume that the weights w_i are known, which is usually the case.

Remark 4. Instead of the constraints (6a), one could use the following constrains:

$$\sum_{i=1}^{r} w_i \frac{X_{ki}}{\pi(X_i;\theta,\gamma)} = X_k^{pop}, k = 1,...,K\ , \quad \sum_{i=1}^{r} w_i \frac{1}{\pi(X_i;\theta,\gamma)} = N\ . \tag{7}$$

A theoretical argument in favor of (7) is that the probabilities $\pi(X_i;\theta,\gamma) = \int \pi(Y_i, X_i;\gamma) f(Y_i \mid X_i;\theta) dY_i$ account for the net effect of X_i on the response, thus yielding less variable H-T estimators for the covariates totals than the use of the probabilities $\pi(Y_i, X_i;\gamma)$, which account also for the effect of the outcome Y_i. Note that these probabilities depend also on the vector parameter θ, thus providing additional information for it.

3.3 Estimation Algorithm with calibration constraints

In order to utilize the additional information provided by knowledge of the population means, we estimate the response model parameters by solving the equations (6), or the equations (6b) or (6c) and (7), and use the following iterative algorithm.

Let $\hat{\theta}^{(0)}$ denote initial values for the vector θ indexing the sample *pdf* $f_S(Y_i \mid X_i;\theta)$.

Step j: For given estimate $\hat{\theta}^{(j)}$ from iteration j, set $\theta = \hat{\theta}^{(j)}$ and solve the set of equations (6), or (6b) or (6c) and (7) as a function of the unknown parameters γ governing the model $\pi(Y_i, X_i;\gamma)$ for response probabilities. This step yields estimates $\hat{\gamma}^{(j+1)}$.

Step j+1: Maximize (5) with respect to θ, with γ equal to $\hat{\gamma}^{(j+1)}$. This step yields new estimate $\hat{\theta}^{(j+1)}$. Continue the iterations until convergence.

We find it convenient to estimate the sample model parameters θ by maximizing the likelihood (5) with 'fixed' parameters γ, and estimate the response model parameters γ by solving the equations (6), or (6b) or (6c) and (7) with 'fixed' θ, since this simplifies the computation of the estimators.

Remark 5. Another possibility of utilizing known covariate totals is by applying an approach proposed by Chang and Kott (2008). By this approach the estimated totals of calibration variables $Z_1,...,Z_q$, which may contain some or all the model covariates are regressed against their known population totals. Thus, in the case that the probability to respond depends on the outcome variable and all the covariates, the method requires that $q \geq K+1$. The major difference between the algorithm described above and this method is that it allows utilizing more than $(K+1)$ known population totals, resulting in more equations than estimated parameters and hence possibly more stable estimators. The authors estimate the unknown parameters by setting the nonlinear regression equations, $\sum_{i=1}^{r} w_i \frac{Z_i}{\pi(Y_i, X_i; \gamma)} = Z^{pop} + \varepsilon$, where Z^{pop} denotes the vector of known population totals of the calibration variables, and ε is a vector of errors. See Chang and Kott (2008) for details.

4. Imputation of missing values and estimation of population totals

Denote by,

$$\begin{aligned} &\hat{f}_S(Y_i \mid X_i) = f_S(Y_i \mid X_i; \hat{\theta}, \hat{\gamma}),\ \hat{\pi}(Y_i, X_i) = \pi(Y_i, X_i; \hat{\gamma}) \\ &\hat{\pi}(X_i) = \pi(X_i; \hat{\theta}, \hat{\gamma}),\ \hat{E}_S(Y_i \mid X_i) = E_S(Y_i \mid X_i; \hat{\theta}, \hat{\gamma}) \end{aligned}, \tag{8}$$

the estimated sample *pdf*, the response probabilities and the estimated expectations, with the estimates of the parameters obtained by one of the methods described in Section 3. The estimates in (8) provide several possibilities for the imputation of the missing values and the estimation of the population total of the outcome variable.

When the covariates for the nonrespondents are unknown, the population total of the outcome can be estimated using the (pseudo) H-T estimator,

$$\hat{Y}_{(1)} = \sum_{i=1}^{r} w_i Y_i / \hat{\pi}(Y_i, X_i)\,. \tag{9}$$

Alternatively, one can use the estimator,

$$\hat{Y}_{(2)} = \sum_{i=1}^{r} w_i \frac{\hat{E}_S(Y_i \mid X_i)}{\hat{\pi}(X_i)}, \tag{10}$$

which uses the response weights that only condition on the covariates.
The estimators (9) and (10) only require knowledge of the covariates for the responding units. If the covariates are known for all the sampled units, another set of plausible estimates is obtained as,

$$\hat{Y}_{(3)} = \sum_{i=1}^{n} w_i Y_i^*\ ;\ \ Y_i^* = Y_i \text{ if } i \in R\ ,\ Y_i^* = Y_i^{imp} \text{ if } i \in R^c\,. \tag{11}$$

The imputed values, Y_i^{imp}, can be computed either as,

$$Y_i^{imp} = E_{R^c}(Y_i \mid X_i) = E(Y_i \mid X_i, i \in R^c)\,, \tag{12}$$

or by generating at random one or more observations from the *pdf* $f_{R^c}(Y_i \mid X_i)$ and taking the average of these observations as the imputed value, using multiple imputation techniques. (Rubin, 1987, Schafer and Schenker, 2000). The *pdf* $f_{R^c}(Y_i \mid X_i)$ is the *pdf* for a *nonresponding* unit with covariates X_i. We emphasize that under NMAR nonresponse the imputations of the missing outcomes should be based on the model holding for the nonrespondents and not on the sample model that assumes full response or the model holding for the respondents. Following Sverchkov and Pfeffermann (2004), the *pdf* for a nonresponding unit can be computed utilizing the relationship,

$$f_{R^c}(Y_i \mid X_i) = \frac{\Pr(R_i = 0 \mid Y_i, X_i, i \in S) f_S(Y_i \mid X_i)}{\Pr(R_i = 0 \mid X_i, i \in S)} = \frac{[1 - \pi(Y_i, X_i)] f_S(Y_i \mid X_i)}{[1 - \pi(X_i)]}\,. \tag{13}$$

In practice, one has to use the estimated *pdf*, $\hat{f}_{R^c}(Y_i \mid X_i) = \frac{[1 - \hat{\pi}(Y_i, X_i)] \hat{f}_S(Y_i \mid X_i)}{[1 - \hat{\pi}(X_i)]}$, as obtained by replacing the unknown parameters by their sample estimates.
The predictor $\hat{Y}_{(3)}$ in (11) assumes that the covariates are known for every unit in the sample. When the covariates are only known for the respondents, we may first predict the missing covariates for the nonrespondents from the

probability function $P_{X|0}(x_i) = \Pr(X_i = x_i \mid R_i = 0, i \in S)$, and then predict the outcome value as described above. By Sverchkov and Pfeffermann (2004), the latter probability function can be expressed as,

$$\begin{aligned}\Pr(X_i = x_i \mid R_i = 0, i \in S) &= \frac{P(R_i = 0 \mid X_i = x_i, i \in S)}{P(R_i = 0 \mid i \in S)} \Pr(X_i = x_i \mid i \in S) \\ &= \frac{P(R_i = 0 \mid X_i = x_i, i \in S) \Pr(X_i = x_i \mid R_i = 1, i \in S) \Pr(R_i = 1 \mid i \in S)}{P(R_i = 0 \mid i \in S) \Pr(R_i = 1 \mid X_i = x_i, i \in S)}\end{aligned} \quad . \tag{14}$$

The use of (14) requires estimating the probability $\Pr(X_i = x_i \mid R_i = 1, i \in S)$. Fitting a parametric model with a large number of covariates is practically formidable, and we use instead the empirical probability $\Pr(X_i = x_i \mid R_i = 1, i \in S) = (1/r) \quad \forall x_i \in R$, (equal probability for each vector covariate observed for the responding units). The probability $P_{X|0}(x_i)$ can be estimated then as,

$$\hat{P}_{X|0}(x_i) = \hat{\Pr}(X_i = x_i \mid R_i = 0, i \in S) = \frac{[1 - \hat{\pi}(x_i)]}{\hat{\pi}(x_i)[\sum_{j=1}^{r}(1/\hat{\pi}(x_j)) - r]}, \; x_i \in R. \tag{15}$$

The estimator $\hat{P}_{X|0}(x_i)$ in (15) is obtained from (14) by estimating $\hat{\Pr}(R_i = 1 \mid X_i = x_i, i \in S) = \hat{\pi}(x_i) = \pi(X_i; \hat{\theta}, \hat{\gamma})$ and $\Pr(R_i = 1 \mid i \in S) = \frac{r}{\sum_{j=1}^{r}(1/\hat{\pi}(x_i))}$, guaranteeing $\sum_{X_i} \hat{\Pr}(X_i = x_i \mid R_i = 0, i \in S) = 1$. Note that the estimator (15) assumes that the plausible covariates in the subsample of the nonrespondents is the same as in the subsample of the respondents.

5. Empirical Study

5.1 Study population and outcome variable

In this section we study the performance of the proposed approach in imputing the missing covariates and outcomes and in estimating the mean population outcome, using a real data set. We compare the results with results obtained when ignoring the response process, assuming that the subsample of respondents is a simple random sample from the original sample, and with results obtained by some of the other approaches proposed in the literature to handle NMAR nonresponse, reviewed in Section 2.

The data used for this study was collected as part of the Household Expenditure Survey carried out by the Israel Central Bureau of Statistics in 2005. The survey collects information on socio-demographic characteristics of each member of the selected Households (HHs), as well as information on the HH income and expenditure. The initial response rate in this survey was 43%, but after many recalls the response rate went up to 90% of the sampled HHs. The HHs were sampled with equal probabilities.

Altogether, the sample consists of 7800 HHs. There were 802 dwellings which did not meet the investigation criteria (vacant, the occupants have another permanent address, etc.); 678 HHs were not investigated due to nonresponse (in all the recalls) and 49 HHs were disqualified at the editing stage, such that the final sample consists of 6271 HHs. For the empirical study of this paper we restrict to HHs where the head of the HH is an employee, aged 25-64 and born in Israel. We only consider HHs where at least one of its members worked during the three months preceding the interview. The head of the HH is the member with the highest income among the members of the HH who worked in these three months. The reduced data set consists of 1721 HHs. The target outcome variable is the *household income per standard person*. The number of standard persons in the HH is defined as follows:

Persons in HH	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Standard persons in HH	1.25	2.00	2.65	3.20	3.75	4.25	4.75	5.20	5.60	Y*

*Y=5.60+0.4×(No. of persons minus 9)

For the empirical study that follows we selected a single sample of respondents from the original sample of the 1721 HHs using a logistic model by which the probability to respond depends strongly on both the income and some of the covariates. The resulting number of respondents is $n_r = 729$.

5.2 Sample model and response probabilities

We assume that the sample distribution of the outcome (under full response) given the covariates is normal;

$$Y_i = X_i'\beta + \varepsilon_i,\ \ \varepsilon_i \sim N(0,\sigma_\varepsilon^2) \tag{16}$$

where Y_i is the log income per standard person in household i and $X_i = [1, X_{i1},..., X_{iK}]'$ is the corresponding vector of covariates. As established by Landsman (2008), the model holding for the responding units (Eq. 4) is identifiable in this case. The covariates considered for this study include characteristics of the head of the HH: Gender, Age, Age^2, No. of years at school and No. of monthly working hours, and characteristics of the HH: No. of earners , HH size and location of the HH. We fitted the sample model (Eq. 16) using all the sample data. The R^2 of the model is 0.60. The values of the regression coefficients are sensible. For example, the values of the coefficients of the education variables increase monotonically as the level of education increases. The number of earners in the household has a strong positive effect on the income, while the size of the household has a strong negative effect. The coefficient of Gender (being a female) is negative. Figure 1 compares the distribution of the estimated regression residuals with the normal distribution with mean zero and (estimated) standard deviation ($\hat{\sigma}_\varepsilon^2 = 0.403$).

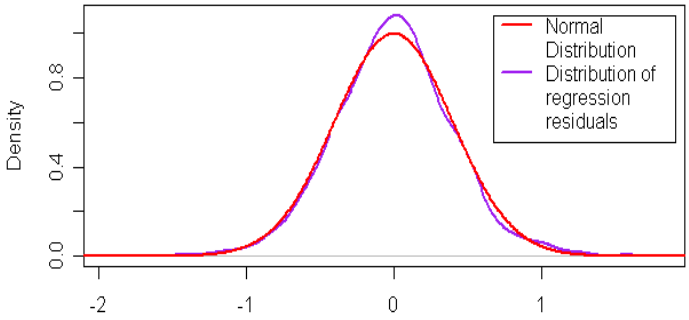

Figure 1: Distribution of regression residuals $\hat{\varepsilon}_i$ and normal distribution with the same variance.

The distribution of the residuals is seen to be close to the normal distribution although with somewhat shorter tails. The Kolmogorov-Smirnov test rejects the hypothesis of normality of the residuals but this test is not really valid in the present case since the test statistic uses the estimated residuals (based on the estimated coefficients) and the estimated residuals are not strictly independent.

We selected the respondents using the logistic model,

$$P(R_i = 1 \mid Y_i, X_i) = [1 + e^{-(\gamma_0 Y_i + X_i'\gamma_1)}]^{-1}. \tag{17}$$

The coefficients $\gamma' = (\gamma_0, \gamma_1')$ were selected so that the probability of response depends strongly on both the income and the covariates.

5.3 Methods considered

We consider the following methods:

1. The proposed method [Equations (5) (6a) and (6c)]. The results obtained when replacing the constraints (6a) by (7) are very similar in the present study and therefore are not shown.
2. Chang and Kott (2008) method described in Remark 5 of Section 3.3.
3. A combination of the proposed method with Chang and Kott (2008) method by which the parameters of the response model are estimated as in 2. In what follows we refer to this method as PCK.
4. Tang *et al.* (2003).
5. Beaumont (2000).

Remark 6. The last two methods require knowledge of the covariates for all the sample units, while the first 3 methods only require knowledge of the covariates for the responding units. The method by Tang et al. (2003) assumes that the probability to respond depends only on the outcome variable. This method does not permit estimation of the outcome distribution for nonresponding units, and hence the imputation of the missing outcomes can be carried out only by random draws from the sample distribution. Beaumont (2000) uses the sample distribution expectation for imputing the missing outcomes. We include the latter two methods in the present study in order to test their robustness to deviations from their underlying conditions.

5.4 Application of proposed approach and other approaches

We assess the performance of the various approaches by comparing the estimates of the response probabilities, the imputations of the missing incomes and the estimates of the sample mean of the income to their known values. The imputation of the missing incomes is carried out under two different scenarios: In scenario 1 we use the known covariates for the nonrespondents and impute the incomes by drawing at random from the estimated distribution $\hat{f}_{R^c}(Y_i \mid X_i) = \hat{f}(Y_i \mid X_i, R_i = 0; \hat{\beta}, \hat{\sigma}_\varepsilon^2, \hat{\gamma})$. In Scenario 2 the covariates for the nonresponding units are taken as unknown and the imputation of the missing incomes is carried out by first predicting the missing covariates using (15), and then imputing the incomes similarly to Scenario 1.

Table 1 shows the bias (BIAS), mean absolute error (MAE) and RMSE of the estimated response probabilities $\pi(Y_i, X_i; \hat{\gamma})$ when estimating the true probabilities $\pi(Y_i, X_i; \gamma)$. Denoting $d_i = [\pi(Y_i, X_i; \hat{\gamma}) - \pi(Y_i, X_i; \gamma)]$, the three measures are defined as: $BIAS = \frac{1}{n}\sum_{i=1}^{n} d_i$; $MAE = \frac{1}{n}\sum_{i=1}^{n} |d_i|$; $RMSE = [\frac{1}{n}\sum_{i=1}^{n} d_i^2]^{1/2}$.

Table 1: Bias, mean absolute error (MAE) and RMSE of estimates of response probabilities under different methods

Method	*BIAS*	*MAE*	*RMSE*
Proposed	-0.013	0.047	0.044
Chang & Kott	0.010	0.081	0.082
Beaumont	-0.020	0.064	0.061

Table 2 compares the percentiles of the estimators $\pi(Y_i, X_i; \hat{\gamma})$ to the percentiles of the true probabilities, $\pi(Y_i, X_i; \gamma)$.

Table 2: Percentiles of empirical distribution of true response probabilities and of empirical distribution of estimated response probabilities

Nominal levels	(0.05)	(0.1)	(0.25)	(0.5)	(0.75)	(0.90)	(0.95)
True percentiles	0.13	0.17	0.26	0.40	0.58	0.71	0.78
Proposed	0.13	0.17	0.25	0.40	0.56	0.70	0.77
Chang & Kott	0.12	0.16	0.27	0.41	0.58	0.73	0.79
Beaumont	0.13	0.16	0.25	0.39	0.55	0.69	0.77

The proposed method is seen to perform much better than the other two methods in terms of the MAE and RMSE measures. The percentiles of the empirical distribution of the estimated response probabilities are close to the percentiles of the empirical distribution of the true response probabilities under all the three methods.

Next we study the performance of the various approaches in imputing the missing outcomes. For the proposed method, PCK and when ignoring the nonresponse we distinguish between the case where the covariates for the nonrespondents are known (full covariate information) and the case where the covariates are only observed for the responding units (missing covariate information). For the methods of Beaumont (2000) and Tang *et al.* (2003) we only consider the case of full covariate information as assumed by these methods. In order to compare the various methods, we generated 300 imputations for each nonresponding unit, and calculated for each of the 300 sets the imputation bias (BIAS), mean absolute error (MAE) and RMSE of the imputed values. Let Y_i^* denote the imputed value for nonresponding unit i as obtained under a given imputation method. The three measures are computed as: $BIAS = \frac{1}{n-r}\sum_{i=1}^{n-r}(Y_i^* - Y_i)$, $MAE = \frac{1}{n-r}\sum_{i=1}^{n-r} |Y_i^* - Y_i|$, $RMSE = [\frac{1}{n-r}\sum_{i=1}^{n-r}(Y_i^* - Y_i)^2]^{1/2}$. Table 3 presents the means of the BIAS, MAE and RMSE over the 300 sets of imputed values.

Table 3: Imputation bias, mean absolute error (MAE) and RMSE over 300 imputations

	Full Covariates Information			*Missing Covariate Information*		
Method	BIAS	MAE	RMSE	BIAS	MAE	RMSE
Proposed	58.57	3750.39	5653.63	134.71	5552.02	7941.52
PCK	296.22	3823.17	5754.87	428.55	5637.91	8064.12
Tang *et al.*	-1085.41	3578.52	5462.94	-	-	-
Beaumont	-825.78	2590.81	4304.47	-	-	-
Ignorable	-1081.50	3581.20	5462.17	-2588.25	5087.77	7267.71

Table 3 indicates that the proposed method yields much smaller biases than the other methods. The use of PCK also yields small biases, but the biases obtained under the methods of Tang *et al*. and Beaumont, which as stated in Remark 6 are not really applicable under the present model are large, as is also the case when ignoring the nonresponse. Notice, however, that the MAE and RMSE measures are actually smaller for the last three methods than for the first two methods, indicating a smaller variability of the imputations.

Figures 2-4 compare the true empirical cumulative distribution of the incomes of the nonresponding units with the means of the estimated empirical distributions over the 300 imputation sets.

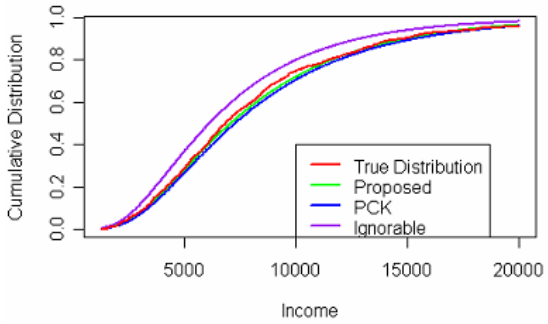

Figure 2: True empirical cumulative distribution and means of estimated empirical cumulative distributions of the incomes over 300 imputation sets. Full covariate information.

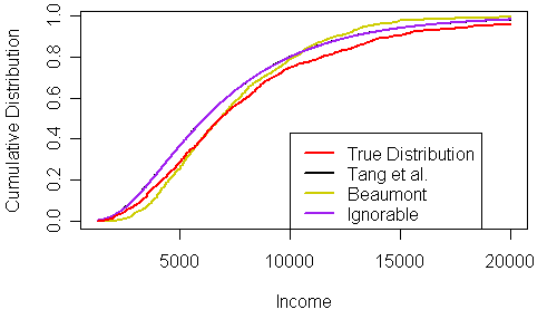

Figure 3: True empirical cumulative distribution and means of estimated empirical cumulative distributions of the incomes over 300 imputation sets. Full covariate information.

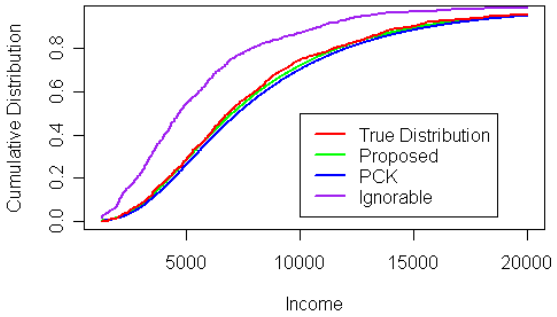

Figure 4: True empirical cumulative distribution and means of estimated empirical cumulative distributions of the incomes over 300 imputation sets. Missing covariate information.

Figures 2-4 illustrate that the use of the methods of Tang *et al*. and Beaumont yield bad estimators for the distribution of the incomes in the subsample of nonresponding units. The same is true of course when ignoring the nonresponse. (The graph for Tang *et al*. in Figure 3 is indistinguishable from the graph obtained when ignoring the nonresponse.) On the other hand, the use of the proposed method or the combination of the proposed method and Chang and Kott yield satisfactory estimators, even when imputing the missing covariates.

The fact that ignoring the nonresponse yields biased estimates for the income distribution is obvious. By ignoring the nonresponse, it is assumed that the distributions of the covariates for the responding and nonresponding units and the corresponding income distributions are the same in the two subsamples, which is not the case. Notice that even if the distribution of the income given the covariates was the same for the responding and nonresponding units, ignoring the nonresponse in the case of missing covariate information would still produce biased estimates for the income distribution, unless the distribution of the covariates is also the same in the two subsamples. The result that Beaumont's method yields biased estimates for the income distribution in the case of full covariate information can be explained by the fact that the imputations obtained by this method have two sources of bias. First, the method imputes the log-income from the estimated sample expectation $E_S(Y_i \mid X_i)$, instead of using the expectation holding for the nonresponding units. Second, in our application we actually impute e^{Y_i}, which requires estimating $E_s(e^{Y_i} \mid X_i)$ under the sample distribution of $Y_i \mid X_i$, but the sample distribution is not specified under this approach. We therefore used the approximation, $\hat{E}_S(e^{Y_i} \mid X_i) = e^{\hat{E}_S(Y_i \mid X_i)}$, which is wrong. Regarding the method of Tang *et al*., we mentioned before that for the application of this method we imputed the missing outcomes in the case of full covariate information by use of the sample distribution. However, since the probability to respond depends strongly on the income given the other covariates in the response model, the sample model of $Y_i \mid X_i$ and the model holding for the nonresponding units are different, which results in large imputation bias.

Very often, the main purpose of adjusting for nonresponse is to reduce the bias in estimating the population mean of the outcome. In what follows we actually consider estimating the true sample mean, which we know. (When sampling with equal probabilities as in the present case, the true sample mean is randomization unbiased for the true population mean.) Table 4 shows the percent error when estimating the mean sample income by the estimators $\hat{\bar{Y}}_{(1)}$ and $\hat{\bar{Y}}_{(2)}$ defined by (9) and (10), and the percent relative bias (PRB) and percent relative RMSE (PRRMSE) of the estimator $\hat{\bar{Y}}_{(3)}$ defined by (11). The estimator $\hat{\bar{Y}}_{(3)}$ was computed for each of the 300 imputation sets. The two measures were calculated as, $PRB(\hat{\bar{Y}}_{(3)}) = \frac{1}{300}\sum_{l=1}^{300}\frac{(\hat{\bar{Y}}_{(3),l} - \bar{Y})}{\bar{Y}} \times 100$, $RRRMSE(\hat{\bar{Y}}_{(3)}) = [\frac{1}{300}\sum_{l=1}^{300}(\frac{\hat{\bar{Y}}_{(3),l} - \bar{Y}}{\bar{Y}})^2]^{1/2} \times 100$, where $\hat{\bar{Y}}_{(3),l}$ denotes the estimator calculated from imputation set l and $\bar{Y}$ is the true sample mean.

Table 4: Percent errors of $\hat{\bar{Y}}_{(1)}$ and $\hat{\bar{Y}}_{(2)}$, and percent relative Bias and percent relative RMSE (in parentheses) of $\hat{\bar{Y}}_{(3)}$ when estimating the mean sample income.

Method	$\hat{\bar{Y}}_{(1)}$	$\hat{\bar{Y}}_{(2)}$	$\hat{\bar{Y}}_{(3)}$	
			Full covariate information	Missing covariate information
Proposed	0.93	-8.73	0.46 (1.08)	1.09 (1.82)
PCK	2.05	-9.24	2.27 (2.46)	3.53 (3.86)
Chang & Kott	2.05	-	-	-
Tang et al.	-	-	-8.81 (8.84)	-
Beaumont	4.28	-	-6.70 (6.70)	-
Ignorable	-21.13	-	-8.78 (8.82)	-21.16 (21.18)

The results in Table 4 show very good performance of the estimators $\hat{\bar{Y}}_{(1)}$ and $\hat{\bar{Y}}_{(3)}$ when using the proposed method and PCK, but the first method performs better. The estimator $\hat{\bar{Y}}_{(1)}$ under Beaumont's approach also has a small error but the estimator $\hat{\bar{Y}}_{(3)}$ has a large relative bias under this method for reasons discussed above. The estimator $\hat{\bar{Y}}_{(1)}$ using the estimated response probabilities computed by the method of Chang and Kott is likewise very accurate. Finally, it is interesting to note that the estimator $\hat{\bar{Y}}_{(2)}$ has a large percentage error in the present case. This result is possibly explained by the fact that $\hat{\bar{Y}}_{(2)}$ uses the estimated response probabilities $\hat{\pi}(X_i) = \pi(X_i; \hat{\beta}, \hat{\sigma}_\varepsilon^2, \hat{\gamma})$, which are less accurate than the estimates $\hat{\pi}(Y_i, X_i) = \pi(Y_i, X_i; \hat{\gamma})$ used for constructing the estimator $\hat{\bar{Y}}_{(1)}$.

6. Conclusions

In this article we develop a general approach for estimation and imputation when the nonresponse is not missing at random (NMAR). By modeling the sample model under full response and the response process, we are able to estimate the distribution of the outcome for the nonresponding units given the corresponding covariates. When the covariates for the nonresponding units are known, we use this distribution for the imputation of the missing values. Otherwise, we impute the missing covariates as well, again accounting for the response process. Estimating the response probabilities allows also estimating population means using Horvitz-Thompson (1952) type estimators. We study the performance of our approach using a real data set that has many missing values after the first interview, but which are later obtained on subsequent interviews.

The proposed approach is model-dependent and its good performance depends on correct specification of the population model and the response process. For any given specification, the goodness of fit of the resulting model holding for the responding units can be assessed by use of classical goodness of fit testing procedures, since the later model relates to the observed data.

There are still outstanding issues that require further investigation before the approach can be recommended for practical applications. We mention in particular variance estimation of the proposed estimators of the finite population totals, extension of the method to the case where some of the covariates are fully observed but other covariates are only observed for the respondents, establishing the consistency of the parameter estimators under the proposed two-step estimation procedure, the development of new model goodness of fit test procedures and studying the robustness of the proposed approach to possible model misspecification.

Acknowledgements

This research is supported by a grant from the United Sates-Israel Binational Science Foundation (BSF). The authors thank the Israel Central Bureau of Statistics for preparing and providing the data used for the empirical study.

References

Beaumont, J.F. (2000). An estimation method for nonignorable nonresponse. *Survey Methodology*, **26**, 131-136.

Chang, T. and P. S. Kott (2008). Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model. *Biometrika*, forthcoming.

Greenlees, J.S., Reece, W.S. and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, **77**, 251-261.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663-685.

Landsman, V. (2008). Estimation of treatment effects in observational studies by fitting models generating the Sample Data. PHD Dissertation submitted to the Hebrew University of Jerusalem, Israel.

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

Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125-134.

Little, R.J.A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471-483.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical analysis with missing data*. John Wiley & Sons, New York; Chichester.

Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric Distributions of Complex Survey Data Under Informative Probability Sampling'. *Statistica Sinica*, **8**, 1087-1114.

Pfeffermann, D. and Sverchkov, M. (2003). Fitting Generalized Linear Models Under informative Sampling. In, *Analysis of survey Data*, Eds. C. Skinner and R. Chambers, New York: Wiley, 175-195.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63** , 581-590.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons. New York: Chichester.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with Survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association*, **97**, 193-200.

Qin, J., Shao, J. and Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing response. *Journal of the American Statistical Association*, **103**, 797-810.

Särndal C.E. and Swensson B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

Schafer, J.L. (1997). *Analysis of incomplete Multivariate Dat*a. London: Chapmanand Hall.

Sverchkov, M. and Pfeffermann, D. (2004). Prediction of finite Population Totals Based on the Sample Distribution'. *Survey Methodology*, 30, 79-92.

Tang T., Little, R.J.A. and Raghunathan, T.E. (2003). Analysis of multivariate missing data with Nonignorable nonresponse. *Biometrica, **90**, 747.*