Imputation and Estimation under NMAR Nonresponse With Limited Covariate Information, Revisited

Danny Pfeffermann[1], Anna Sikov[2]

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

[2] Department of Statistics, Hebrew University, Jerusalem, 91905, Israel.

Abstract

In this research we develop and apply new methods for handling not missing at random (NMAR) 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, which can be tested. Our methods utilize information on the population means of some or all the auxiliary variables in the models. In a JSM paper last year we developed an algorithm for estimating the parameters governing the two models. We also showed how to estimate the distributions of the missing covariates and outcomes, and used them for imputing the missing values for the nonresponding units and for estimating population means. In this paper we outline conditions for the convergence of the proposed algorithm and study the properties of the estimators obtained from its application. We consider different approaches of estimating the variance of estimators of population means and propose a test statistic for selecting covariates to the model of the response probabilities. The new developments are illustrated using simulated data and a real data set collected as part of the Household Expenditure Survey carried out by the Israel Central Bureau of Statistics in 2005.

Key Words: Bootstrap, Calibration, Horvitz-Thompson estimator, Multiple imputation, Nonrespondents distribution, Respondents distribution.

1. Introduction

Most of the methods dealing with nonresponse assume either explicitly or implicitly that the missing values are “missing at random” (MAR), and that the auxiliary (explanatory) variables are observed for both the respondents and the nonrespondents. These assumptions, however, are not always met in practice. In this paper we consider the often practical situation where the probability to respond depends on the outcome value, solely or in addition to depending on the model explanatory variables. For example, the probability to observe income may depend on the income level, as well as on 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 mostly consider the case of ‘unit nonresponse’, where the auxiliary (covariate) information for the nonrespondents is likewise unobserved, except for the population

totals of some or all of these variables. The totals of the covariates are often available from administrative or census records.

In a JSM presentation last year (Hereafter P-S (2008)), we proposed a new approach for handling NMAR nonresponse, which does not require knowledge of the auxiliary variables (covariates) for the nonrespondents. We illustrated the effectiveness of the proposed method in estimating the unknown model parameters and the imputation of the missing covariates and outcomes, and consequently in estimating the population mean of the target outcome. This is achieved by deriving the model holding for the outcomes of the responding units as a combination of a model assumed for the outcome variable under complete response (the 'sample model'), and a model assumed for the response probabilities. The resulting 'respondents model' defines the likelihood for the observed outcomes. In order to utilize the additional information provided by the population totals of the covariates, we add calibration constraints, which match pseudo probability weighted estimates of the totals of the covariates with their known population values. The weights used for these estimates are the inverse of the postulated response probabilities. The unknown model parameters are then estimated by an iterative algorithm which maximizes the likelihood with respect to the parameters governing the sample model, and solves the calibration constraints with respect to the parameters of the response probabilities. We repeat the details of the algorithm in Section 3 and then outline the conditions for its convergence. We also discuss the properties of the resulting estimators.

Having estimated the parameters of the model for the response probabilities, we predict the population mean of the outcome values by use of Horvitz-Thompson (H-T, 1952)) type estimators, utilizing the estimated response probabilities. Alternatively, when the covariates are observed for all the sampled units, we can estimate the conditional distribution of the outcome values for the non responding units given their respective covariates, and then use this distribution for imputing the missing outcomes. Combining the observed and imputed values provides another predictor of the outcome population mean. In the case of missing covariate information, this can be done by first imputing the missing covariate values. In the present paper we describe parametric and resampling methods for estimating the variances of the proposed estimators of the population mean.

We develop a new test statistic for testing whether the covariates included in the model for the response probabilities together with the outcome variable are sufficient for the computation of these probabilities. The test uses the estimated probabilities for estimating the totals of other covariates not yet included in the model by the H-T estimator, and compares these estimates with their known totals (when available). When estimating the standard deviation of the estimators of the covariates' totals we account for the calibration constraints used for estimating the parameters governing the model of the response probabilities.

We illustrate our methods using data collected as part of the Household Expenditure Survey (HES) carried out by the Israel Central Bureau of Statistics in 2005.

2. Existing Approaches

In this section we repeat the literature review of P-S (2008), so as to make the paper self contained. See the previous paper for the performance of some of the approaches described below (not repeated in this article). 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*), $f_U(Y_i \mid X_i;\theta)$, governed by the 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 the outcomes and possibly the covariates are unobserved. The response process is assumed to be independent between units. The observed sample of respondents can be viewed therefore as the outcome 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; Rubin, 1976, 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 complete 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 paper 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 unit in the sample. 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 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) consider two ways of formulating the joint distribution in this case, where we suppress 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 probability. 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* $f_{R^C}(Y_i \mid X_i) = f(Y_i \mid X_i, R_i = 0)$, accounting this way 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 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 use of the *pdf* $f_S(Y_i \mid X_i)$, instead of the *pdf* $f_{R^C}(Y_i \mid X_i)$.

Pattern-mixture models specify,

$$f(Y_i, R_i \mid X_i) = f(Y_i \mid X_i, R_i) \Pr(R_i \mid X_i), \tag{3}$$

where $f(Y_i \mid X_i, R_i)$ defines the *pdf* for the respondents $(R_i = 1)$ and the nonrespondents $(R_i = 0)$, and $\Pr(R_i \mid X_i)$ models the response probability given the covariates (the 'propensity scores'). 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 $\Pr(R_i = 1 \mid Y_i, X_i)$ 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 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 model for the response probabilities that can depend on the outcome value and estimate the parameters governing this model by regressing the H-T estimators of the totals of the calibration variables against the corresponding known totals, with the response probabilities defined by their values under the model. Having estimated the response probabilities, the use of this approach allows estimating the population totals of target variables of interest, but it does not allow imputing the missing outcomes, since no model is assumed for the outcome values.

3. The Respondents Distribution and Parameter 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,

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

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, 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. As with selection models, the use of the respondents' *pdf* requires modeling the sample *pdf*, $f_S(Y_i \mid X_i)$ and the response probability, $\pi(Y_i, X_i)$. 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 indexing the distribution under complete response and the parameters γ indexing the response probabilities by maximizing the respondents' likelihood,

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

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

In what follows we assume knowledge of the population size, N, and the totals $X^{pop} = (X_1^{pop}, ..., X_K^{pop})$ of the covariates contained in the model for the responding units. This additional information is not part of the likelihood in (5). Our experience so far shows that the response model does not contain all the covariates included in the sample model. Let $X = (X_1, ..., X_K) = (X^{(1)}, X^{(2)})$, where $X^{(1)} = (X_1, ..., X_m)$ and $X^{(2)} = (X_{m+1}, ..., X_K)$, and suppose that $\pi(Y_i, X_i; \gamma) = \Pr(R_i = 1 \mid Y_i, X_i; \gamma) = \pi(Y_i, X_i^{(1)}; \gamma)$. We assume that the sample model belongs to the family of generalized linear model (GLM) with the linear predictor $X'\beta = X^{(1)\prime}\beta_1 + X^{(2)\prime}\beta_2$. (The vector $\beta' = (\beta_1', \beta_2')$ is part of the vector parameter θ.)

Let $w_i = 1/\pi_i$ denote the sampling weights. We utilize the knowledge of the totals by imposing the following calibration constraints:

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

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

$$\sum_{i=1}^{r} w_i \frac{\beta_2' X_i^{(2)}}{\pi(Y_i, X_i^{(1)};\gamma)} = \beta_2' X^{(2),pop} \ . \quad (6b)$$

The left hand sides of (6a) and (6b) are H-T type estimators of the corresponding population totals, with the 'selection probabilities' defined by the products $\Pr(i \in R \mid Y_i, X_i) = \pi_i \times \pi(Y_i, X_i^{(1)};\gamma)$.

3.3 Estimation Algorithm, Properties of Estimators

In P-S (2008) we described an iterative estimation algorithm, which alternates between maximum likelihood estimation of the parameters underlying the sample model for given values of the parameters of the model for the response probabilities, and the solution of the calibration equations with respect to the parameters of the model for the response probabilities, for given values of the parameters of the sample model. The "given" parameters on each iteration are the corresponding estimates from the previous iteration.

Let $l(Y_i, X_i;\theta,\gamma) = \frac{\partial \log(L_{\text{Resp}})}{\partial \theta}$ with the likelihood L_{Resp} defined by (5); denote by $h(Y_i, X_i^{(1)};\theta,\gamma)$ the system of equations (6a) and (6b) and let $(\hat{\theta}', \hat{\gamma}')$ define the estimators obtained by application of the algorithm.

Theorem: Suppose that:

i) The population model belongs to the family of generalized linear models,

ii) $0 < \pi(Y_i, X_i^{(1)};\gamma) < 1$,

iii) $l(Y_i, X_i;\theta,\gamma)$ and $h(Y_i, X_i^{(1)};\theta,\gamma)$ have continuous and bounded first and second order derivatives in a neighborhood of the true parameters (θ_0, γ_0).

Then, as $N \to \infty, n \to \infty$ such that $\frac{N}{n} < \infty$, the algorithm converges in probability to (θ_0, γ_0) in the neighborhood, implying consistency of $(\hat{\theta}, \hat{\gamma})$. Also, $\sqrt{n}(\hat{\theta}, \hat{\gamma}) \to N((\theta_0, \gamma_0), \Sigma)$ for some matrix Σ.

The proof of the theorem can be obtained from the authors.

4. Imputation of Missing Values and Estimation of Population Means

Denote by,

$$\hat{f}_S(Y_i \mid X_i) = f_S(Y_i \mid X_i;\hat{\theta},\hat{\gamma}), \ \hat{\pi}(Y_i, X_i^{(1)}) = \pi(Y_i, X_i^{(1)};\hat{\gamma}), \ \hat{E}_S(Y_i \mid X_i) = E_S(Y_i \mid X_i;\hat{\theta},\hat{\gamma}) \ , \quad (7)$$

the estimates of the sample *pdf*, the response probabilities and the sample expectations. The estimates in (7) provide several possibilities for the imputation of the missing values and the estimation of the population mean of the outcome variable.

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

$$\hat{\bar{Y}}_{(1)} = \frac{1}{N} \sum_{i=1}^{r} w_i Y_i / \hat{\pi}(Y_i, X_i^{(1)}) \ . \quad (8)$$

If the covariates are known for all the sampled units, another set of estimates is obtained as,

$$\hat{\bar{Y}}_{(2)} = \frac{1}{N} \sum_{i=1}^{n} w_i Y_i^* \ ; \ \ Y_i^* = Y_i \ \text{if} \ i \in R \ , \ Y_i^* = Y_i^{imp} \ \text{if} \ i \in R^c \ . \quad (9)$$

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

$$Y_i^{imp} = E_{R^c}(Y_i \mid X_i) = E(Y_i \mid X_i, i \in R^c), \tag{10}$$

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. The *pdf* for a nonresponding unit is computed utilizing the relationship,

$$f_{R^c}(Y_i \mid X_i) = \frac{\Pr(R_i = 0 \mid Y_i, X_i^{(1)}, i \in S) f_S(Y_i \mid X_i)}{\Pr(R_i = 0 \mid X_i, i \in S)} = \frac{[1 - \pi(Y_i, X_i^{(1)})] f_S(Y_i \mid X_i)}{[1 - \pi(X_i)]}, \tag{11}$$

where $\pi(X_i) = \int \Pr(R_i = 1 \mid Y_i, X_i^{(1)}, i \in S) f_S(Y_i \mid X_i) dY_i$. In practice, one has to use the estimated *pdf*, obtained by replacing the unknown parameters by their estimates.

Remark 2. It is important to emphasize that we don't assume any model for the outcomes of the nonresponding units. This model is defined mathematically by the relationship (11). The sample model, $f_S(Y_i \mid X_i)$, and the model for the response probabilities, $\pi(Y_i, X_i^{(1)})$, define the model holding for the outcomes of the responding units and this model can be validated by application of classical goodness of fit test statistics since it refers to the observed data.

The predictor $\hat{\bar{Y}}_{(2)}$ in (9) 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,

$$\begin{aligned} P_{X|0}(x_i) &= \frac{P(R_i = 0 \mid X_i = x_i, i \in S)}{P(R_i = 0 \mid i \in S)} \Pr(X_i = x_i \mid i \in S) \\ &= \frac{P(R_i = 0 \mid X_i = x_i, i \in S) \Pr(X_i = x_i \mid R_i = 1, i \in S) \Pr(R_i = 1 \mid i \in S)}{P(R_i = 0 \mid i \in S) \Pr(R_i = 1 \mid X_i = x_i, i \in S)}. \end{aligned} \tag{12}$$

Estimating $\Pr(X_i = x_i \mid R_i = 1, i \in S) = \frac{1}{r} \quad \forall x_i \in R$ and $\Pr(R_i = 1 \mid i \in S) = \frac{r}{\sum_{j=1}^{r}(1/\hat{\pi}(x_i))}$, the probability $P_{X|0}(x_i)$ can be estimated as,

$$\hat{P}_{X|0}(x_i) = \frac{[1 - \hat{\pi}(x_i)]}{\hat{\pi}(x_i)[\sum_{j=1}^{r}(1/\hat{\pi}(x_j)) - r]}, \; x_i \in R. \tag{13}$$

Remark 3. The estimator (13) assumes that the covariates in the subsample of the nonrespondents take the same values as in the subsample of the respondents (but with different probabilities). In practice, the estimate $\Pr(X_i = x_i \mid R_i = 1, i \in S) = \frac{1}{r}$ can be replaced by a 'smoothed' estimator, using more advanced density estimation methods.

5. Estimation of Variances of Estimators of Population Mean

In Section 4 we considered several estimators of the population mean of the outcome variable. In order to estimate the variance of these estimators, we can apply a parametric bootstrap procedure, distinguishing between estimation of the conditional variance given the observed covariates (and thus conditioning also on the number of

respondents), and the unconditional variance over all possible samples of respondents (and thus also over all possible numbers of respondents). The bootstrap procedure for estimating the conditional variances consists of the following steps:

1. Generate a large number of samples of outcomes from the estimated respondents' distribution $f_R(Y_i \mid X_i; \hat{\theta}, \hat{\gamma})$ with fixed (original) covariates X_i.
2. For each new sample, re-estimate (θ, γ) and then compute the estimators $\hat{\bar{Y}}_{(1)}$ and $\hat{\bar{Y}}_{(2)}$ (two versions), using the new parameter estimators.
3. Denote by $\hat{\bar{Y}}_{(k)}^{(b)}$, $k = 1,2$ the estimators obtained for bootstrap sample b, $b = 1,...,B$. Estimate,

$$\hat{Var}(\hat{\bar{Y}}_{(k)}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\bar{Y}}_{(k)}^{(b)} - \bar{Y}_{(k)})^2;\ \bar{Y}_{(k)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\bar{Y}}_{(k)}^{(b)},\ k = 1,2. \qquad (14)$$

For estimating the unconditional variances we first generate the outcomes for the whole population using the estimated distribution $f_S(Y_i \mid X_i; \hat{\theta}, \hat{\gamma})$, and then select respondents with probabilities $\pi(Y_i, X_i^{(1)}; \hat{\gamma})$. In this case the covariates and the sample sizes are not fixed. The rest of the computations are as before.

Another way of estimating the variance of the H-T type estimator $\hat{\bar{Y}}_{(1)}$ is by computing the conditional variance,

$$Var(\hat{\bar{Y}}_{(1)}) = Var[\tilde{Y}_{(1)} \mid \tilde{T}_x^t = (N, X_1^{pop}, ..., X_m^{pop}, \hat{\theta}^{(2)\prime} X^{pop,(2)})^t], \qquad (15)$$

where $\tilde{Y}_{(1)} = \frac{1}{N}\sum_{i=1}^{r} w_i Y_i / \pi(Y_i, X_i^{(1)}; \gamma)$, and

$\tilde{T}_x = [\sum_{i=1}^{r} w_i \frac{1}{\pi(Y_i, X_i^{(1)}; \gamma)}, \sum_{i=1}^{r} w_i \frac{X_{1i}}{\pi(Y_i, X_i^{(1)}; \gamma)}, ..., \sum_{i=1}^{r} w_i \frac{X_{mi}}{\pi(Y_i, X_i^{(1)}; \gamma)}, \sum_{i=1}^{r} w_i \frac{\hat{\theta}^{(2)\prime} X_i^{(2)}}{\pi(Y_i, X_i^{(1)}; \gamma)}]^t$. This variance accounts for the calibration equations used for estimating the model parameters and hence the response probabilities. Denote $\sigma_{11} = Var(\tilde{Y}_{(1)})$, $\Sigma_{22} = Var(\tilde{T}_x)$ and $\sigma'_{12} = Cov(\tilde{Y}_{(1)}, \tilde{T}_x)$. Assuming $\tilde{Y}_{(1)} \cong \delta^t \tilde{T}_x + \varepsilon,\ E(\varepsilon \mid \tilde{T}_x) = 0$ for some vector δ, (e.g., by assuming asymptotic normality of $(\tilde{Y}_{(1)}, \tilde{T}_x)$),

$$Var(\hat{\bar{Y}}_{(1)}) = \sigma_{11} - \sigma'_{12}\Sigma_{22}^{-1}\sigma_{12}. \qquad (16)$$

The variance components in (16) and hence the variance of the estimator $\hat{\bar{Y}}_{(1)}$ can be estimated by 'design-based' methods, or by the joint model-design distribution with the unknown model parameters replaced by their original sample estimators.

Finally, the variance of $\hat{\bar{Y}}_{(2)}$ that uses observed and imputed values can be estimated also using the multiple imputations method. Suppose that M outcomes are imputed for each nonresponding unit j. Let $\hat{\bar{Y}}_{(2),m}$ denote the estimator $\hat{\bar{Y}}_{(2)}$, computed from the observed outcomes and the m-th set of imputed values, $m = 1,...,M$. Following the theory of multiple imputations, the variance of $\hat{\bar{Y}}_{(2)} = \sum_{m=1}^{M} \hat{\bar{Y}}_{(2),m}$ is estimated as,

$$\hat{Var}(\hat{\bar{Y}}_{(2)}) = (1 + M^{-1})\hat{B} + \hat{V};\ \hat{B} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\bar{Y}}_{(2),m} - \hat{\bar{Y}}_{(2)})^2,\ \hat{V} = \frac{1}{M}\sum_{m=1}^{M}\hat{V}_m, \qquad (17)$$

where $\hat{V}_m = \hat{Var}(\hat{\bar{Y}}_{(2),m})$.

6. Testing Which Covariates to Include in Response probabilities Model

Let Z be a variable not included in response model $\pi(Y_i, X_i^{(1)};\gamma)$. Suppose that $Z^{tot} = \sum_{i=1}^{N} Z_i$ is known. The 'standard' way of testing whether Z should be included in the response model is to add it to the other covariates, refit the model holding for the respondents and then test the significance of the coefficient of Z. This procedure can be time consuming. Below we consider an alternative test, which does not require refitting the model. If the probabilities $\pi(Y_i, X_i^{(1)};\gamma)$ do not depend on Z, then we expect that the H-T type estimator of Z^{tot} will be unbiased. Thus, we may test the hypothesis $H_0 : E\left(\sum_{i=1}^{r} \frac{\omega_i Z_i}{\pi(Y_i, X_i^{(1)};\gamma)}\right) = Z^{tot}$. A plausible test statistic is,

$$U = \{[\sum_{i=1}^{r} \frac{\omega_i Z_i}{\pi(Y_i, X_i^{(1)};\hat{\gamma})} - Z^{tot}] / \hat{Std}(\sum_{i=1}^{r} \frac{\omega_i Z_i}{\pi(Y_i, X_i^{(1)};\hat{\gamma})})\} \overset{Asymp}{\sim} N(0,1) . \qquad (18)$$

The variance of $\sum_{i=1}^{r} \omega_i Z_i / \pi(Y_i, X_i^{(1)};\hat{\gamma})$ is the conditional variance, given the calibration constraints and it can be estimated similarly to the estimation of the variance of the estimator $\hat{\bar{Y}}_{(1)}$ described in Section 5, (Eqs. 15 and 16).

The procedure can be extended for testing simultaneously whether several additional covariates need to be added to the model and/or for designing an appropriate stepwise algorithm.

7. Empirical Results

7.1 Study Population and Outcome Variable

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 increased to 90% of the sampled HHs. The HHs were sampled with equal probabilities. In what follows 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 these members. The target outcome variable is the *household income per standard person*.

For the present study we define the responding HHs to be the HHs that responded on the first interview. The nonresponding HHs are the HHs which did not respond on the first interview but responded on one of the later interviews, such that the data for both the responding and the nonresponding HHs are actually known. The total number of HHs in our sample is n=1721, with $r = 631$ responding HHs and $n - r = 1090$ nonresponding HHs (but responding on one of the later recalls).

7.2 Sample Model and Response Probabilities

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

$$\log(Y_i) = X_i'\beta + \varepsilon_i, \;\; \varepsilon_i \sim N(0, \sigma_\varepsilon^2) , \qquad (19)$$

where Y_i is the income per standard person in household i and $X_i = [1, X_{i1}, ..., X_{iK}]'$ is the corresponding vector of covariates. The covariates include characteristics of the head of the HH: gender, age, occupation and number of years at school; and HH characteristics: number of earners, HH size and district.

The response probabilities given the outcome and the covariates are modeled by the logistic function, that is,

$$P(R_i = 1 | Y_i, X_i^{(1)}) = [1 + e^{-(\gamma_0 Y_i + \gamma_1' X_i^{(1)})}]^{-1}. \quad (20)$$

As established by Landsman (2008), the model holding for the responding units (Eq. 4) resulting from a normal sample distribution and a logistic model for the response probabilities, with at least one different covariate, is identifiable. Most of the covariates included in the sample model, as well as the outcome variable log(income) are nonsignificant when included in the model (20). However, removing the nonsignificant covariates from the model makes the log(income) variable significant. As shown below, the resulting model contains much less covariates.

Tables 1 and 2 show the estimated coefficients for the two models as obtained when fitting the models separately for all the sample data, and when fitting the respondents' model (4) to only the responding units, using the estimation algorithm described in Section 3.

Table 1: Sample model fitted to all the sampled HH (Respondents and Nonrespondents"), and based only on responding HH.

Coeff.	Cons.	Gender	Age	Dist. 21	Dist.41	Dist.42	Dist. 43
All HH	7.32	-0.13	0.02	-0.18	0.17	0.13	0.17
Respond.	7.22	-0.14	0.02	-0.10	0.15	0.10	0.16

Coeff.	Dist. 44	Dist.51	Dist. 52	Earners	HHsize	Occ.0	Occ.1
All HH	0.18	0.23	0.09	0.24	-0.14	0.44	0.22
Respond.	0.17	0.28	0.15	0.26	-0.13	0.45	0.24

Coeff.	Occ.2	Occ.3	Occ.4	School10	School12	School16	σ_ε^2
All HH	0.44	0.21	0.15	-0.36	-0.15	0.17	0.401
Respond.	0.38	0.26	0.15	-0.36	-0.15	0.20	0.404

Table 2: Model for response probabilities fitted to all the sampled HH (Respondents and "Nonrespondents"), and based only on responding HH.

Coeff.	Cons.	Log(Y)	Gender	Dist.43	Dist.44	Dist.53	HHsize
All HH	0.91	-.21	-0.20	0.88	-0.58	-0.77	0.10
Respond.	1.38	-.21	-0.26	0.91	-0.59	-0.79	0.12

The values of the coefficients in the two tables show that the coefficients can be estimated sufficiently accurately based only on the model holding for the responding units with 631 respondents. When fitting the sample model (Eq. 19) to all the sample data, we obtained $R^2 = 0.60$ with residual variance $\hat{\sigma}_\varepsilon^2 = 0.401$. The values of the regression coefficients are sensible. For example, the coefficients of the education variables increase 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 the same standard deviation $\sigma_\varepsilon^2 = 0.401$. The distribution of the residuals is seen to be close to the normal distribution, although with somewhat shorter tails.

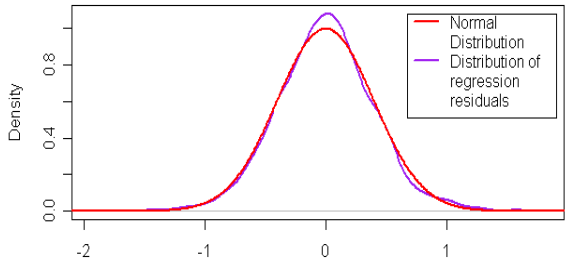

Figure 1: Distribution of estimated regression residuals and normal distribution with mean zero and the same variance ($\sigma_\varepsilon^2 = 0.401$).

7.3 Imputation of Missing Outcomes

Next we show the performance of the proposed approach in imputing the missing outcomes. The imputations were 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 $\hat{f}_{R^c}(Y_i \mid X_i) = f(Y_i \mid X_i, R_i = 0; \hat{\beta}, \hat{\sigma}_\varepsilon^2, \hat{\gamma})$. We imputed 5 values for each unit and then averaged the 5 imputations. 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 imputing the missing covariates using Eq. 13, and then imputing the incomes similarly to Scenario 1. Figures 2 and 3 compare the true empirical cumulative distribution of the incomes of the nonresponding units with the means of the estimated empirical distributions over the 5 imputation sets. Also shown in the two figures is the cumulative distribution of the imputed values when ignoring the nonresponse process, imputing the missing covariates by drawing at random from their empirical distribution for the responding HH and imputing the missing incomes given the covariates by drawing at random from the estimated sample distribution.

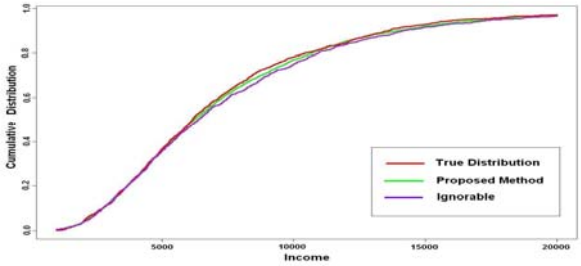

Figure 2: True empirical cumulative distribution and means of estimated empirical cumulative distributions of the incomes over 5 imputation sets. Known covariates.

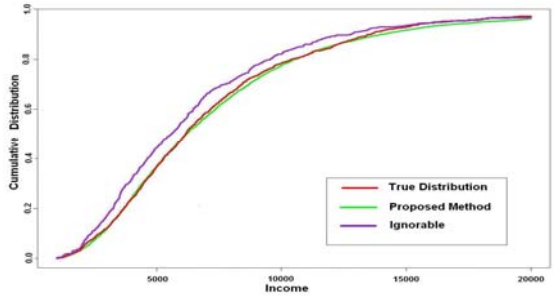

Figure 3: True empirical cumulative distribution and means of estimated empirical cumulative distributions of the incomes over 5 imputation sets. Missing covariates.

Figures 2 and 3 show that the proposed approach yields imputations with distribution that is close to the true distribution. On the other hand, ignoring the nonresponse yields biased imputations, particularly when the covariates for the nonresponding units are likewise unknown. 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 unknown covariates for the nonresponding units would still produce biased estimates for the income distribution, since the nonresponse for some of the covariates cannot be ignored. For example, Table 3 shows the percentage of HH by size for the responding and nonresponding units. The HH size is one of the important covariates in both the models (19) and (20) (Tables 1 and 2).

Table 3: Percent of households by size in household expenditure survey.

HH size	1	2	3	4	5	6+
Resp.	6.18	13.63	19.33	26.94	20.60	13.31
NonResp.	12.39	19.00	17.34	24.40	17.34	9.54

7.4 Estimation of Mean Sample Income and Variance of Estimators

In Section 4 we proposed two different estimators of the population mean of the outcome variable and in Section 5 we considered alternative ways of estimating their variance. Tables 4 and 5 summarize the results obtained when estimating the true sample mean of the incomes. Table 4 presents the estimated standard errors (SE) when conditioning on the observed covariates (and hence also on the number of respondents). Table 5 presents the unconditional SE estimators. For both cases we used bootstrap samples as described in Section 5. Also shown in the two tables is the mean and variance over all bootstrap samples of the H-T estimator, which uses the 'true' probabilities to respond, $\pi(Y_i, X_i; \hat{\gamma})$, that is, when the probabilities to respond are not reestimated for each of the bootstrap samples. This estimator, denoted by $\hat{\bar{Y}}_{(1,P-K)}$, does not take into account the known totals of the covariates via the calibration constraints. The estimator $\hat{\bar{Y}}_{(2)}$ that uses the imputed values is calculated under Scenario 1, where we assume that the covariates are known for the nonresponding units, (denoted by $\hat{\bar{Y}}_{(2,C-K)}$), and under Scenario 2, where the covariates for the nonresponding units are also imputed (denoted by $\hat{\bar{Y}}_{(2,C-UK)}$). In both tables we also show the results obtained when estimating the variance by the multiple imputations method (Eq. 17).

Table 4: Estimation of sample mean of income ($\bar{Y} = 7244.46$). Conditional SE. 100 bootstrap samples.

Estimator	Estimate		Standard Error	
	Original sample of respondents	Mean over bootstrap samples	Parametric bootstrap	Mean of Multiple Imputation
$\hat{\bar{Y}}_{(1,P-K)}$	----	7347.00	213.20	----
$\hat{\bar{Y}}_{(1)}$	7381.75	7345.00	182.37	----
$\hat{\bar{Y}}_{(2,C-UK)}$	7392.73	7356.20	127.40	132.69
$\hat{\bar{Y}}_{(2,C-K)}$	7318.05	7316.38	122.09	123.87

Table 5: Estimation of sample mean of income ($\bar{Y} = 7244.46$). Unconditional SE. 500 bootstrap samples.

Estimator	Estimate		Standard Error	
	Original sample of respondents	Mean over bootstrap samples	Parametric bootstrap	Mean of Multiple Imputation
$\hat{\bar{Y}}_{(1,P-K)}$	----	7335.90	346.25	----
$\hat{\bar{Y}}_{(1)}$	7381.75	7317.91	187.51	----
$\hat{\bar{Y}}_{(2,C-UK)}$	7392.73	7344.37	175.32	158.80
$\hat{\bar{Y}}_{(2,C-K)}$	7318.05	7344.45	173.00	158.98

Tables 4 and 5 illustrate that all the estimators of the mean population income overestimate the true mean, but with the largest bias being less than 1.5%. Notice that the mean of the incomes computed from only the responding units is 6842.22, an underestimation of 5.5%. As anticipated, the standard errors of the estimators are smaller when conditioning on the observed covariates (Table 4), than in the case where the standard errors are taken over all possible samples of respondents (Table 5). Also, the standard errors are somewhat smaller when the covariates for the nonresponding units are known than in the case that they have to be imputed. The estimator $\hat{\bar{Y}}_{(1,P-K)}$, which does not use the calibration constraints has a much larger variance than the other estimators, illustrating the advantage of modifying the sampling weights by use of calibration constraints. The multiple imputations method estimates fairly well the conditional variance in table 4, but underestimates the unconditional variance.

Finally, for estimating the unconditional standard error of the H-T type estimator $\hat{\bar{Y}}_{(1)}$ we also computed for each of the 500 bootstrap samples the estimator (16), using the joint response-model distribution with estimated parameters (the distribution over all possible samples of respondents and the sample distribution). The mean of the SE estimators turned out to be 188.46, which is very close to the empirical standard error of 187.51 over all the bootstrap samples. The standard error estimator based on the original sample is 187.27.

7.5 Testing Which Covariates to Include in Response Probability Model

In section 6 we proposed a new test statistic, defined by (18), for testing which covariates to include in the response probability model. The performance of the test statistic should be assessed by its distribution under the null hypothesis that it should not be included in the model, and by its power in rejecting the null hypothesis. To this end, we computed the empirical distribution of the test statistic for different covariates, using the parametric bootstrap samples generated for estimating the unconditional variances described in Section 7.4.

Figure 4 compares the empirical histograms of the test statistic for all the covariates that are not included in model with the standard normal distribution (the asymptotic distribution under H_0).

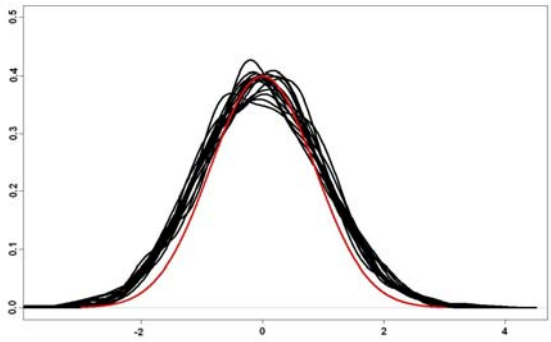

Figure 4: Histograms of test statistic for covariates not included in response probability model. 500 bootstrap samples. The red curve is the standard normal *pdf*.

Figure 5 compares the empirical histograms of the test statistic for the five covariates included in the model (Table 2) with the standard normal distribution. For this experiment we dropped each covariate in turn and recomputed the response probabilities based on the remaining covariates in the model.

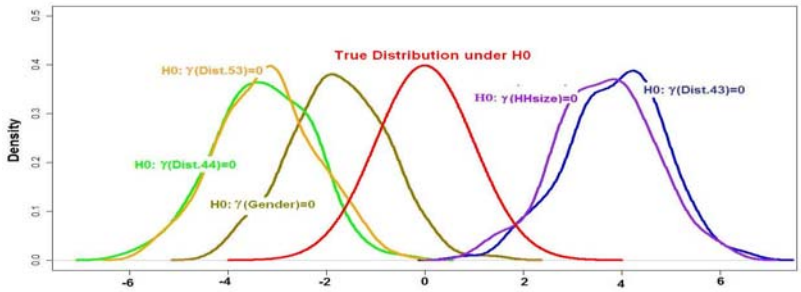

Figure 5: Histograms of test statistic for covariates included in response probability model. 150 bootstrap samples. The red curve is the standard normal *pdf*.

Figures 4 and 5 illustrate good performance of the test statistic in the present application. For covariates that were not included in the model the empirical distribution of the test statistic is sufficiently close to the standard normal *pdf*, while the empirical distributions for covariates that are included in the model have means far from zero.

8. Final Remark

We developed a general approach for imputation and estimation when the nonresponse is not NMAR. The proposed approach is model-based and its good performance is likely to depend in general on correct specification of the population model and the model for the response probabilities. NMAR nonresponse is a difficult problem and making strong assumptions is inevitable. The advantage of our approach, however, is that for given specifications of the two models, the goodness of fit of the resulting model holding for the responding units can be tested by use of classical goodness of fit testing procedures, since the latter model refers to the observed data. Several tests of this kind have been established and illustrated in Landsman (2008). We developed additional tests and we are presently investigating their performance.

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*, **95**, 555-571.

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, 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 Sikov, A. (2008). Estimation and imputation under nonignorable nonresponse with missing covariate information. In *JSM Proceedings*, Section on Survey Research Methods, Alexandria, VA: American Statistical Association. 90-101.

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: Chapman and 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**, 747-764.