UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Working Paper M10/04
Methodology

Imputation And Estimation Under Nonignorable Nonresponse For Household Surveys With Missing Covariate Information

Danny Pfeffermann, Anna Sikov

Abstract

In this paper 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 totals of some or all of the auxiliary variables in the two models, but we do not require that the auxiliary variables are observed for the nonresponding units. We develop an algorithm for estimating the parameters governing the two models and show how to estimate the distributions of the missing covariates and outcomes, which are then used for imputing the missing values for the nonresponding units and for estimating population means and the variances of the estimators. We also consider several test statistics for testing the model fitted to the observed data and study their performance, thus validating the proposed procedure. 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.

Imputation and Estimation under Nonignorable Nonresponse for Household Surveys with Missing Covariate Information

Danny Pfeffermann[1] and Anna Sikov[2]

[1]Hebrew University of Jerusalem, Israel, and
Southampton Statistical Sciences Research Institute, UK.

[2] Hebrew University of Jerusalem, Israel.

Abstract

In this paper 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 totals of some or all of the auxiliary variables in the two models, but we do not require that the auxiliary variables are observed for the nonresponding units. We develop an algorithm for estimating the parameters governing the two models and show how to estimate the distributions of the missing covariates and outcomes, which are then used for imputing the missing values for the nonresponding units and for estimating population means and the variances of the estimators. We also consider several test statistics for testing the model fitted to the observed data and study their performance, thus validating the proposed procedure. 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 type estimator, Nonrespondents distribution, Respondents distribution.

Acknowledgement: This research is supported by a grant from the U.S.-Israel Bi-national Science Foundation.

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, and possibly also on 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 nonresponse depends on the missing outcomes. We consider mostly 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.

We propose a new approach for handling NMAR nonresponse, which does not require knowledge of the covariates for the nonrespondents. We assume a model for the outcome variable under complete response (the 'sample model'), and a model for the response probability, which is allowed to depend on the outcome and auxiliary variables. 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 products of the sampling weight (inverse of the sample inclusion probability) and the inverse of the response probability under the model. 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 prove the convergence of the algorithm and discuss the properties of the resulting estimators.

Having estimated the model parameters, 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 for the nonresponding units, the missing values of the covariates are imputed as well from their approximate distribution. The variances of the proposed estimators are estimated by parametric and resampling methods. Finally, we test the combined model fitted for the responding units by using standard tests that compare the cumulative hypothesized distribution with the corresponding empirical distribution, and by testing moments of the hypothesized model.

The various procedures considered in this article are illustrated using data collected as part of the Household Expenditure Survey (HES) carried out by the Israel Central Bureau of Statistics in 2005. The initial response rate in this survey was 43%, but after many recalls, the final response rate went up to 90%. This survey provides therefore a rare opportunity of comparing the imputed values after the first interview with the actual values obtained from the recalls.

2. EXISTING APPROACHES TO DEAL WITH NMAR NONRESPONSE

In this section we review briefly some of the approaches proposed in the literature to deal with NMAR nonresponse. Let y_i denote the value of an outcome variable Y, associated with unit i belonging to a sample $S=\{1,...,n\}$, drawn from a finite population $U=\{1,...,N\}$ by probability sampling with known inclusion probabilities $\pi_i=\Pr(i\in s)$. Let $x_i=(x_{1i},...,x_{pi})$ denote the values of *p* auxiliary variables (covariates) associated with unit i. In what follows we assume that the population outcomes are independent realizations from distributions with probability density functions (*pdf*), $f_P(y_i|x_i;\theta)$, governed by an unknown vector parameter θ. Let $R=\{1,...,r\}$ define the sample of respondents with observed outcomes and covariates, and $R^c=\{r+1,...,n\}$ define the subsample of nonrespondents, for which at least the outcomes are missing. The response process is assumed to be independent between units.

In the present study we assume that the sampling process is noninformative such that under complete response, $f_S(y_i\mid x_i)=f(y_i\mid x_i,i\in S)=f_P(y_i\mid x_i)\ \forall i$. Most of the approaches considered 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 full response are distinct from the parameters governing the response process, the nonresponse can be ignored for likelihood and Bayesian 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 consider situations where the probability to respond may depend also on the outcome value even after conditioning on the covariates. 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 outcomes are not missing at random (NMAR).

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 (does not respond) to the outcome variable. A possible way to deal with the nonresponse in this situation 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.

A- *Selection Models* specify,

$$f(y_i, R_i \mid x_i; \theta, \gamma) = \Pr(R_i \mid y_i, x_i; \gamma) f_S(y_i \mid x_i; \theta), \tag{2}$$

where $f_S(y_i \mid x_i; \theta)$ defines the sample *pdf* (model), $\Pr(R_i \mid y_i, x_i; \gamma)$ models the response process and θ and γ denote the (distinct) parameters of the two models. Assuming that the outcomes are independent given the covariates, the *full likelihood* takes in this case the form,

$$L = \prod_{i=1}^{r} \Pr(R_i = 1 \mid y_i, x_i; \gamma) f_S(y_i \mid x_i; \theta) \prod_{i=r+1}^{n} \Pr(R_i = 0 \mid x_i; \theta, \gamma), \tag{3}$$

where $\Pr(R_i = 0 \mid x_i; \theta, \gamma) = 1 - \int \Pr(R_i = 1 \mid y_i, x_i; \gamma) f_S(y_i \mid x_i; \theta) dy_i$. The response probability is often modeled as,

$$\Pr(R_i = 1 \mid y_i, x_i; \gamma) = g(\gamma_0 + \gamma_1^t x_i + \gamma_2 y_i), \tag{4}$$

for some function *g* taking values in the range (0,1) (see below).

Suppressing for convenience the parameters from the notation, the missing sample values can be imputed in this case by the expectations $E_{R^c}(Y_i \mid x_i) = E(Y_i \mid x_i, R_i = 0)$, which can be calculated using Bayes theorem as,

$$E_{R^c}(Y_i \mid x_i) = \int_{-\infty}^{\infty} y_i f(y_i \mid x_i, R_i = 0) dy_i = \int_{-\infty}^{\infty} y_i \frac{P(R_i = 0 \mid y_i, x_i) f_S(y_i \mid x_i)}{\int_{-\infty}^{\infty} P(R_i = 0 \mid y_i, x_i) f_S(y_i \mid x_i) dy_i} dy_i \cdot \tag{5}$$

In practice, the probabilities and densities in (5) are replaced by their estimates, obtained by estimating the unknown parameters. Alternatively, the imputed values can be obtained 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. 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), the use of the likelihood in (3) is 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. The author estimates the parameters γ by maximizing the likelihood $L = \prod_{i=1}^{r} P(R_i = 1 \mid y_i, x_i; \gamma) \prod_{i=r+1}^{n} P(R_i = 0 \mid x_i; \theta, \gamma)$ with respect to γ assuming that θ is 'known', and the parameters θ by solving weighted least square equations, assuming that γ is 'known'. The procedure is carried out iteratively, with the 'known' values on a given iteration defined by the estimates obtained on the previous iteration, and with the weights defined by the inverse response probabilities as computed on the previous iteration. A drawback of this method is that the probability $P(R_i = 0 \mid x_i; \theta, \gamma)$ cannot actually be calculated, since the sample *pdf* of $Y_i \mid x_i$ is not specified. The author deals with this problem by expanding $P(R_i = 1 \mid y_i, x_i, \gamma)$ around the mean $E_S(Y_i \mid x_i)$, but this amounts to assuming a MAR nonresponse.

B- *Pattern-mixture* models specify,

$$f(y_i, R_i \mid x_i; \psi_m^{(l)}, \psi_r) = f(y_i \mid x_i, R_i; \psi_m^{(l)}) \Pr(R_i \mid x_i; \psi_r), \tag{6}$$

where $f(y_i \mid x_i, R_i; \psi_m^{(l)}), l = 0,1$ define the *pdf*s of *Y* under the different patterns of the missing data, $(R_i = 0, R_i = 1)$, and $\Pr(R_i \mid X_i; \psi_r)$ models the response probability given the covariates, with $\psi_m^{(l)}$ and ψ_r denoting the corresponding unknown parameters. The likelihood takes in this case the form,

$$L = \prod_{i=1}^{r} f(y_i \mid x_i, R_i = 1; \psi_m^{(1)}) \Pr(R_i = 1 \mid x_i; \psi_r) \prod_{i=r+1}^{n} \Pr(R_i = 0 \mid x_i; \psi_r). \quad (7)$$

A major drawback of pattern-mixture models is that the model holding for the nonrespondents, $f(y_i \mid x_i, R_i = 0; \psi_m^{(0)})$, cannot be extracted from the models $f(y_i \mid x_i, R_i = 1; \psi_m^{(1)})$ and $\Pr(R_i \mid x_i; \psi_r)$ fitted under this approach, and hence it is not clear how to impute the missing outcomes unless under strong assumptions, which are generally hard to test. Little (1993, 1994) discusses plausible relationships between the parameters governing the models holding for the respondents and nonrespondents and provides examples for the application of selection and pattern-mixture models. Rubin (1987) discusses selection and pattern-mixture models from a Bayesian perspective.

Tang *et al*. (2003) propose a 'pseudo-likelihood' method that uses the conditional *pdf,* $g_S(x_i \mid y_i)$ for the respondents. Application of this method requires specification of the sample *pdf* $f_S(y_i \mid x_i)$, and of the marginal *pdf* $g_S(x_i)$, which can be replaced by the empirical sample distribution. The method does not require a parametric model for the response probability but it assumes that it depends only on the outcome. The likelihood takes now the form,

$$L = \prod_{i=1}^{r} g_S(x_i \mid y_i; \theta, \eta) = \prod_{i=1}^{r} \frac{f_S(y_i \mid x_i; \theta) g_S(x_i; \eta)}{\int f_S(y_i \mid x; \theta) g_S(x; \eta) dx}. \quad (8)$$

Note that although the product is only over the responding units, estimation of the *pdf* $g_S(x_i)$ requires that the covariates are known for all the sample units. The authors point out that this method is robust to misspecification of the response process but the use of this approach does not allow imputing the missing outcomes based on the distribution $f_{R^c}(y_i \mid x_i) = f(y_i \mid x_i, R_i = 0)$.

So far we considered methods that require that 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 defined as,

$$L = \prod_{i=1}^{r} \Pr(R_i = 1 \mid y_i, x_i; \gamma)\, p_i (1-\lambda)^{n-r}\,, \tag{9}$$

where $\lambda = \Pr(R_i = 1)$ and $p_i = dF_S(y_i, x_i)$ is the 'jump' of the joint cumulative distribution $F_S(y_i, x_i)$ at (y_i, x_i), $i = 1,...,r$, which, however, is not defined. The empirical likelihood is maximized under the constraints,

$$\sum_{i=1}^{r} p_i [\Pr(R_i = 1 \mid y_i, x_i; \gamma) - \lambda] = 0,\ \sum_{i=1}^{r} p_i (x_i - \bar{X}^{pop}) = 0;\ p_i \geq 0,\ \sum_{i=1}^{r} p_i = 1\,. \tag{10}$$

The use of this method addresses the problem of missing covariate information by using the unconditional response probability $\lambda = \Pr(R_i = 1)$ in the likelihood, and it accounts for the known population means of the covariates. However, our experience so far shows that the performance of this procedure depends on having sufficiently accurate initial values for the response model parameters and the Lagrange multipliers used for the constrained maximization.

Chang and Kott (2008) propose an approach for estimating the response probabilities that uses 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 unknown parameters of this model by regressing the H-T estimators of the totals of the calibration variables against the corresponding known totals. The weights used for the H-T estimators are the product of the sampling weights and the inverse of the response probabilities under the model. The use of this approach allows estimating population totals of interest, but it does not allow imputation of the missing data, since no model is assumed for the outcome values.

3. THE RESPONDENTS DISTRIBUTION AND PARAMETER ESTIMATION

3.1 *The respondents distribution and its relation to the sample distribution*

In what follows we denote by x_i the covariates included in the population model and by v_i the covariates included in the response model. Let $z_i = (x_i \cup v_i)$.

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

$$f_R(y_i \mid z_i) = f(y_i \mid z_i, i \in S, R_i = 1) = \frac{\Pr(R_i = 1 \mid y_i, v_i, i \in S)}{\Pr(R_i = 1 \mid z_i, i \in S)}\ f_S(y_i \mid x_i)\,, \tag{11}$$

where $\Pr(R_i=1|z_i, i\in S)=\int \Pr(R_i=1|y_i, v_i, i\in S) f_S(y_i \mid x_i) dy_i$ and $f_S(y_i \mid x_i)$ is the sample *pdf* under complete response. As noted before, in this article we assume that the sample *pdf* and the population *pdf* are the same.

Remark 1. As in selection models, the use of the respondents' model requires modeling the sample *pdf*, $f_S(y_i \mid x_i)$ and the response probability, $\Pr(R_i=1 \mid y_i, v_i, i\in S)$. Notice, however, that the resulting respondents' model can be tested, since it relates to the data observed for the responding units (see Section 7).

By (11), if the sample outcomes and the response are independent between the units, one can estimate the parameters (θ, γ) by maximizing the 'respondents likelihood',

$$L_{\text{Resp}}=\prod_{i=1}^{r} f(y_i \mid z_i, R_i=1, i\in S;\theta,\gamma)=\prod_{i=1}^{r}\frac{\Pr(R_i=1|y_i, v_i, i\in S;\gamma) f_S(y_i|x_i;\theta)}{\Pr(R_i=1|z_i, i\in S;\theta,\gamma)}. \qquad (12)$$

The notable property of the likelihood (12) is that it does not require knowledge of the covariates for nonresponding units, or modeling the distribution of the sampled covariates. As shown later, estimation of the parameters in (12) permits imputing the missing values and estimating the finite population mean of the outcome variable (or any other variable).

3.2 *The respondents' likelihood for Generalized Linear Sample Models* (*GLM*)

The GLM is defined as,

$$f_S(y_i|x_i;\beta,\phi)=\exp\{a(\phi)[y_i\sum_{k=0}^{p}\beta_k x_{ki}-g(\sum_{k=0}^{p}\beta_k x_{ki})+d(y_i)]+\eta(\phi, y_i)\}, \qquad (13)$$

where $x_{ko}=1$, $\theta=(\beta,\phi)$ defines the set of unknown parameters and $g(.)$, $a(.)$, $d(.)$ and $\eta(.)$ are known real functions with $g(\cdot)$ strictly increasing and differentiable.

In what follows we assume that $\eta(\phi, y_i)=\eta(\phi)$. The log of the respondents' likelihood in (12) for the GLM in (13) can be written then as,

$$\begin{aligned} l_{\text{Resp}} &= a(\phi)\sum_{i=1}^{r}[y_i\sum_{k=0}^{p}\beta_k x_{ki}-g(\sum_{k=0}^{p}\beta_k x_{ki})+d(y_i)]+r\eta(\phi) \\ &+\sum_{i=1}^{r}\log[\Pr(R_i=1|y_i, v_i, i\in S;\gamma)]-\sum_{i=1}^{r}\log[\Pr(R_i=1|z_i, i\in S;\beta,\phi,\gamma)] \end{aligned}. \qquad (14)$$

Denote $\pi(y_i, v_i;\gamma)=\Pr(R_i=1|y_i, v_i, i\in S;\gamma)$. Taking the derivatives of the log-likelihood with respect to β and ϕ, we obtain after some tedious algebra the following equations:

$$l_k = \sum_{i=1}^{r} [y_i - E_R(Y_i \mid z_i; \beta, \phi, \gamma)] x_{ki} = 0,\ k = 0, ..., p\,, \tag{15a}$$

$$l_{p+1} = \sum_{i=1}^{r} [d(y_i) - E_R(d(Y_i) \mid z_i; \beta, \phi, \gamma)] = 0\,, \tag{15b}$$

where $E_R(Y_i \mid z_i; \beta, \phi, \gamma) = \int y_i f_R(y_i \mid z_i; \beta, \phi, \gamma) dy_i = \int y_i \frac{\pi(y_i, v_i; \gamma) f_S(y_i \mid x_i; \beta, \phi) dy_i}{\int \pi(y_i, v_i; \gamma) f_S(y_i \mid x_i; \beta, \phi) dy_i}$.

Let $\gamma = (\gamma_0, ..., \gamma_{q+1})$. Taking derivatives of the log-likelihood (14) with respect to γ and assuming that the order of integration and differentiation can be interchanged yields the equations,

$$\sum_{i=1}^{r} E_R\left(\frac{\partial \pi(y_i, v_i; \gamma)}{\partial \gamma_k} \frac{1}{\pi(y_i, v_i; \gamma)} \mid z_i\right) = \sum_{i=1}^{r} \left(\frac{\partial \pi(y_i, v_i; \gamma)}{\partial \gamma_k} \frac{1}{\pi(y_i, v_i; \gamma)}\right),\ k = 0, ..., q\,. \tag{16}$$

For example, if the response probability is logistic such that $\pi(y_i, v_i; \gamma) = [1 + \exp(-(\gamma_0 + v_i'\gamma + \gamma_{q+1} y_i)]^{-1}$, we obtain the following equations:

$$\begin{aligned}
&\sum_{i=1}^{r} E_R(\pi(y_i, v_i; \gamma) \mid z_i) = \sum_{i=1}^{r} \pi(y_i, v_i; \gamma)\ ;\ \sum_{i=1}^{r} v_{1i} E_R(\pi(y_i, v_i; \gamma) \mid z_i) = \sum_{i=1}^{r} v_{1i} \pi(y_i, v_i; \gamma) \\
&\qquad\qquad \vdots \\
&\sum_{i=1}^{r} v_{qi} E_R(\pi(y_i, v_i; \gamma) \mid z_i) = \sum_{i=1}^{r} v_{qi} \pi(y_i, v_i; \gamma)\ ;\ \sum_{i=1}^{r} E_R(y_i \pi(y_i, v_i; \gamma) \mid z_i) = \sum_{i=1}^{r} y_i \pi(y_i, v_i; \gamma)
\end{aligned} \tag{17}$$

The solution of the equations (15a), (15b) and (16), (or 17 in the case of logistic response probabilities), yields the maximum likelihood estimators (MLE) for (β, ϕ, γ).

3.3 *Calibration constraints*

In what follows we assume knowledge of the population totals of all the covariates included in the response model and at least one of the covariates included in the sample model. We later relax this requirement. The additional information contained in the population totals is not part of the likelihood (14). We utilize this information by imposing the following constraints. Let the sample pdf be the GLM defined by (13) and denote by $Z^{pop} = (Z_1^{pop}, ..., Z_{t^*}^{pop})$

$= (V_1^{pop},...,V_q^{pop}, X_1^{pop},..., X_{p^*}^{pop})$ the known population totals, where $p^* \le p = \dim(x_i)$, $t^* \le t = \dim(z_i)$. The calibration constraints are,

$$\sum_{i=1}^{r} w_i \frac{v_{ki}}{\pi(y_i, v_i; \gamma)} = V_k^{pop}, k = 1,..., q; \ \sum_{i=1}^{r} w_i \frac{1}{\pi(y_i, v_i; \gamma)} = N, \tag{18a}$$

where $\{w_i = (1/\pi_i) = 1/\Pr(i \in S)\}$ are the sampling weights. When the response model has an intercept, we use the additional constraint,

$$\sum_{i=1}^{r} w_i \frac{\tilde{\beta}' \tilde{x}_i}{\pi(y_i, v_i; \gamma)} = \tilde{\beta}' \tilde{X}^{pop} \tag{18b}$$

where $\tilde{x}_i = (x_{1i},..., x_{p^*i})$, $\tilde{X}^{pop} = (X_1^{pop},..., X_{p^*}^{pop})$ and $\tilde{\beta}$ is the vector of coefficients of $\tilde{x}_i$ in the sample model. Notice that if $E_S(Y_i \mid x_i; \beta, \phi) = \sum_{k=0}^{p} \beta_k x_{ki}$, (e.g., the sample model is normal) and $p = p^*$, the constraint (18b) implies, $\sum_{i=1}^{r} w_i \frac{E_S(Y_i \mid x_i)}{\pi(y_i, v_i; \gamma)} = \sum_{j=1}^{N} E_S(Y_j \mid x_j) = \sum_{j=1}^{N} E_p(Y_j \mid x_j)$, since we assume that the population and sample models are the same.

Remark 2. The left hand sides of (18a) and (18b) 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 $\pi(y_i, v_i; \gamma)$ (Särndal and Swensson, 1987).

3.4 *Estimation algorithm, properties of estimators*

In order to utilize the additional information provided by knowledge of the population totals, we replace Eqs. (16) by Eqs. (18a) and (18b), and use the following iterative algorithm.

Let $(\beta^{(0)}, \phi^{(0)})$ denote initial values for the vector (β, ϕ) indexing the sample *pdf* $f_s(Y_i | X_i; \beta, \phi)$.

Step j: For given $(\hat{\beta}^{(j)}, \hat{\phi}^{(j)})$ from iteration j, set $(\beta, \phi) = (\hat{\beta}^{(j)}, \hat{\phi}^{(j)})$ and solve the set of equations (18a) and (18b) as a function of the unknown parameters γ indexing the model $\pi(y_i, v_i; \gamma)$ of the response probabilities. This step yields new estimators $\hat{\gamma}^{(j+1)}$.

Step j+1: Solve (15a) and (15b) with respect to (β, ϕ), with γ equal to $\gamma^{(j+1)}$. This step yields new estimators $(\hat{\beta}^{(j+1)}, \hat{\phi}^{(j+1)})$. Continue the iterations until convergence.

Our experience so far shows that the use of this algorithm simplifies the computation of the estimators and is more stable than the solution of the likelihood equations (15a), (15b) and (16). It also utilizes the additional information provided by the known totals of the covariates. The solution of the Equations (15a) and (15b) for fixed γ is outlined in Appendix A.

Let $l(\theta,\gamma)=\dfrac{\partial \log(L_{\text{Resp}})}{\partial\theta}$ with L_{Resp} defined by (12), denote by $h(\theta,\gamma)$ the system of equations (18a) and (18b) and let $(\hat{\theta}',\hat{\gamma}')$ define the estimators obtained by application of the algorithm.

Theorem 1: Suppose that:

I) The population (sample) model belongs to the family of generalized linear models,

II) $0<\pi(y_i,v_i;\gamma)<1$, with bounded first derivatives with respect to γ.

III) The functions $l(\theta,\gamma)$ and $h(\theta,\gamma)$ are continuous and twice differentiable with respect to (θ,γ) in a compact neighborhood of the solution (θ_0,γ_0).

IV) The matrices $\dfrac{\partial l(\theta,\gamma)}{\partial\theta}$ and $\dfrac{\partial h(\theta,\gamma)}{\partial\gamma}$ are nonsingular in the neighborhood of $(\tilde{\theta},\tilde{\gamma})$.

Then, as $N\to\infty, n\to\infty$ such that $(N/n)<\infty$ the estimator $(\hat{\theta}',\hat{\gamma}')$ converges in probability to the solution of the equations 15(a)-15(b), 18(a)-18(b).

The theorem is proved in Appendix B.

Next we establish the consistency and asymptotic normality of the estimator $\hat{\xi}=(\hat{\theta}',\hat{\gamma}')'$. For this, note first that the equations (15a)-(15b), (18a)-(18b) can be written as $\dfrac{1}{n}\sum_{i=1}^{n}\varphi(R_i,y_i,z_i;\beta,\phi,\gamma)=0$, where as before, $\theta=(\beta,\phi)$ and R_i is the response indicator. Denote also by $\tilde{\xi}=(\tilde{\theta}',\tilde{\gamma}')'$ the true vector parameter. In the theorem below all the expectations are taken over all possible samples of respondents and all possible outcomes under the sample distribution.

Theorem 2: Suppose that:

(i) $\tilde{\xi}$ is an interior point of the parameter space, (ii) $\varphi(R_i,y_i,z_i;\xi)$ is continuously differentiable in a neighborhood δ of $\tilde{\xi}$, (iii) $E[\varphi(R_i,y_i,z_i,\tilde{\xi})]=0$ and $\tilde{\xi}$ is the unique solution of the

equations $E[U(\xi)]=0$, (iv) $E[\varphi(R_i,y_i,z_i,\tilde{\xi})\varphi(R_i,y_i,z_i,\tilde{\xi})']<\infty$, (v) $E\left[\sup_{\xi\in\delta}\left\|\frac{\partial\varphi(y_i,z_i,\xi)}{\partial\xi}\right\|\right]<\infty$.

Then the estimator $\hat{\xi}=(\hat{\theta},\hat{\gamma})$ is consistent for $\tilde{\xi}=(\tilde{\theta},\tilde{\gamma})$ and $\sqrt{n}(\hat{\xi}-\tilde{\xi})\xrightarrow{D}N[0,V(\tilde{\xi})]$.
The theorem is proved in Appendix C.

Another possibility of utilizing known covariate totals for estimating the parameters γ governing the model for the response probabilities is by applying an approach proposed by Chang and Kott (2008). By this approach, the H-T estimators of totals of calibration variables $C_1,...,C_K$, which may contain some or all of the covariates in the response model are regressed against their known population totals. Thus, in the case that the probability to respond depends on the outcome variable and $q+1$ covariates (including an intercept), the method requires that $K\geq q+2$. The major difference between the calibration equations in (18) and this method is that it allows utilizing more population totals than the totals of the variables included in Z. In particular, population totals of variables not included in the model for the response probabilities may be used. This results in more equations than estimated parameters and hence possibly more stable estimators.

Let c_i denote the values of the calibration variables for unit i. Chang and Kott (2008) estimate the unknown parameters by setting the nonlinear regression equations, $C^{pop}=\sum_{i=1}^{r}w_i\frac{c_i}{\pi(y_i,v_i;\gamma)}+\varepsilon^*$ where $C^{pop}=\sum_{j=1}^{N}c_j$ and ε^* is a vector of errors. The parameters γ are estimated by applying the iterative algorithm,

$$\hat{\gamma}^{(j+1)}=\hat{\gamma}^{(j)}+\left\{\hat{H}(\hat{\gamma}^{(j)})^T\hat{V}^{-1}(\hat{\gamma}^{(j)})\hat{H}(\hat{\gamma}^{(j)})\right\}^{-1}\hat{H}(\hat{\gamma}^{(j)})^T\hat{V}^{-1}(\hat{\gamma}^{(j)})(C^{pop}-\sum_{i=1}^{r}w_i\frac{c_i}{\pi(y_i,v_i;\hat{\gamma}^{(j)})}),\quad(19)$$

where $\hat{H}(\hat{\gamma}^{(j)})=\frac{\partial[\sum_{i=1}^{r}w_i\frac{c_i}{\pi(y_i,v_i;\gamma)}]}{\partial\gamma}\Big|\gamma=\hat{\gamma}^{(j)}$ and $\hat{V}^{-1}(\hat{\gamma}^{(j)})$ is the inverse of an estimator for the quasi-randomization variance of $\sum_{i=1}^{r}w_i\frac{c_i}{\pi(y_i,v_i;\gamma)}$, computed at $\gamma=\hat{\gamma}^{(j)}$.

Remark 3. Chang and Kott (2008) do not assume a model for the outcome so that their approach is restricted to estimation of the model of the response probabilities and it cannot be used for imputation. However, the following theorem holds (the proof can be obtained from the authors).

Theorem 3:

Let γ^* be the estimator obtained by application of (19), and θ^* be the solution of the equations $l(\theta, \gamma^*) = 0$ (Eqs. 15(a)-15(b) with $\gamma = \gamma^*$). Suppose that $0 < \pi(y_i, v_i; \gamma) < 1$ with bounded first derivatives with respect to γ. Then under some added regularity conditions $(\theta^*, \gamma^*) \xrightarrow{p} (\tilde{\theta}, \tilde{\gamma})$, and $\sqrt{n}(\theta^*, \gamma^*) \to N[(\tilde{\theta}, \tilde{\gamma}), \Sigma^*]$ for some fixed matrix Σ^*.

Remark 4. The obvious advantage of the use of (19) instead of (18a) and 18(b) for the estimation of (θ, γ) is that it does not require knowledge of the population totals of all the covariates featuring in the model for the response probabilities and at least one of the covariates featuring in the model for the outcome variable. On the other hand, our experience so far shows that the use of (18a) and 18(b) yields better parameter estimates and better imputations when the totals required for the use of these equations are known.

4. IMPUTATION AND ESTIMATION OF POPULATION MEANS

Denote by

$$\hat{f}_{R^c}(y_i \mid z_i) = \frac{\Pr(R_i = 0 \mid y_i, v_i, i \in S; \hat{\gamma}) f_S(y_i \mid x_i; \hat{\theta})}{\Pr(R_i = 0 \mid z_i, i \in S; \hat{\theta}, \hat{\gamma})}, \tag{20}$$

$$\hat{\pi}(y_i, v_i) = \pi(y_i, v_i; \hat{\gamma}), \hat{E}_{R^c}(Y_i \mid z_i) = E_{R^c}(Y_i \mid z_i; \hat{\theta}, \hat{\gamma}), \tag{21}$$

the estimated *pdf*s for the nonresponding units, the response probabilities and the expectations for the nonresponding units respectively. The expectation $E_{R^C}(Y_i \mid z_i)$ in (23) is with respect to the *pdf* $f_{R^c}(y_i \mid z_i)$. The estimates in (20) and (21) 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,

$$\hat{\bar{Y}}_1 = \frac{1}{N} \sum_{i=1}^{r} w_i y_i / \hat{\pi}(y_i, v_i). \tag{22}$$

When the covariates are known for all the sampled units, another estimator is obtained as,

$$\hat{\bar{Y}}_2 = \frac{1}{N} \sum_{i=1}^{n} w_i y_i^* \; ; \quad y_i^* = y_i \text{ if } i \in R \; , \; y_i^* = y_i^{imp} \text{ if } i \in R^c. \tag{23}$$

The imputed values can be computed either as,

$$y_i^{imp} = \hat{E}_{R^c}(Y_i \mid z_i), \tag{24}$$

or by generating one or more random observations from the *pdf* $\hat{f}_{R^c}(y_i \mid z_i)$ and taking the average of these observations as the imputed value, using multiple imputation techniques (Rubin, 1987, Schafer and Schenker, 2000).

Remark 5. It is important to emphasize that no model is assumed for the outcomes of the nonresponding units. This model is defined mathematically by the relationship (20). The sample model, $f_S(y_i \mid x_i;\theta)$, and the model for the response probabilities, $\pi(y_i, v_i;\gamma)$, define the model holding for the outcomes of the responding units and this model can be validated by application of goodness of fit test statistics since it refers to the observed data (see section 6).

The predictor $\hat{\bar{Y}}_{(2)}$ in (23) assumes that the covariates are known for every unit in the sample. When the covariates are only known for the respondents, we can first impute the missing covariates for the nonrespondents from the probability function $P_{Z|0}(z_i) = \Pr(Z_i = z_i \mid R_i = 0, i \in S)$, and then predict the outcome value as described above. By Sverchkov and Pfeffermann (2004), the latter probability function can be expressed as,

$$\begin{aligned} P_{Z|0}(z_i) &= \frac{P(R_i = 0 \mid Z_i = z_i, i \in S)}{P(R_i = 0 \mid i \in S)} \Pr(Z_i = z_i \mid i \in S) \\ &= \frac{P(R_i = 0 \mid Z_i = z_i, i \in S) \Pr(Z_i = z_i \mid R_i = 1, i \in S) \Pr(R_i = 1 \mid i \in S)}{P(R_i = 0 \mid i \in S) \Pr(R_i = 1 \mid Z_i = z_i, i \in S)}. \end{aligned} \tag{25}$$

Estimating $\hat{\Pr}(Z_i = z_i \mid R_i = 1,\ i \in S) = (1/r) \quad \forall i \in R$ and $\hat{\Pr}(R_i = 1 \mid i \in S) \quad = r / [\sum_{j=1}^{r} [1/\hat{\pi}(z_j)]$, the probability $P_{Z|0}(z_i)$ can be estimated as,

$$\hat{P}_{Z|0}(z_i) = \frac{[1 - \hat{\pi}(z_i)]}{\hat{\pi}(z_i)[\sum_{j=1}^{r} (1/\hat{\pi}(z_j)) - r]},\ z_i \in R. \tag{26}$$

Remark 6. The estimator (26) assumes that the covariates in the subsample of the nonrespondents take the same values as in the subsample of the respondents (although with different frequencies). Note that $\sum_{j=1}^{r} \hat{P}_{Z|0}(z_j) = 1$. When the dimension of z_i is small, the estimate $\hat{\Pr}(Z_i = z_i \mid R_i = 1, i \in S)$ can be enhanced by use of a 'smoothed' estimator, using more advanced density estimation methods.

5. ESTIMATION OF VARIANCES OF ESTIMATORS OF POPULATION MEANS

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 for the respondents (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 *B* of samples of outcomes of size r from the estimated respondents' distribution $f_R(y_i \mid z_i; \hat{\theta}, \hat{\gamma})$ with fixed (original) covariates z_i.
2. For each new sample, re-estimate (θ, γ) and then compute the estimators $\hat{\bar{Y}}_{(1)}$ and $\hat{\bar{Y}}_{(2)}$ using the new parameter estimators.
3. Estimate,

$$\hat{Var}(\hat{\bar{Y}}_{(k)}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\bar{Y}}_{(k)}^{(b)} - \bar{Y}_{(k)})^2;\ \bar{Y}_{(k)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\bar{Y}}_{(k)}^{(b)},\ k = 1, 2, \tag{27}$$

where $\hat{\bar{Y}}_{(k)}^{(b)}$ denotes the estimators obtained for bootstrap sample $b = 1, ..., B$.

For estimating the unconditional variances we first impute the missing covariates for the nonrespondents, if they are missing, using Eq. (26). Next we generate the outcomes for the whole sample using the estimated sample distribution, $f_S(y_i \mid x_i; \hat{\theta})$ and then select respondents with probabilities $\pi(y, v_i; \hat{\gamma})$. In this case the number of respondents and their covariates change from one bootstrap sample to the other. The whole process is repeated *B* times. The rest of the computations are the same as for the conditional variances.

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

$$Var(\hat{\bar{Y}}_{(1)}) = Var[\tilde{\bar{Y}}_{(1)} \mid \tilde{T}_{v,x} = (N, V_1^{pop}, ..., V_q^{pop}, \hat{\tilde{\beta}}' \tilde{X}^{pop})'\,]; \tag{28}$$

$$\tilde{\bar{Y}}_{(1)} = \frac{1}{N}\sum_{i=1}^{r}\frac{w_i y_i}{\pi(y_i, v_i; \gamma)},\ \tilde{T}_{v,x} = [\sum_{i=1}^{r} w_i \frac{1}{\pi(y_i, v_i; \gamma)}, \sum_{i=1}^{r} w_i \frac{v_{1i}}{\pi(y_i, v_i; \gamma)}, ..., \sum_{i=1}^{r} w_i \frac{v_{qi}}{\pi(y_i, v_i; \gamma)}, \sum_{i=1}^{r} w_i \frac{\hat{\tilde{\beta}}' \tilde{x}_i}{\pi(y_i, v_i; \gamma)}]'.$$

This variance accounts for the calibration equations used for estimating the model parameters (Eqs. 18(a) and 18(b)) and hence the response probabilities. (Deville and Tille, 2005 propose a

similar variance estimator in a different context). Denote $\sigma_{11} = Var(\tilde{\tilde{Y}}_{(1)})$, $\Sigma_{22} = Var(\tilde{T}_{v,x})$ and $\sigma'_{12} = Cov(\tilde{\tilde{Y}}_{(1)}, \tilde{T}_{v,x})$. Assuming $\tilde{\tilde{Y}}_{(1)} \cong \delta' \tilde{T}_{v,x} + \varepsilon,\ E(\varepsilon \mid \tilde{T}_{v,x}) = 0$ for some vector δ, (e.g., by assuming asymptotic normality of $(\tilde{\tilde{Y}}_{(1)}, \tilde{T}_{v,x})$),

$$Var(\hat{\tilde{Y}}_{(1)}) = \sigma_{11} - \sigma'_{12}\Sigma_{22}^{-1}\sigma_{12}\,. \tag{29}$$

The variance components in (29) and hence the variance of the estimator $\hat{\tilde{Y}}_{(1)}$ can be computed and estimated with respect to the randomization distribution over all possible samples of respondents, or over all possible samples of respondents and all possible outcomes under the sample model, with the unknown model parameters replaced by their original sample estimates. The apparent advantage of the estimator (29) is that it does not require resampling procedures.

Remark 7. In principle, the variance of $\hat{\tilde{Y}}_{(2)}$, which uses observed and imputed values can be estimated also using the multiple imputation theory (Rubin, 1987, Schafer and Schenker, 2000). However, empirical results obtained so far show that a textbook application of this method in the present context does not produce well behaved estimators, requiring some extra adjustments that are still under investigation.

6. TESTING THE GOODNESS OF FIT OF THE MODEL

As noted before, the *pdf* (11), which is fitted for the responding units can be validated (tested) since it refers to the observed data. In fact, one faces the classical problem of having a random sample from a hypothesized *pdf* which has to be validated. In what follows we consider several goodness of fit test statistics that seem appropriate for our problem.

6.1. *Classical Tests*

Suppose first that the true model parameters (θ, γ) are known. Denote by $U_i(y) = \int_{-\infty}^{y} f_R(t \mid z_i; \theta, \gamma)dt$ the hypothesized cumulative sample distribution function (*cdf*) of $y_i \mid z_i$, $i = 1,...,r$. For an absolutely continuous *cdf* the random variables $U_i(\cdot)$ are independent Uniform [0,1] variables since the responses y_i are independent given the covariates z_i. Denote by $u_1,...,u_r$ the values of $U_1,...,U_r$ at the sample values $y_1,...,y_r$ respectively, and let

F_{emp} define the empirical distribution of $u_1,...,u_r$. Following Landsman (2008), we apply three classical goodness of fit tests to the ordered values $u_{(1)},...,u_{(r)}$. The tests are:

$$\text{Kolmogorov-Smirnov:} \quad KS = \max_i | F_{Emp}(u_{(i)}) - u_{(i)} |, \tag{30}$$

$$\text{Cramer-von Misses:} \qquad CM = \frac{1}{12r} + \sum_{i=1}^{r} [u_{(i)} - \frac{2i-1}{2r}]^2, \tag{31}$$

$$\text{Anderson-Darling:} \quad AD = -r - \frac{1}{r}\sum_{i=1}^{r} [(2i-1)\ln(u_{(i)}) + (2r+1-2i)\ln(1-u_{(i)})]. \tag{32}$$

As discussed in Babu and Feigelson (2006), the KS statistic is sensitive to large-scale differences in location and shape between the model and the empirical distribution, the CM statistic is sensitive to small-scale differences in the shape and the AD statistic is sensitive to differences near the tails of the distribution.

So far we assumed known parameter values. When the test statistics are computed with estimated parameters, the asymptotic distribution of the three statistics depends in a complex way on the hypothesized model, the true model parameters and the method of estimation. Correct critical values can be obtained in this case by use of parametric bootstrap. The procedure consists of generating a large number of samples from the estimated hypothesized model, re-estimating the unknown parameters from each bootstrap sample and then computing the corresponding test statistics. The bootstrap distribution of these statistics provides approximate critical values for the null distribution with correct order of error. See Babu and Rao (2004) for regularity conditions validating the use of this procedure.

6.2. *Other Tests*

In addition to the classical tests considered above, we propose additional tests that compare the theoretical moments of the fitted distributions with their HT estimators. In what follows we illustrate the use of these statistics for the case where the population *pdf* is normal, but the tests can be modified to other population distributions.

Under normality of the population *pdf*, $Y_i = x_i'\beta + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and we can test, for example, $H_0^{(3)} : \mu^{(3)} = E(\varepsilon^3) = 0$, or $H_0^{(4)} : \mu^{(4)} = E(\varepsilon^4) = 3$, using the following test statistics:

$$H_0^{(3)}: \; C_{(3)} = \frac{1}{N\hat{\sigma}^3\sqrt{\hat{V}^{(3)}}} \sum_{i=1}^{r} w_i \frac{(y_i - x_i'\hat{\beta})^3}{\pi(y_i, v_i; \hat{\gamma})}, \tag{33}$$

$$H_0^{(4)}: C_{(4)} = \frac{1}{N\hat{\sigma}^4}[\frac{1}{\sqrt{\hat{V}^{(4)}}}\sum_{i=1}^{r} w_i \frac{(y_i - x_i^t\hat{\beta})^4}{\pi(y_i, v_i; \hat{\gamma})} - 3], \quad (34)$$

where $\hat{V}^{(3)} = \hat{Var}(\sum_{i=1}^{r} w_i \frac{(y_i - x_i'\hat{\beta})^3}{\pi(Y_i, v_i; \hat{\gamma})})$ and $\hat{V}^{(4)} = \hat{Var}(\sum_{i=1}^{r} w_i \frac{(y_i - x_i'\hat{\beta})^4}{\pi(y_i, v_i; \hat{\gamma})})$ are the conditional variances given the calibration constraints 18(a)-18(b). Critical values for the test statistics $C_{(3)}$ and $C_{(4)}$ can be obtained by parametric bootstrap, similarly to the procedure described in Section 6.1. Alternatively, for large r one can use the standard normal approximation by application of an appropriate central limit theorem.

7. APPLICATION OF METHODS TO HOUSEOLD EXPENDITURE SURVEY

7.1 *Study Population and Outcome Variable*

In this section we illustrate and study the performance of the proposed approach by using data collected as part of the Household Expenditure Survey (HES) carried out by the Israel Central Bureau of Statistics in 2005. The survey collects information on socio-demographic characteristics of each member of the sampled households (HH), as well as information on the HH income and expenditure. The HHs were sampled with equal probabilities by a two-stage sampling design. The initial response rate in this survey was as low as 43%, but after many recalls it increased to 90% of the sampled HHs. 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. After removing 4 HHs as outliers, the total sample size is $n = 1717$, with $r = 629$ responding HHs and $n - r = 1088$ nonresponding HHs, so that for our sample the response rate is 37%. The head of the HH is the member with the highest income among its 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. This allows comparing the imputed values with the corresponding true values, assuming that the reported incomes are not affected by being collected at a later interview. As noted above, the HHs were sampled with equal probabilities

and we assume therefore that the population model and the sample model under full response are the same.

7.2 *Sample Model and Response Probabilities*

We assume (and validate in Section 7.5) that the sample distribution of the outcome (under full response) given the covariates is *lognormal*, and that the response probabilities given the outcome and the covariates can be modeled by the *logistic* function, that is,

$$y_i = x_i'\beta + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma_\varepsilon^2), \tag{35}$$

$$P(R_i = 1 \mid y_i, v_i) = [1 + e^{-(\gamma_0 y_i + \gamma_1' v_i)}]^{-1}, \tag{36}$$

where y_i is the log(income) per standard person in household i and x_i and v_i are the corresponding vector covariates. The covariates include characteristics of the head of the HH: gender, age, occupation ('Occ.') and number of years at school ('Sch.'), as well as HH characteristics: number of earners ('earners'), HH size ('HHsize') and district of residence ('Dist.'). Most of the covariates included in the sample model (35), and in particular the outcome variable log(income) are nonsignificant when included in the response model (36). However, removing the nonsignificant covariates from the model makes the log(income) variable significant and the resulting model contains much fewer covariates.

Tables 1 and 2 show the estimated coefficients of the models (35) and (36) as obtained when fitting the models separately to all the sample data (respondents and 'nonrepondents'), and when fitting the respondents' model (11) to only the responding units, using the algorithm described in Section 3.4. For the application of the algorithm we took the true population totals of the covariates included in the logistic response model to be the corresponding sample totals.

The values of the coefficients in the two tables show that they can be estimated sufficiently accurately based only on the model holding for the responding units. When fitting the sample model (Eq. 35) to all the sample data, we obtained $R^2 = 0.612$ with residual variance $\hat{\sigma}_\varepsilon^2 = 0.394$. The estimator of σ_ε^2 from fitting the respondents model is $\hat{\sigma}_\varepsilon^2 = 0.393$. 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.

Table 1: Sample model when fitted to all sampled HH (Respondents and Nonrespondents), and when fitting the respondents' model to the responding HHs.

Coeff.	Const.	Gender	Age	Dist. 21	Dist.41	Dist.42	Dist. 43	Dist. 44
All HH	7.29	-0.12	0.02	-0.18	0.16	0.13	0.19	0.18
Respond.	7.18	-0.13	0.02	-0.10	0.14	0.10	0.19	0.16

Coeff.	Earners	HHsize	Occ.0	Occ.1	Occ.4	Sch.10	Sch.12	Sch.16
All HH	0.25	-0.14	0.44	0.23	0.15	-0.36	-0.14	0.16
Respond.	0.27	-0.14	0.45	0.26	0.15	-0.36	-0.14	0.19

Table 2: Model for response probabilities when fitted to all sampled HH (Respondents and "Nonrespondents"), and and when fitting the respondents' model to the responding HHs.

Coeff.	Const.	$Log(y)$	Gender	Dist.43	Dist.44	Dist.53	HHsize
All HH	1.00	-.21	-0.21	0.86	-0.58	-0.77	0.10
Respond.	1.35	-.26	-0.20	0.90	-0.59	-0.79	0.12

Figure 1 compares the empirical distribution of the estimated sample model residuals with the normal distribution with mean zero and same standard deviation, $\hat{\sigma}_{\varepsilon}^2 = 0.394$. The distribution of the estimated residuals is seen to be close to the normal distribution, although with somewhat shorter tails, which can be explained by the fact that the estimated residuals are not independent. The normality assumption is tested and validated in section 7.5.

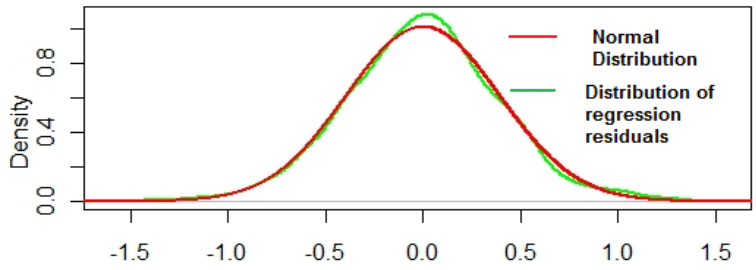

Figure 1: Distribution of estimated regression residuals and normal distribution with mean zero and same variance ($\hat{\sigma}_{\varepsilon}^2 = 0.394$).

7.3 *Imputation of Missing Outcomes*

Next we show the performance of the proposed approach in imputing the missing incomes. 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 *pdf* $\hat{f}_{R^c}(y_i \mid z_i) = f(y_i \mid z_i, i \in S, R_i = 0; \hat{\beta}, \hat{\sigma}_\varepsilon^2, \hat{\gamma})$ (Eq. 24). We imputed 5 values for each unit and 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. 26, 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, that is, when imputing the missing covariates by drawing at random from their empirical distribution for the responding HHs and imputing the missing incomes given the covariates by drawing at random from the estimated sample distribution.

Figures 2 and 3 show that application of our 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.

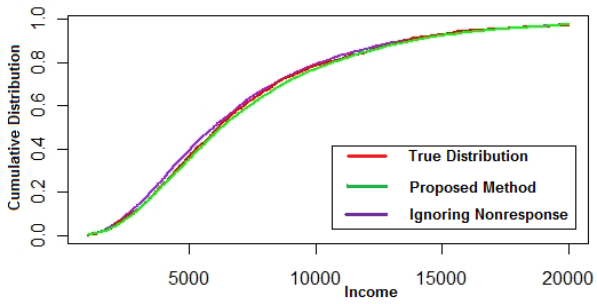

Figure 2: True empirical cumulative distribution and means of estimated empirical cumulative distributions of the incomes over 5 imputation sets. Known covariates.

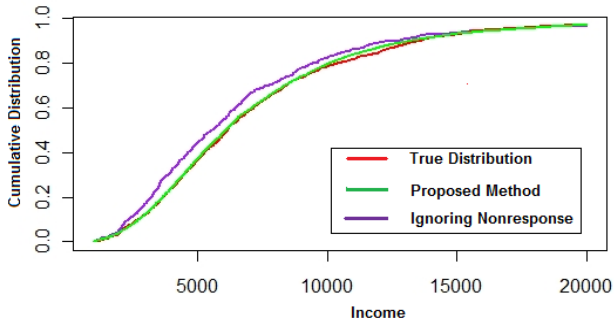

Figure 3: True empirical cumulative distribution and means of estimated empirical cumulative distributions of the incomes over 5 imputation sets. Missing covariates.

It is important to mention 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 process cannot be ignored for some of the covariates. For example, Table 3 shows the percentage of HHs by size for the responding and nonresponding units. The HH size is an important covariate in both the models (35) and (36) (Tables 1 and 2).

Table 3: Distribution of HH size in sub-samples of responding and nonresponding HHs

HH size	1	2	3	4	5	6+
Respond.	6.18	13.63	19.33	26.94	20.60	13.31
NonRespond.	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 considered two 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 (Std) when conditioning on the observed covariates for the respondents (and hence also on the number of respondents). Table 5 presents the unconditional Std 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 that uses the 'true' probabilities to respond, $\pi(y_i, v_i; \hat{\gamma})$, that is, when the probabilities to respond are not re-estimated 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 $\hat{\bar{Y}}_{(2,C-UK)}$).

Table 4: Estimation of sample mean of income (True $\bar{Y} = 7215.06$). Conditional Std. 500 bootstrap samples.

Estimator	Estimate		Standard Error
	Original sample of respondents	Mean over bootstrap samples	
$\hat{\bar{Y}}_{(1,P-K)}$	----	7303.65	174.16
$\hat{\bar{Y}}_{(1)}$	7332.30	7299.17	147.38
$\hat{\bar{Y}}_{(2,C-UK)}$	7311.06	7297.09	146.58
$\hat{\bar{Y}}_{(2,C-K)}$	7272.26	7265.53	140.81

Table 5: Estimation of sample mean of income (True $\bar{Y} = 7215.06$). Unconditional Std. 500 bootstrap samples.

Estimator	Estimate		Standard Error
	Original sample of respondents	Mean over bootstrap samples	
$\hat{\bar{Y}}_{(1,P-K)}$	----	7246.26	347.39
$\hat{\bar{Y}}_{(1)}$	7332.30	7248.88	179.83
$\hat{\bar{Y}}_{(2,C-UK)}$	7311.06	7308.37	152.43
$\hat{\bar{Y}}_{(2,C-K)}$	7272.26	7304.99	148.00

Tables 4 and 5 illustrate that all the estimators of the mean population income overestimate the true mean, but with the largest bias in the two tables being 1.6%. In comparison, the mean of the incomes computed from only the responding units is 6822.42, an underestimation of 5.4%. 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 (the estimator $\hat{\bar{Y}}_{(2,C-K)}$) than in the case that they have to be imputed (the estimator $\hat{\bar{Y}}_{(2,C-UK)}$). Finally, 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.

For estimating the unconditional standard error of the H-T estimator $\hat{\bar{Y}}_{(1)}$ we also computed for each of the 500 bootstrap samples the estimator (29), using the distribution over all possible samples of respondents and all possible outcomes. The mean of the Std estimators turned out to be 184.78, which is very close to the empirical standard error of 179.83 over all the bootstrap samples. The standard error estimator based on the original sample is 185.24.

7.5 *Testing the model assumptions*

In this section we study the performance of the test statistics (30)-(34) by considering three different combinations of the true distribution of the sample model residuals and the fitted (assumed) model:

I- The true residual distribution is $N(0,\hat{\sigma}_\varepsilon^2)$ (Model 1), and the fitted distribution is $N(0,\sigma_\varepsilon^2)$.

II- The true residual distribution is a mixture of $N(0.5\hat{\sigma}_\varepsilon^2,\hat{\sigma}_\varepsilon^2)$ and $N(-0.5\hat{\sigma}_\varepsilon^2,0.5\hat{\sigma}_\varepsilon^2)$ with equal probabilities (Model 2), while the fitted distribution is $N(0,\sigma_\varepsilon^2)$.

III- The true residual distribution is a mixture of $N(0.7\hat{\sigma}_\varepsilon^2,0.51\hat{\sigma}_\varepsilon^2)$ and $N(-0.7\hat{\sigma}_\varepsilon^2,0.51\hat{\sigma}_\varepsilon^2)$ with equal probabilities (Model 3), while the fitted distribution is $N(0,\sigma_\varepsilon^2)$.

For all the three cases we sampled the respondents using the logistic model, which was assumed also under the misspecified distributions. Figure 3 shows the three true sample models.

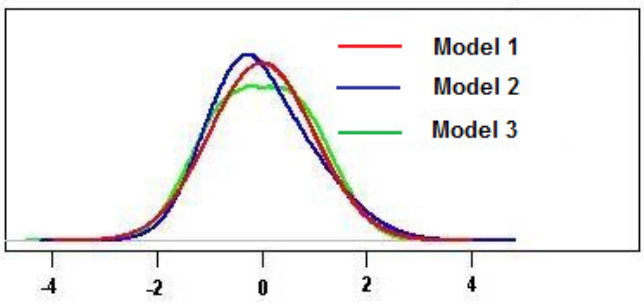

Figure 4: Sample models used for studying the performance of the test statistics

In order to study the performance of the test statistics we performed the following experiment for each of the three models:

1. Generate 250 new samples of outcomes from the true respondents' distribution under the model with parameters $(\hat{\theta},\hat{\gamma})$ and fixed (original) covariates z_i.

2. For each new sample:

2i- Re-estimate (θ,γ) assuming that the true sample distribution is normal,

2ii- Compute the test statistics (30)-(34),

2iii- Generate 250 new samples from the respondents' distribution assuming that the sample distribution is normal, using as parameters the estimates from (2i). Then,

3- For each new sample generated in 2iii,

3i- Re-estimate (θ,γ) assuming that the true sample distribution is normal and compute the test statistics (30)-(34).

3ii- Compute the distribution of each test statistic based on the 250 values in 3i.

Table 6 compares the empirical distribution of the five test statistics under Model 1 as obtained in Step 2ii, with the nominal values computed in Step 3ii. Denoting the ordered values of any one of the test statistics obtained in Step 3ii by $u_{(1)}<\ldots<u_{(250)}$, the critical value for nominal level α_j was defined as $u_{(250\alpha_j)}$ when $250\alpha_j$ is an integer, and $u_{[250\alpha_j]+1}$ otherwise, where $[\cdot]$ defines the integer number. The value of any given statistic in a cell corresponding to nominal level α_j is the percentage of samples that the statistic was between the critical values for nominal levels α_{j-1} and α_j ($\alpha_0=0$).

Table 6: Empirical and theoretical distribution of test statistics under Model 1

Test	Nominal levels											
	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	1.00
KS	0.040	0.046	0.088	0.079	0.109	0.097	0.112	0.096	0.146	0.097	0.039	0.053
AD	0.048	0.046	0.068	0.095	0.093	0.113	0.109	0.144	0.074	0.109	0.043	0.058
CM	0.056	0.031	0.084	0.110	0.072	0.080	0.125	0.133	0.097	0.101	0.059	0.051
$C_{(3)}$	0.043	0.035	0.088	0.120	0.109	0.120	0.088	0.107	0.079	0.102	0.055	0.053
$C_{(4)}$	0.042	0.043	0.060	0.126	0.098	0.135	0.143	0.108	0.093	0.092	0.027	0.053

In general, the empirical distribution of all the statistics is sufficiently close to the nominal values, thus validating the parametric bootstrap procedure described above for calculating the critical values under an assumed model. The goodness of fit of the empirical distributions to the nominal values was tested by the Pearson Chi-square statistic with 11 degrees of freedom, yielding p-values of 0.67 for KS, 0.45 for AD, 0.47 for CM, 0.89 for $C_{(3)}$ and 0.13 $C_{(4)}$.

Table 7 exhibits the proportion of samples that each of the test statistics rejects the misspecified distribution for the responding units, which assumes that the sample distribution is normal when in fact the true sample distribution is as defined under Model 2 or Model 3. For the test statistics defined by (30)-(32) we used one sided tests. For the test statistics defined by (33)-(34) we used two-sided tests. The proportions in Table 7 estimate the powers of the various tests in rejecting the misspecified model.

Table 7: Proportion of samples that each test statistic rejects the misspecified model for different nominal significance levels

Test	Model 2				Model 3			
	Significance level				Significance level			
	0.01	0.025	0.05	0.10	0.01	0.025	0.05	0.10
KS	0.832	0.892	0.936	0.960	0.245	0.549	0.637	0.775
AD	0.936	0.964	0.984	0.988	0.588	0.725	0.784	0.853
CM	0.924	0.948	0.980	0.988	0.490	0.696	0.765	0.843
$C_{(3)}$	0.876	0.932	0.956	0.984	0.000	0.000	0.020	0.088
$C_{(4)}$	0.112	0.188	0.264	0.356	0.480	0.647	0.716	0.823

When the true sample distribution is skewed as under Model 2, the three classical test statistics and the statistic $C_{(3)}$ that is designed for testing the skewness of the distribution have very good power properties, with powers higher than 0.9 for significance levels that are equal or higher than 0.025. As could be anticipated, the test $C_{(4)}$ that is designed to test the fourth moment has very low power in this case. The powers of all the test statistics except $C_{(4)}$ reduce when the true sample distribution is symmetric but flatter than the normal distribution, as under Model 3.

Nonetheless, all the test statistics except for $C_{(3)}$, and in particular AD and CM still have acceptable powers in this case for significance level equal or higher than 0.05.

The power of test statistics depends on the distance between the true model and the misspecified model and by Figure 4 the distances in our case are not really large. Although more work should be invested in developing new model testing techniques, the results in Tables 6 and 7 suggest that the goodness of fit of models fitted for the responding units can be tested adequately.

Finally, we applied the test procedure defined by Steps 2ii and 3ii in order to validate the normal/logistic model fitted to the original household expenditure data (Eqs. (35)-(36)). For this, we generated 500 samples from the fitted respondent's distribution using the parameter estimates in Tables 1 and 2 (estimated based only on the responding units). Table 8 exhibits the p-values obtained for the 5 test statistics.

Table 8: P-values when testing the model fitted to the original sample (Eqs. 37, 38)

Test	KS	AD	CM	$C_{(3)}$	$C_{(4)}$
p-value	0.262	0.098	0.122	0.256	0.108

With the usual type II error in mind and recalling the powers exhibited in Table 7 under the misspecified models, the p-values in Table 8 support the normal/logistic model fitted to the data, as already suggested by the other empirical results shown in previous sections.

APPENDIX A, SOLUTION OF EQUATIONS (15a)-(15b)

In order to solve the equations (15a), (15b) for given vector coefficient γ we use the Newton-Raphson algorithm. The second derivatives are as follows:

$$\frac{\partial l_j}{\partial \beta_k} = -a(\phi)\sum_{i=1}^{r} Var_R(Y_i \mid z_i;\beta,\phi,\gamma)x_{ji}x_{ki}, \quad j,k = 0,...,p$$

$$\frac{\partial l_j}{\partial \phi} = -\frac{\partial a(\phi)}{\partial \phi}\left\{\sum_{i=1}^{r}\left(Var_R(Y_i \mid z_i;\beta,\phi,\gamma)\sum_{s=0}^{p}\beta_s x_{si} + COV_R(Y_i, d(Y_i) \mid z_i;\beta,\phi,\gamma)\right)x_{ki}\right\} \; j = 0,...,p$$

$$\frac{\partial l_{p+1}}{\partial \beta_k} = -a(\phi)\sum_{i=1}^{r} COV_R(Y_i, d(Y_i) \mid z_i;\beta,\phi,\gamma)x_{ki}$$

$$\frac{\partial l_{p+1}}{\partial \phi} = -\frac{\partial a(\phi)}{\partial \phi}\left\{\sum_{i=1}^{r}\left(Var_R(d(Y_i)\mid z_i;\beta,\phi,\gamma) + COV_R(Y_i, d(Y_i)\mid z_i;\beta,\phi,\gamma)\sum_{s=0}^{p}\beta_s x_{si}\right)\right\}$$

where Var_R and Cov_R are the variance and covariance with respect to the distribution holding for the responding units. Denote,

$$X' = \begin{bmatrix} 1,\ldots,1\ ,\ 0,\ldots,0 \\ X_1,\ldots,X_r\ ,\ 0,\ldots,0 \\ 0,\ \ldots,0\ ,\ \ 1,\ldots,1 \end{bmatrix}, \quad Y' = [y_1,\ldots,y_r, d(y_1),\ldots,d(y_r)],$$

$$\mu' = \{E_R(Y_1\mid z_1),\ldots,E_R(Y_r\mid z_r),\ E_R[h(Y_1)\mid z_1],\ldots,E_R[h(Y_r)\mid z_r]\},$$

$$V_1 = Diag\left\{Var_R(Y_i\mid z_i)\right\},\ V_2 = Diag\left\{Var_R(d(Y_i)\mid z_i)\right\},\ i=1,\ldots,r$$

$$C = Diag\left\{COV_R((d(Y_i),Y_i)\mid z_i)\right\},\ S = Diag[\sum_{k=0}^{p}\beta_k x_{ki}],\ i=1,\ldots,r\ .$$

Application of the Newton-Raphson method yields,

$$\begin{pmatrix}\beta\\ \phi\end{pmatrix}^{(m+1)} = \begin{pmatrix}\beta\\ \phi\end{pmatrix}^{(m)} - [A^{-1}(X^T W X)^{-1} X^T (Y-\mu)], \tag{A1}$$

where A is a diagonal matrix of dimension $(p+2)(p+2)$, with all the elements on the main diagonal being $-a(\phi)$ except for the last element that is $-\frac{\partial a(\phi)}{\partial \phi}$ and $W = \begin{bmatrix} V_1 & C+SV_1 \\ C & V_2+SC \end{bmatrix}$.

APPENDIX B, PROOF OF THEOREM 1

We need to solve the equations $U(\xi)=0$, where $\xi' = (\theta,\gamma) = (\theta_0,\ldots,\theta_{p+1},\gamma_0,\ldots,\gamma_{q+1})$ and $U = (l',h')' = (l_1,\ldots,l_p,h_1,\ldots,h_q)'$. The functions $l(\xi) = [l_0(\xi),\ldots,l_{p+1}(\xi)]'$ are defined by (15a)-15(b) with $\theta = (\beta,\phi)$. The functions $h(\xi) = [h_1(\xi),\ldots,h_{q+2}(\xi)]'$ are defined by (18a)-(18b). The true vector parameter $\tilde{\xi} = (\tilde{\theta},\tilde{\gamma})$ is the unique solution of the estimating equations $E[U(\xi)]=0$, where the expectation is taken over all possible samples of respondents and all possible outcomes of the responding units under the sample distribution. We now show that the solution $\hat{\xi}$ of the algorithm of Section 3.4 converges in probability to the solution of the equations $U(\xi)=0$, which we denote by $\xi_0 = (\theta_0',\gamma_0')'$. Consider the familiar Newton-Raphson algorithm. Application of this algorithm to the present problem requires solving iteratively until convergence the equations,

$$\xi^{(m+1)} = \xi^{(m)} - A_m^{-1} U(\xi^{(m)}), \tag{B1}$$

where $\xi^{(m)}$ is the solution on the m^{th} iteration. The matrix A_m is defined as, $A_m = \begin{pmatrix} \frac{\partial l(\theta^{(m)}, \gamma^{(m)})}{\partial \theta^{(m)}} & \frac{\partial l(\theta^{(m)}, \gamma^{(m)})}{\partial \gamma^{(m)}} \\ \frac{\partial h(\theta^{(m)}, \gamma^{(m)})}{\partial \theta^{(m)}} & \frac{\partial h(\theta^{(m)}, \gamma^{(m)})}{\partial \gamma^{(m)}} \end{pmatrix}$, where $\frac{\partial l(\theta^{(m)}, \gamma^{(m)})}{\partial \theta^{(m)}}$ is the matrix of partial derivatives of $l(\theta, \gamma)$ with respect to θ evaluated at $(\theta^{(m)}, \gamma^{(m)})$, and similarly for the other block matrices. The estimation algorithm in Section 3.4 splits instead the system $U(\xi) = 0$ into the two systems, $l(\xi) = 0$ and $h(\xi) = 0$, and solves them iteratively until convergence as follows:

Apply one Newton-Raphson iteration to the equations $l(\xi) = 0$ with respect to θ for given γ, and one Newton-Raphson iteration to the equations $h(\xi) = 0$ with respect to γ for given θ, where the given values of θ and γ are the solutions from the previous iteration. The updating equations in this case can be written as,

$$\xi^{(m+1)} = \xi^{(m)} - B_m^{-1} U(\xi^{(m)}), \tag{B2}$$

where $B_m = \begin{pmatrix} \frac{\partial l(\theta^{(m)}, \gamma^{(m)})}{\partial \theta^{(m)}} & 0 \\ 0 & \frac{\partial h(\theta^{(m)}, \gamma^{(m)})}{\partial \gamma^{(m)}} \end{pmatrix}$. Splitting the system of equations into the two sub-systems is advantageous for large dimensional systems, since it saves the computation of $2pq$ (possibly complicated) partial derivatives. Define matrices M_{01} and M_{02} as follows:

$$M_{01} = \left[\frac{\partial l(\theta_0, \gamma_0)}{\partial \theta_0} \right]^{-1} \frac{\partial l(\theta_0, \gamma_0)}{\partial \gamma_0} \left[\frac{\partial h(\theta_0, \gamma_0)}{\partial \gamma_0} \right]^{-1} \frac{\partial h(\theta_0, \gamma_0)}{\partial \theta_0},$$

$$M_{02} = \left[\frac{\partial h(\theta_0, \gamma_0)}{\partial \gamma_0} \right]^{-1} \frac{\partial h(\theta_0, \gamma_0)}{\partial \theta_0} \left[\frac{\partial l(\theta_0, \gamma_0)}{\partial \theta_0} \right]^{-1} \frac{\partial l(\theta_0, \gamma_0)}{\partial \gamma_0}.$$

Suppose now that the conditions of Theorem 1 hold and that $\underset{N \to \infty, n \to \infty}{P\lim} \| M_{01} \| = \lambda_{01} < 1$, $\underset{N \to \infty, n \to \infty}{P\lim} \| M_{02} \| = \lambda_{02} < 1$, where $\|\cdot\|$ defines the Euclidian norm. We later check the fulfillment of the conditions for the model used for the empirical study in Section 7.

Proof of Theorem 1:

It is known that the Newton-Raphson algorithm has a quadratic rate of convergence, implying, $\left\|\varepsilon^{(m+1)}\right\| < c\left\|\varepsilon^{(m)}\right\|^2$, where $\varepsilon^{(m)} = (\xi^{(m)} - \xi_0)$ and c is a constant. It follows that,

$$\left\|\varepsilon^{(m)} - A_m^{-1}U(\xi^{(m)})\right\| < c\left\|\varepsilon^{(m)}\right\|^2. \tag{B3}$$

Next, rewrite the equations B(2) as $\xi^{(m+1)} = \xi^{(m)} - B_m^{-1}A_m A_m^{-1}U(\xi^{(m)})$. The rate of convergence of the proposed algorithm can be derived therefore as,

$$\varepsilon^{(m+1)} = \varepsilon^{(m)} - B_m^{-1}A_m A_m^{-1}U(\xi^{(m)}) = \varepsilon^{(m)} + B_m^{-1}A_m((\varepsilon^{(m)} - A_m^{-1}U(\xi^{(m)})) - \varepsilon^{(m)})$$

$$= (I - B_m^{-1}A_m)\varepsilon^{(m)} + B_m^{-1}A_m(\varepsilon^{(m)} - A_m^{-1}U(\xi_m)), \text{ or,}$$

$$\left\|\varepsilon^{(m+1)}\right\| \le \left\|I - B_m^{-1}A_m\right\| \cdot \left\|\varepsilon^{(m)}\right\| + \left\|B_m^{-1}A_m\right\| \cdot \left\|\varepsilon^{(m)} - A_m^{-1}U(\xi^{(m)})\right\|. \tag{B4}$$

By (B3), $\left\|\varepsilon^{(m+1)}\right\| \le \left\|I - B_m^{-1}A_m\right\| \cdot \left\|\varepsilon^{(m)}\right\| + c\left\|B_m^{-1}A_m\right\|\left\|\varepsilon^{(m)}\right\|^2$ and hence for $\varepsilon^{(m)}$ sufficiently small,

$$\left\|\varepsilon^{(m+1)}\right\| \le \left\|H_m\right\| \cdot \left\|\varepsilon^{(m)}\right\|, \tag{B5}$$

where $H_m = H(\theta^{(m)}, \gamma^{(m)}) = I - B_m^{-1}A_m$. It follows that,

$$\left\|\varepsilon^{(m+1)}\right\| \le \left\|H_{m-1}H_m\right\| \cdot \left\|\varepsilon^{(m-1)}\right\|. \tag{B6}$$

Now,

$$H_m = \begin{bmatrix} 0_{p\times p} & -\left[\frac{\partial l(\theta^{(m)}, \gamma^{(m)})}{\partial \theta^{(m)}}\right]^{-1} \frac{\partial l(\theta^{(m)}, \gamma^{(m)})}{\partial \gamma^{(m)}} \\ -\left[\frac{\partial h(\theta^{(m)}, \gamma^{(m)})}{\partial \gamma^{(m)}}\right]^{-1} \frac{\partial h(\theta^{(m)}, \gamma^{(m)})}{\partial \theta^{(m)}} & 0_{q\times q} \end{bmatrix} \quad \text{and let}$$

$H_0 = H(\theta, \gamma)\Big|_{(\theta_0, \gamma_0)}$. By Taylor expansion of H_m around (θ_0, γ_0), $H_m \approx H_0 + D_0 \mathrm{E}^{(m)}$, where the elements of D_0 and $\mathrm{E}^{(m)}$ are defined as follows: Let h_{ijk} be the derivative $\partial H_{ij} / \partial \xi_k$, where H_{ij} denotes the $(i, j)^{\text{th}}$ element of the matrix H_0. The matrix D_0 is obtained from H_0 by replacing each element H_{ij} by the row vector $(h_{ij0}, ..., h_{ijp}, h_{ij,p+1}, ..., h_{ij,p+q+3})$. The matrix $E^{(m)}$ is $E^{(m)} = I_{p+q+4} \otimes \varepsilon^{(m)}$ where $\otimes$ defines the Kronecker product. Note that $\dim(D_0) = (p+q+4) \times (p+q+4)^2$, and $\dim(\mathrm{E}^{(m)}) = (p+q+4)^2 \times (p+q+4)$. It follows that,

$\left\|\varepsilon^{(m+1)}\right\| \le \left\|H_{m-1}H_m\right\| \cdot \left\|\varepsilon^{(m-1)}\right\| \approx \left\|\left(H_0 + D_0\mathrm{E}^{(m-1)}\right)\left(H_0 + D_0\mathrm{E}^{(m)}\right)\right\| \cdot \left\|\varepsilon^{(m-1)}\right\| \approx \left\|H_0^2\right\| \cdot \left\|\varepsilon^{(m-1)}\right\|$, and hence by (B6),

$$\| \varepsilon^{(m+1)} \| \le \| H_0^2 \|^k \cdot \| \varepsilon^{(m-2k+1)} \|, \tag{B7}$$

where $H_0^2 = \begin{bmatrix} M_{01} & 0_{q\times p} \\ 0_{p\times q} & M_{02} \end{bmatrix}$. Since we assume, $\underset{N\to\infty,n\to\infty}{P\lim} \| M_{01} \| = \lambda_{01} < 1$, $\underset{N\to\infty,n\to\infty}{P\lim} \| M_{02} \| = \lambda_{02} < 1$, we obtain $\underset{N\to\infty,n\to\infty}{P\lim} \| H_0^2 \|^k = 0$ as $k \to \infty$, and

$$\| \varepsilon^{(m+1)} \| = \| \xi^{(m+1)} - \xi_0 \| \xrightarrow{p} 0 \quad \text{for } m - 2k \to \infty,\ k \to \infty, \tag{B8}$$

showing that the solution $\xi^{(m)}$ of the proposed algorithm converges in probability to ξ_0. *QED*

It remains to show that the conditions of the theorem are satisfied by the equations (15a)-(15b) and (18a)-(18b) as obtained for the model defined by (35)-(36). It is easy to show that in this case the functions $l(\theta,\gamma)$ and $h(\theta,\gamma)$ satisfy the conditions I) and III), provided that $0 < \pi(y_i, v_i; \gamma) < 1$ with bounded first derivatives. For example, in (38) $\pi(y_i, v_i; \gamma)$ is logistic and denoting $v_{0,i} = 1$, $v_{q+1,i} = y_i$, $\dfrac{\partial \pi(y_i, v_i; \gamma)}{\partial \gamma_l} = \dfrac{v_{li} e^{-(v_i'\gamma + \gamma_{q+1} y_i)}}{\left(1 + e^{-(v_i'\gamma + \gamma_{q+1} y_i)}\right)^2}$ for $l = 0,...,(q+1)$. In order to show that the norms of M_{01} and M_{02} converge in probability to limits smaller than 1, we assume (r1)- $\sum_{i=1}^{r} w_i = O(N^\delta)$ for $0.5 < \delta \le 1$, (r2)- $\sum_{j=1}^{N} w_j = O(N)$ and (r3)- $Var[\sum_{i=1}^{r} w_i \dfrac{x_{ki}}{\pi(y_i, v_i; \gamma)}] = O(N)$. These are standard requirements in sample surveys. As a simple example suppose that the sample is drawn by simple random sampling without replacement and the response probabilities are $\pi(y_i, v_i; \gamma) = (r/n)$. In this case, $w_i = (N/n)$ and $\sum_{j=1}^{N} w_j = N^2/n = O(N)$ since we assume that (N/n) is bounded. Clearly, $\sum_{i=1}^{r} w_i = r(N/n) = O(N^\delta)$ for $0.5 < \delta \le 1$ and the condition (r3) is also satisfied as long as $(N/r) = (N/n)(n/r)$ is bounded.

Suppose for convenience that $\dim(\mathrm{x}_i) = 2$ and $\dim(\mathrm{v}_i) = 1$. Then, $\theta = (\beta_0, \beta_1, \beta_2, \phi)'$ and $\gamma = (\gamma_0, \gamma_1, \gamma_2)'$. The matrix $\dfrac{\partial h(\theta,\gamma)}{\partial \theta}|_{(\theta_0,\gamma_0)}$ for the functions $h(\theta,\gamma)$ in (18a)-(18b) is then a 3×4 matrix with all of its elements equal to zero except for the $(3,3)^{\text{th}}$ element, which equals

$\sum_{i=1}^{r} w_i \frac{\tilde{x}_i}{\pi(y_i, v_i; \gamma)} - \tilde{X}^{pop}$. Next consider the derivatives of the functions in (18a)-(18b) with respect to γ. Denoting as above $v_{0,i} = 1$, $v_{q+1,i} = y_i$, we have that $\frac{\partial}{\partial \gamma_l} \sum_{i=1}^{r} w_i \frac{1}{\pi(y_i, v_i; \gamma)}$ $= -\sum_{i=1}^{r} w_i \frac{1}{\pi(y_i, v_i; \gamma)^2} \frac{\partial \pi(y_i, v_i; \gamma)}{\partial \gamma_l}$ $l = 0, ..., (q+1)$. Note that under the logistic model $\frac{1}{\pi(y_i, v_i; \gamma)^2} \frac{\partial \pi(y_i, v_i; \gamma)}{\partial \gamma_l} = O(1)$ and $\sum_{i=1}^{r} w_i = O(N^{\delta})$ by (r1) and therefore the whole expression is $O(N^{\delta})$. Some further algebra shows that the only nonzero entry of the matrix $\left[\frac{\partial h(\theta, \gamma)}{\partial \gamma}\right]^{-1} \frac{\partial h(\theta, \gamma)}{\partial \theta}$ is a constant times $N^{-\delta}[\sum_{i=1}^{r} w_i \frac{\tilde{x}_i}{\pi(y_i, v_i; \gamma)} - \tilde{X}^{pop}]$ and by Chebyshev inequality and the condition (r3), $P[N^{-\delta}(|\sum_{i=1}^{r} w_i \frac{\tilde{x}_i}{\pi(y_i, v_i; \gamma)} - \tilde{X}^{pop}| > \varepsilon)] \xrightarrow[N \to \infty]{P} 0$, if $\delta > 0.5$. It follows that all the elements of $\left[\frac{\partial h(\theta, \gamma)}{\partial \gamma}\right]^{-1} \frac{\partial h(\theta, \gamma)}{\partial \theta} \xrightarrow[N \to \infty]{P} 0$ and hence the norms of M_{01} and M_{02} converge in probability to limits smaller than 1, as assumed for the proof.

APPENDIX C, PROOF OF THEOREM 2

The consistency of the estimator follows from a result by Huber (1967), which states that under the conditions of the theorem, as $n \to \infty$ the estimator ξ_0 solving the equations $U(\xi) = 0$ is consistent for $\tilde{\xi}$, the solution of the equations $E[U(\xi)] = 0$. By Theorem 1, the estimator $\hat{\xi}$ converges in probability to ξ_0, establishing its consistency under the same conditions.

The asymptotic normality of the estimator $\hat{\xi}$ follows from a result by Newey and McFadden (1994), which states that under the same conditions,

$$\sqrt{n}(\hat{\xi} - \tilde{\xi}) \xrightarrow{D} N[0, V(\tilde{\xi})], \tag{C1}$$

with the variance matrix $V(\tilde{\xi})$ defined as $V(\tilde{\xi}) = A(\tilde{\xi})^{-1} B(\tilde{\xi})[A(\tilde{\xi})^{-1}]'$, where by abbreviating

$$\varphi_i = \varphi(R_i, y_i, z_i, \tilde{\xi}),\ A_n(\tilde{\xi}) = \frac{1}{n}\sum_{i=1}^{n} E(-\nabla\varphi_i), \qquad B_n(\tilde{\xi}) = \frac{1}{n}\sum_{i=1}^{n} E(\varphi_i\varphi_i'); \qquad A(\tilde{\xi}) = \lim_{n\to\infty} A_n(\tilde{\xi}),$$

$B(\tilde{\xi}) = \lim_{n\to\infty} B_n(\tilde{\xi})$ and $\nabla\varphi_i$ is the matrix of first derivatives of φ_i with respect to $\tilde{\xi}$.

The equations (15a)-(15b), (18a)-(18b) satisfy the conditions of the theorem, thus establishing the consistency and asymptotic normality of the estimator $\hat{\xi}$. *QED*.

8. REFERENCES

Babu, G.J., and Feigelson, E.D. (2006). Astrostatistics: goodness-of-fit and all that! In: *Astronomical Data Analysis Software and Systems XV*, ASP Conference Series, Eds. C. Gabriel, C. Arviset, D. Ponz and E. Solano **35**, pp. 127-136.

Babu, G.J., and Rao, C.R. (2004). Goodness-of-Fit tests when parameters are estimated. *Sankhya, Series* A, **66**, 63-74.

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.

Deville J.C., and Tille, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, **128**, 569-591.

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.

Huber, P.J. (1967). The behavior of maximum likelihood Estimates under nonstandard conditions. *Proceedings of the 5th Berkeley Symposium*, **1**, 221-233.

Landsman, V. (2008). Estimation of treatment effects in observational studies by fitting models generating the Sample Data. PHD Dissertation, Department of Statistics, 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*. New York: Wiley.

Newey, W.K., and McFadden, D. (1994). Large sample estimation and hypothesis testing. In: R.F Engle and D.L. McFadden, Editors. *Handbook of Econometrics,* **4**, 2111-2245.

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.

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.

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

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

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.

Schafer, J.L., and Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association*, **95**, 144-154.

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., Raghunathan, T.E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, **90**, 747.