Dealing with nonresponse in survey sampling: a latent modeling approach

Alina Matei* M. Giovanna Ranalli†

July 19, 2012

Abstract

Nonresponse is present in almost all surveys and can severely bias estimates. It is usually distinguished between unit and item nonresponse: in the former, we completely fail to have information from a unit selected in the sample, while in the latter, we observe only part of the information on the selected unit. Unit nonresponse is usually dealt with by reweighting: each unit selected in the sample has associated a sampling weight and an unknown response probability; the initial sampling weight is multiplied by the inverse of estimated response probability. Item nonresponse is usually dealt with by imputation. By noting that for a particular survey variable, we just have observed and unobserved values, in this work we exploit the connection between unit and item nonresponse. In particular, we assume that the factors that drive unit response are the same as those that drive item response on selected variables of interest. Response probabilities are then estimated by using a logistic regression with a latent covariate that measures such *will to respond* and that can explain part of the unknown behavior of a unit to participate in the survey. The latent covariate is estimated using latent trait models. Such approach is particularly relevant for sensitive items and, therefore, can handle non-ignorable nonresponse. Auxiliary information known for both respondents and nonrespondents can be included either in the latent variable model or in the logistic model. The approach can be also used when auxiliary information is not available, and we focus here on this case. The theoretical properties of the proposed estimators are sketched and simulations studies are conducted to illustrate their finite size sample performance.

Key words: unit nonresponse, item nonresponse, latent trait models, response propensity, non-ignorable nonresponse.

1 Introduction

Nonresponse is an increasingly common problem in surveys. It is a problem because, firstly, it causes missing data. It is usually distinguished between two types of missing data: *unit nonresponse*, when for a selected sample element the information is missing on all variables surveyed with the questionnaire, and *item nonresponse*, when the information is missing on at least one, but not all the survey variables. Nonresponse is a problem, secondly and more importantly, because such missingness is a potential and likely source of bias for survey estimates.

*Institute of Statistics, University of Neuchâtel, Pierre à Mazel 7, 2000, Neuchâtel, Switzerland, alina.matei@unine.ch and Institute of Pedagogical Research and Documentation Neuchâtel, Switzerland

†Dept. of Economics, Finance and Statistics, University of Perugia, Italy, giovanna.ranalli@stat.unipg.it

Randomness in survey sampling is guaranteed by a probabilistic sampling design. In the presence of unit nonresponse, another source of randomness is added, which is generated by an unknown response mechanism. The sampling design associates to each unit in the population an inclusion probability, whose inverse, for sampled units, determines its sampling weight. Each unit in the population has also associated a response probability to respond to the survey. The response probability is unknown and several methods have been proposed to estimate it either explicitly, using response propensity modeling like logistic regression models (see e.g. Kim and Kim, 2007), or implicitly, using response homogeneity groups or more generally calibration (see Särndal and Lundström, 2005, for an overview). Once estimates are computed, a commonly used method to deal with unit nonresponse is reweighting: sampling weights of the respondents are adjusted by the inverse of the estimated response probability providing new weights. The estimation of the response probabilities typically requires the availability of auxiliary information, either in the form of the value of some auxiliary variables for all units in the originally selected sample or of their population mean or total.

Three types of nonresponse are commonly considered in survey practice (see Little and Rubin, 1987): missing completely at random (MCAR), when the response indicator to the survey is independent of all other variables in the survey; missing at random (MAR), when the response indicator depends only on some characteristics observed in the survey and available also for nonrespondents; missing not at random (MNAR), when the response indicator depends on characteristics of interest observed only on the respondents or completely unobserved. Missing data mechanisms assuming an MCAR or a MAR response mechanism are called ignorable; all other mechanisms are non-ignorable.

We focus here on non-ignorable nonresponse, that is typical of surveys with sensitive questions (concerning drug abuse, sexual attitudes, politics, income etc). Various approaches have been proposed in survey sampling literature to deal with non-ignorable nonresponse. These approaches can be roughly divided into likelihood based methods and reweighting methods. Note that all these methods make use of observed auxiliary information. Survey problems with non-ignorable nonrespondents have been discussed e.g. by Greenlees et al. (1982), Little and Rubin (1987), Beaumont (2000), Qin et al. (2002), Zhang (2002). Copas and Farewell (1998) introduce into the British National Survey of Sexual Attitudes and Lifestyles, in addition to the variables of interest, a variable called 'enthusiasm-to-respond' to the survey, which is expected to be related to probabilities of unit and item response. They propose a method to estimate these probabilities using this variable to achieve unbiased estimates of population parameters. An approach based on the use of latent variables for modeling nonignorable nonresponse is given in Biemer and Link (2007), that extend the ideas in Drew and Fuller (1980) and use a discrete latent variable based on call history data available for all sample units. The latent variable is computed using some indicators of level of effort based on call attempts.

We focus on unit nonresponse and propose a method of reweighting to reduce nonresponse bias in the case of non-ignorable nonresponse. The method does not require the availability of auxiliary information, on the sample or population level, but different assumptions are made. First, it is assumed that item nonresponse is present in the survey and that it affects m variables of particular interest. Thus a response indicator can be defined for each variable (item) ℓ, for $\ell = 1, \ldots, m$, taking value 1 if item ℓ is observed on unit k and 0 otherwise. Next, the response indicators are manifestations of an underlying continuous scale which determines a latent variable that is related to the response propensity of the units and to the variable of interest. It is possible to compute such latent variable for all units in the sample, not only for the respondents, and thus to use it as auxiliary information in a response probability estimation procedure. The outcome of this estimation procedure can finally be used in a reweighting fashion. No information about the nonrespondents and no auxiliary information (like sex, age, marital status) are required in the proposed method.

The use of continuous latent variables to model item nonresponse is also considered in Moustaki and Knott (2000). In this paper, we take a different perspective and use latent variable models to address non-ignorable unit nonresponse also when auxiliary information is not available. We propose to use a latent variable called here 'will to respond to the survey', which is expected to be related to the probability of unit response, similar to the case of the 'enthusiasm-to-respond' variable as defined by Copas and Farewell (1998). This latent variable is suggestible to be computed in attitude surveys or surveys where the income is one of the variable of interest. As we have mentioned, our method does not need to use any auxiliary information about respondents and non-respondents, but we link unit nonresponse to item nonresponse via a latent variable and use this variable exactly like an auxiliary information in the process of estimating the response probabilities.

Following Moustaki and Knott (2000), '*weighting through latent variable modeling is expected to perform well under non-ignorable nonresponse where conditioning on observed covariates only is not enough.*' Moreover, in the absence of any covariate, we expect that an estimator based on the proposed weighting system with latent variable will perform better in terms of bias reduction than the naive estimator computed on the set of respondents.

The paper is organized as follows. Section 2 introduces the survey framework and notation, and reviews some classical estimators in the presence of nonresponse. In Section 3, we present how we can use latent variables as auxiliary information in nonresponse adjusted estimators. In Section 4, the practical properties of the proposed nonresponse adjusted estimators are evaluated via simulations. In Section 5, we summarize our conclusions.

2 Framework

Let U be a finite population of size N, indexed by k from 1 to N. Let s denote the set of sample labels, so that $s \subset U$. The sample size is denoted by n. The sample s is drawn using a probabilistic sampling design $p(s)$ such that

$$p(s) \geq 0, \text{ for all } s \in \mathcal{S} \text{ and } \sum_{s \in \mathcal{S}} p(s) = 1,$$

where $\mathcal{S}$ is the set of all samples s of size n. Let $\pi_k = \sum_{s; s \ni k} p(s)$ be the probability of including unit k in the sample. It is assumed that $\pi_k > 0$, $k = 1, \ldots, N$. Not all units selected in s respond to the survey. Denote by $r \subseteq s$ the set of respondents, and by $\bar{r} = s \setminus r$ the set of nonrespondents. The response mechanism is given by the distribution $q(r|s)$ such that for every fixed s we have

$$q(r|s) \geq 0, \text{ for all } r \in \mathcal{R}_s \text{ and } \sum_{s \in \mathcal{R}_s} q(r|s) = 1,$$

where $\mathcal{R}_s = \{r | r \subseteq s\}$.

Under unit nonresponse we define the response indicator $R_k = 1$ if unit $k \in r$ and 0 if $k \in \bar{r}$. Thus $r = \{k \in s | R_k = 1\}$. We assume that these random variables are independent of one another and of the sample selection mechanism; this is the pseudo-randomization model of Oh and Scheuren (1983). Since only the units in r are observed, a response model is used to estimate the probability of responding to the survey, $p_k = P(k \in r | k \in s) = P(R_k = 1 | k \in s)$, which is a function of the sample and must be positive. It follows that R_k are independent Bernoulli variables with expected value equal to p_k. We assume that the units respond independently of each other and of s and so

$$q(r|s) = \prod_{k \in r} p_k \prod_{k \in \bar{r}} (1 - p_k).$$

Suppose that in the survey there are m survey variables (items) of particular interest. Each respondent is exposed to these m questionnaire items, labelled $\ell = 1, \ldots, m$. Suppose that the goal is to estimate the population total of the variables of interest and, in particular, of the variable of interest y_j, i.e. $Y_j = \sum_{k=1}^{N} y_{kj}$, with y_{kj} being the value taken by y_j on unit k. In the ideal case, if the response distribution $q(r|s)$ is known, then the p_k's would be known and available to estimate Y_j in a two-phase fashion. Thus, the combined weights $1/(\pi_k p_k)$ could be computed for all $k \in r$ and used in the following Horvitz-Thompson (HT) estimator

$$\widehat{Y}_{j,\mathrm{2pHT}} = \sum_{k \in r} \frac{y_{kj}}{\pi_k p_k}. \tag{1}$$

It is unbiased for Y_j, under the quasi-randomization approach of Oh and Scheuren (1983), where the statistical properties are evaluated using the joint distribution of the sampling design and the response mechanism. Since the p_k's are unknown in practice, $\widehat{Y}_{j,\mathrm{2pHT}}$ cannot be computed.

The naive estimator

$$\widehat{Y}_{j,\mathrm{naive}} = N \Big(\sum_{k \in r} \frac{y_{kj}}{\pi_k} \Big) / \Big(\sum_{k \in r} \frac{1}{\pi_k} \Big), \tag{2}$$

does not adjust for nonresponse and is sometimes computed in surveys. Following Särndal and Lundström (2005, p. 68), its bias can very large, but "*may still find use if there is no better auxiliary information and/or if there is strong reason to believe that the nonresponse occurs at random.*" When this is not the case, the p_k's must be estimated and a nonresponse adjusted estimator is constructed by replacing p_k with an estimate $\widehat{p}_k$ in (1). One obtains a Horvitz-Thompson type estimator defined by

$$\widehat{Y}_{j,\mathrm{HT}} = \sum_{k \in r} \frac{y_{kj}}{\pi_k \widehat{p}_k}. \tag{3}$$

Different methods to compute $\widehat{p}_k$ have been proposed in the literature. All of these methods are based on the use of auxiliary information known on the population or sample level.

In the case of non-ignorable nonresponse the variable of interest is itself the cause (or one of the causes) of the response behavior, and a covariance between the former and the response probability is produced through a direct causal relation (see Groves, 2006). In such a case, the response probability p_k should be modeled for $k \in s$ using logistic regression as follows

$$p_k = P(R_k = 1|y_{kj}) = \frac{1}{1 + \exp(-(a_0 + a_1 y_{kj}))}, \tag{4}$$

or as follows

$$p_k = P(R_k = 1|y_{kj}, \mathbf{z}_k) = \frac{1}{1 + \exp(-(a_0 + a_1 y_{kj} + \mathbf{z}_k' \boldsymbol{\alpha}))}, \tag{5}$$

where $\mathbf{z}_k = (z_{k1}, \ldots, z_{kt})'$ is a vector with the values taken by $t \geq 1$ covariates on unit k, and a_0, a_1 and $\boldsymbol{\alpha}$ are parameters.

Nonresponse bias in the unadjusted respondent total of the variable of interest y_j depends on the covariance between the values y_{kj} and p_k (see Bethlehem, 1988). An example of covariate that reduces the covariance between y_{kj} and p_k is the interest in the survey topic, such as knowledge, attitudes, and behaviors related to the survey topic (see Groves et al., 2006). The set of covariates $\mathbf{z}_k$ should be also related to the variable of interest y_j to reduce sampling variance (Little and Vartivarian, 2005).

Since y_{kj} is only observed on the respondents, Models (4) and (5) cannot be estimated. Therefore, usually, the values of $\mathbf{z}_k$ that are known for both respondents and nonrespondents and are related to the y_{kj}'s by a 'hopefully strong regression' (Cassel et al., 1983) are used in the following model

$$p_k = P(R_k = 1|\mathbf{z}_k) = \frac{1}{1 + \exp(-(a_0 + \mathbf{z}_k' \boldsymbol{\alpha}))}. \tag{6}$$

Then, maximum likelihood can be used to fit Model (6) using the data $(R_k, \mathbf{z}_k)$ for $k \in s$. This leads to estimate $\widehat{a}_0$ and $\widehat{\boldsymbol{\alpha}}$ and to the estimated response probabilities $\widehat{p}_k = 1/(1 + \exp((-(\widehat{a}_0 + \mathbf{z}_k' \widehat{\boldsymbol{\alpha}}))))$ to be used in (3). Such procedure provides some protection against nonresponse bias if $\mathbf{z}_k$ is a powerful predictor of the response probability and/or of the variable of interest (Kim and Kim, 2007).

In what follows we propose an adjustment weighting system based on an auxiliary variable that measures the propensity of each unit to participate to the survey. To this end, further assumptions on the response model are introduced in order to assume a dependence of the p_k's on one latent auxiliary variable that is connected to the propensity scores of Rosenbaum and Rubin (1983). The proposed approach can be used when no other auxiliary information is available on $k \in s$.

3 Latent variables as auxiliary information

To obtain a measure of response propensities, we consider the case in which item nonresponse on the variables of interest is also present. Then, following Chambers and Skinner (2003, p.278): '*from a theoretical perspective the difference between unit and item nonresponse is unnecessary. Unit nonresponse is just an extreme form of item nonresponse*', we assume that item response on the variables of interest is driven on respondents by the same attitude and factors that drive unit response. Then, latent variable models can be used to estimate such factors and use them as covariates in a logistic response model. Suppose that item nonresponse affects the aforementioned m survey variables (items) of particular interest. Each respondent is exposed to such m questionnaire items, labelled $\ell = 1, \ldots, m$. A second response indicator is introduced for each item ℓ. For each item ℓ and each unit k, a binary variable $x_{k\ell}$ is defined that takes value 1 if unit k answers to item ℓ and 0 otherwise. Let $\mathbf{x}_k = (x_{k1}, \ldots, x_{k\ell}, \ldots, x_{km})'$ denote the vector of response indicators for unit k to the m items and let $\mathbf{y}_k = (y_{k1}, \ldots y_{k\ell}, \ldots, y_{km})'$ be the study variable vector for unit k. Thus $y_{k\ell}$ is the response value of unit k to item ℓ and $x_{k\ell}$ is the response indicator to the same item.

Suppose the $x_{k\ell}$'s are related to an assumed underlying latent continuous scale; they are the indicators of a latent variable denoted by θ_k. De Menezes and Bartholomew (1996) call the variable θ_k the 'tendency to respond' to the survey. We call it here the 'will to respond to the survey' of unit k. A latent trait model with a single latent variable is used to compute θ_k for each $k \in s$ (we will see later how; see Section 3.2). Assume for now that θ_k is known on all sample units and, as with usual auxiliary information, can be used as a covariate. Without any other covariate, Model (6) is rewritten as

$$p_k = P(R_k = 1|\theta_k) = \frac{1}{1 + \exp(-(\alpha_0 + \alpha_1 \theta_k))}. \tag{7}$$

The covariate θ_k can be viewed as a variable explaining the behavior related to the survey topic, and thus having good properties to reduce the covariance between y_{kj} and p_k and the nonresponse bias (see Groves et al., 2006).

Note that, in Model (7) auxiliary information takes the form of a latent variable. Of course, in case other suitable auxiliary information is available, it can be inserted in the model as supplementary covariates. Now, to estimate the parameters of Model (7) the value of θ_k has to be available for all units in the sample. The following sections provide details on how to obtain estimated values of θ_k for both respondents and nonrespondents.

3.1 Computing response propensities using latent trait models

The variable θ_k can be computed using a latent trait model. In general, latent variable models are multivariate regression models that link continuous or categorical responses to unobserved covariates. A latent trait model is a factor analysis model for binary data (see Bartholomew et al., 2002; Skrondal and Rabe-Hesketh, 2007).

We start by creating the matrix with elements $\{x_{k\ell}\}_{k\in s;\ell=1,\ldots,m}$. Figure 1 shows a schematic of the indicators $x_{k\ell}$ for respondents and nonrespondents.

$$\begin{array}{rc}
 & \text{items} \\
 & \downarrow \\
\text{units} \rightarrow & \left.\boxed{\begin{array}{cccc} 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ & & \vdots & \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{array}}\right\} r \\
\text{units} \rightarrow & \left.\boxed{\begin{array}{cccc} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}}\right\} \bar{r}
\end{array}$$

Figure 1: Schematic representing variables $x_{k\ell}$ for the sets r and $\bar{r}$

We assume that the factors that drive unit response are the same as those that drive item response on selected variables of interest. That is item nonresponse is assumed nonignorable. Let $q_{k\ell}$ be the probability of response of unit k for item ℓ, for all $\ell = 1, \ldots, m$ and $k \in r$. As in the case of unit nonresponse, $q_{k\ell}$ should be modelled as a function of the variable of interest using logistic regression as follows

$$q_{k\ell} = P(x_{k\ell} = 1 | y_{k\ell}, \theta_k) = \frac{1}{1 + \exp(-(\beta_{\ell 0} + \beta_{\ell 1}\theta_k + \beta_{\ell 2} y_{k\ell}))}, \tag{8}$$

for $\ell = 1, \ldots, m$, and $k \in r$. where $\beta_{0\ell}, \beta_{1\ell}$ and $\beta_{2\ell}$ are parameters. Since $y_{k\ell}$ is known only for units with $x_{k\ell} = 1, k \in r$, Model (8) cannot be estimated. As in the case of unit nonresponse, we propose to estimate $q_{k\ell}$ as a function of an auxiliary variable related to the variable of interest, that is θ_k. Model (8) is rewritten

$$q_{k\ell} = \frac{1}{1 + \exp\left(-(\beta_{\ell 0} + \beta_{\ell 1}\theta_k)\right)} \quad \text{for } \ell = 1, \ldots, m \text{ and } k \in r. \tag{9}$$

Note that Model (9) is not an ordinary logistic regression model, because the θ_k's are unobservable values taken by a latent variable. Latent trait models can be used in this case to estimate $q_{k\ell}$, θ_k and the model parameters. Note that in the area of educational testing and psychological measurement, latent trait modelling is termed Item Response Theory.

The Rasch model (Rasch, 1960) is a first simple latent trait model that is well known in the psychometrical literature and used for analyzing data from assessments to measure variables such as abilities and attitudes. It takes the following form

$$q_{k\ell} = \frac{1}{1 + \exp\left(-(\beta_{\ell 0} + \beta_{1}\theta_k)\right)} \quad \text{for } \ell = 1, \ldots, m \text{ and } k \in r. \tag{10}$$

The parameters $\beta_{\ell 0}$ are estimated for each item ℓ and reflect the extremeness (easiness) of item ℓ: larger values correspond to a larger probability of a positive response at all points in the latent space. The parameter β_1 is known as the 'discrimination' parameter and can be fixed to some arbitrary value without affecting the likelihood as long as the scale of the individuals' propensities is allowed to be free. In many situations the assumption that item discriminations are constant across items is too restrictive. The two-parameter logistic (2PL) model generalizes the Rasch model by allowing the slopes to vary. Specifically, the 2PL model assumes the form given in equation (9). The parameters $\beta_{\ell 1}$ are now estimated for each item ℓ and are typically assumed to be positive. They provide a measure of how much information an item provides about the latent variable θ_k. A further generalization to Model (9) is considered in the literature – the 3PL model – that includes another parameter, the *guessing* parameter, to model the probability that a subject with a latent variable tending to $-\infty$ responds to an item. Such extension does not seem necessary in the context at hand and will not be considered further.

Latent trait models typically rely on the following assumptions. The first one is the so-called *conditional independence* assumption, which postulates that item responses are independent given the latent variable (i.e. the latent variable accounts for all association among the observed variables x_ℓ). Following Bartholomew et al. (2002, p. 181) 'the assumption of conditional independence can only be tested indirectly by checking whether the model fits the data. A latent variable model is accepted as a good fit when the latent variables account for most of the association among the observed responses.'

A second assumption of Models (10) and (9) is that of *monotonicity*: as the latent variable θ_k increases, the probability of response to an item increases or stays the same across intervals of θ_k. In other words, for two values of θ_k, say a and b, and arbitrarily assuming that $a < b$, monotonicity implies that $P(x_{k\ell} = 1|\theta_k = a) < P(x_{k\ell} = 1|\theta_k = b)$ for $\ell = 1, \ldots, m$. Larger values of θ_k are associated with a greater chance of a response to each item.

Finally, the third, and possibly strongest, assumption of Models (10) and (9) is that of *unidimensionality*, implying that a single latent variable fully explains the willingness of unit k to answer the questionnaire. All these basic assumptions imply that the dependence between the variables $x_{k\ell}$ may be explained by the latent variable θ_k which represents the units' willingness and that the probability that an unit k responds to a given item increases with θ_k.

In what follow we use the two-parameter logistic (2PL) model given in (9). Let $\boldsymbol{\beta} = (\beta_{\ell 0}, \beta_{\ell 1})'$. Usually the following distributional assumption on θ_k is made: $\theta_k \sim N(0, 1)$. Given θ_k, the response indicators $x_{k\ell}$ are independent and their conditional probability for a given θ_k and $\boldsymbol{\beta}$ is

$$f(\mathbf{x}_k|\theta_k, \boldsymbol{\beta}) = \prod_{\ell=1}^{m} f_\ell(x_{k\ell}|\theta_k, \boldsymbol{\beta}) = \prod_{\ell=1}^{m} q_{k\ell}^{x_{k\ell}}(1 - q_{k\ell})^{1-x_{k\ell}},$$

where f_ℓ is the conditional probability of response indicator $x_{k\ell}$ for the ℓ-th item. The likelihood of $\mathbf{x}_k$ is given by

$$f(\mathbf{x}_k|\boldsymbol{\beta}) = \int_{-\infty}^{\infty} f(\mathbf{x}_k|\theta_k, \boldsymbol{\beta})h(\theta_k)d\theta_k,$$

where h is the standard normal density function. Conditional to the estimated value of θ_k, the probabilities $q_{k\ell}$ are estimated by the model using

$$q_{k\ell} = \int_{-\infty}^{\infty} f_\ell(x_{k\ell}|\theta_k, \boldsymbol{\beta})h(\theta_k)d\theta_k. \tag{11}$$

Different goodness-of-fit measures have been proposed in the literature (see Bartholomew et al., 2002). One of them is based on two-way and three-way margins of the response items. Discrepancies between the expected (E) and observed (O) counts in these tables are measured using the statistic $(O - E)^2/E$. Response items or pairs of response items where the model does not fit well are revealed using the table margins. Note that the residuals $(O - E)^2/E$ are not independent and they cannot be summed to give an overall test statistics distributed as a chi-squared (see Bartholomew et al., 2002, p. 186). This measure of goodness-of-fit is used in Section 4.

Assuming a simple random sampling (srs) without replacement of size n, Model (9) can be fitted by maximum likelihood, maximizing the following function assuming conditional independence among the response indicators $x_{k\ell}$:

$$\ln(\mathcal{L}) = \sum_{k\in r} \ln(f(\mathbf{x}_k)).$$

Bartholomew and Knott (1999) discuss an extended EM algorithm to estimate Model (9). In the case of non-srs, where sampling weights w_k are available for all respondents, a sample-weighted pseudo-log-likelihood can be defined as

$$\sum_{k\in r} w_k \ln(f(\mathbf{x}_k)).$$

Maximizing the pseudo-log-likelihood simultaneously with respect to $\beta_{\ell 0}, \beta_{\ell 1}$, and $x_{k\ell}$ provides design-consistent estimates of the underlying population parameters under suitable conditions (Pfeffermann, 1993).

Estimates $\widehat{\theta}_k$, $\widehat{\beta}_{\ell 0}$ and $\widehat{\beta}_{\ell 1}$ of θ_k, $\beta_{\ell 0}$ and $\beta_{\ell 1}$, respectively, are provided fitting Model (9). We use the estimate $\widehat{\theta}_k$ as covariate in Model (7) instead of the unknown value of θ_k. Model (7) is rewritten

$$p_k = P(R_k = 1|\widehat{\theta}_k) = \frac{1}{1+\exp(-(\alpha_0 + \alpha_1\widehat{\theta}_k))}. \tag{12}$$

It is important to test whether another latent variable may describe the responding willingness of units in a survey. In particular, other than the 'will-to-respond to the questionnaire', there may be also an 'ability' to answer to some items of the questionnaire for unit k, given, for example, by being more or less informed on the topic surveyed by those items. Model (9) can be extended to fit latent trait models with several latent variables, allowing also for the incorporation of nonlinear terms between them as discussed in Rizopoulos and Moustaki (2008). With two latent variables θ_k, γ_k and an interaction term, Model (9) is rewritten as

$$q_{k\ell} = \frac{1}{1+\exp\left(-(\beta_{\ell 0} + \beta_{\ell 1}\theta_k + \beta_{\ell 2}\gamma_k + \beta_{\ell 3}\theta_k\gamma_k)\right)}. \tag{13}$$

An alternative solution to estimate p_k is investigated in Matei (2010). Let $S_k = \sum_{\ell=1}^{m} x_{k\ell}$ be the raw score for unit k, i.e. the number of items unit k has responded to: if $k \in \bar{r}$, then $S_k = 0$;

if $k \in r$, then $S_k > 0$. Then p_k can be estimated by modelling $P(S_k > 0)$. By the conditional independence assumption we have

$$\begin{aligned} P(S_k > 0) &= 1 - P(S_k = 0) \\ &= 1 - P\left(\cap_{\ell=1}^{m}(x_{k\ell} = 0)\right) \\ &= 1 - \prod_{\ell=1}^{m}(1 - q_{k\ell}), \end{aligned}$$

where $x_{k\ell} \sim Bernoulli(q_{k\ell})$. The response probability p_k is then estimated by

$$\widehat{p}_k = 1 - \prod_{\ell=1}^{m}(1 - \widehat{q}_{k\ell}),$$

where $\widehat{q}_{k\ell}$ is computed using Model (9), for $\ell = 1, \ldots, m$. This solution has, however, a drawback: if there exists at least one $\widehat{q}_{k\ell}$ close to 1, then $\widehat{p}_k$ is close to 1, and there is no effect in the reweighting system.

3.2 Computing $\widehat{\theta}_k$ for non-respondents

To compute a value $\widehat{\theta}_k$ for $k \in \bar{r}$, we assume again that unit nonresponse is just an extreme form of item nonresponse. Thus, a nonrespondent does not answer to any item ℓ and thus $x_{k\ell} = 0$, for all $\ell = 1, \ldots, m$. Model (9) allows the computation of $\widehat{\theta}_k$ for this particular case, providing the same value of $\widehat{\theta}_k$ for all $\in \bar{r}$. Thus, Model (9) allows the computation of $\widehat{\theta}_k$ for all $k \in s$. This is a very important feature for the estimation of the response probabilities p_k using Model (12).

Sometimes the problem of separation is present, and Model (12) cannot be fitted directly. The phenomenon of separation is observed in the fitting process of a logistic regression model when the likelihood converges to a finite value while at least one parameter estimate diverges to (plus or minus) infinity. Separation occurs here since the zero cases ($R_k = 0$ for $\in \bar{r}$) have the same value of the covariate $\widehat{\theta}_k$. To overcome this problem two practical solutions can be used:

1. add a jitter to the value of $\widehat{\theta}_k$ (some noise to $\widehat{\theta}_k$ for the cases with $R_k = 0$),

2. add some artificial cases with $R_k = 0$ and with different values of θ_k randomly generated from $N(0, 1)$ distribution and then conduct the analysis in the usual fashion on the resulting data (see also Clogg et al., 1991).

3.3 Adjustment for item and unit nonresponse and corresponding estimators

Recall that we have a variable of particular interest y_j and that item nonresponse is present for it. Let $r_y = \{k \in r | x_{kj} = 1\}$ be the set of respondents for variable y_j. If we wish to estimate the population total $Y_j = \sum_{k=1}^{N} y_{kj}$ of y_j, then a naive estimator that does not correct neither for unit nor for item nonresponse is given by

$$\widehat{Y}_{j,\text{naive}} = N \sum_{k \in r_y} \frac{y_{kj}}{\pi_k} / \sum_{k \in r_y} \frac{1}{\pi_k}. \tag{14}$$

Imputation is often used to handle item nonresponse in survey practice. It consists in replacing missing values of items by imputed values. The naive estimator that uses imputed values y^*_{kj} of y_{kj} when $x_{kj} = 0$ is given by

$$\widehat{Y}^{\text{imp}}_{j,\text{naive}} = N \Big(\sum_{k \in r_y} \frac{y_{kj}}{\pi_k} + \sum_{k \in r \setminus r_y} \frac{y^*_{kj}}{\pi_k} \Big) / \Big(\sum_{k \in r} \frac{1}{\pi_k} \Big). \tag{15}$$

Reweighting item responders is another approach to handle item nonresponse. Moustaki and Knott (2000) propose to weight item responders by the inverse of the fitted probability of item response $\widehat{q}_{k\ell}$, assuming $\widehat{q}_{k\ell} > 0$. Therefore, a possible adjustment weight for item and unit nonresponse associated to unit $k \in r_y$ is given by $1/(\widehat{p}_k \widehat{q}_{kj})$. By considering different combinations of the aforementioned methods we propose the following estimators:

- a three-phase estimator adjusted for item and unit nonresponse via reweighting,

$$\widehat{Y}_{j,pq} = \sum_{k \in r_y} \frac{y_{kj}}{\pi_k \widehat{p}_k \widehat{q}_{kj}}; \tag{16}$$

- a two-phase estimator estimator adjusted for unit nonresponse via reweighting and for item nonresponse using imputation (y^*_{kj} for y_{kj} when $x_{kj} = 0$),

$$\widehat{Y}^{\text{imp}}_{j,p} = \sum_{k \in r_y} \frac{y_{kj}}{\pi_k \widehat{p}_k} + \sum_{k \in r \setminus r_y} \frac{y^*_{kj}}{\pi_k \widehat{p}_k}; \tag{17}$$

- a sort of model assisted compromise between estimators (16) and (17) given by

$$\widehat{Y}^{\text{imp}}_{j,pq} = \sum_{k \in r} \frac{y^*_{kj}}{\pi_k \widehat{p}_k} + \sum_{k \in r_y} \frac{y_{kj} - y^*_{kj}}{\pi_k \widehat{p}_k \widehat{q}_{kj}}. \tag{18}$$

  This estimator is inspired by that proposed in Kim and Park (2006) in the context of ratio imputation.

The properties of the proposed estimators (16)-(18) depend on the assumptions about the unit and the item nonresponse mechanisms. In particular, all three estimators assume a second phase of sampling with unknown response probabilities. If we ignore estimation of θ_k in Model (12), the results in Kim and Kim (2007) on consistency of the two-phase estimator that uses estimated response probabilities hold here as well when considering maximum likelihood estimates for the parameters α_0 and α_1.

As of the imputation process, estimator $\widehat{Y}_{j,pq}$ in (16) follows a fully weighted approach (see e.g. Särndal and Lundström, 2005, Chap. 12) and is a three-phase estimator with estimated unit and item response probabilities. Again, ignoring estimation of the latent variable θ_k and using conditional maximum likelihood estimates for the parameters $\beta_{\ell 0}$ and $\beta_{\ell 1}$ in Model (9), estimator $\widehat{Y}_{j,pq}$ will be consistent if the models for unit and item nonresponse probabilities are correctly specified. On the other hand, estimator $\widehat{Y}^{\text{imp}}_{j,p}$ in (17) follows a combined approach with reweighing for unit nonresponse and imputation for item nonresponse. In this case, the estimator will be consistent if the unit response probability model and the imputation model assumed to obtain predicted values y^*_{kj} both hold. Finally, estimator $\widehat{Y}^{\text{imp}}_{j,pq}$ takes a sort of model-assisted perspective in including the item nonresponse process. Therefore, it is expected to be double protected against imputation model misspecification. In particular, it is enough that either the imputation model assumed to obtain predicted values y^*_{kj} or the item response probability model for q_{kj} is correctly specified in order to achieve consistency. This is particularly relevant in our context in which no particular emphasis is put on the choice of the imputation model.

4 Simulation results

We evaluate the performance of the estimators presented in Section 3.3 by means of Monte Carlo simulations under two different settings. The first one uses a real data set as the population and considers variables of interest that are all binary, while the second one uses simulated population data with variables of interest that are all continuous. Results from the first setting are presented in Section 4.1, while those from the second setting are presented in Section 4.2.

In both settings the following estimators have been considered.

- $HT = \sum_{k \in s} y_{kj}/\pi_k$: the Horvitz-Thompson estimator in the case of full response has been computed as a benchmark for bias and variance in absence of nonresponse. It reduces to

N times the sample mean under srs without replacement (the design employed in both simulation settings).

- $\widehat{Y}_{j,\text{naive}}$: the naive estimator in (14) has been computed as a benchmark for the other estimators, in which no explicit action is taken to adjust for unit and item nonresponse. Note that for srs without replacement, it reduces to $\widehat{Y}_{j,\text{naive}} = N \sum_{k \in r_y} y_{kj}/n_{r_y}$, where n_{r_y} is the size of the set r_y, and it is the same as the Horvitz-Thompson estimator adjusted for unit nonresponse that assumes uniform response probabilities estimated by n_{r_y}/n.

- $\widehat{Y}^{\text{imp}}_{j,\text{naive}}$: the naive estimator that uses imputed values to treat item nonresponse considered in (15).

- $\widehat{Y}_{j,pq}$: the three-phase estimator proposed in Section 3.3, equation (16).

- $\widehat{Y}^{\text{true}}_{j,pq}$: the three-phase estimator that uses the true values for the response probabilities p_k and q_{kj} has been also computed for comparison with $\widehat{Y}_{j,pq}$ to understand the effect of estimating the response probabilities.

- $\widehat{Y}^{\text{imp}}_{j,p}$: the two-phase estimator that adjusts for item nonresponse via imputation proposed in (17).

- $\widehat{Y}^{\text{imp}}_{j,pq}$: the model assisted compromise proposed in (18).

- $\widehat{Y}^{\text{imp}}_{j,p\text{LC}}$: the estimator proposed by Laaksonen and Chambers (2006) and given by in our context by

$$\widehat{Y}^{\text{imp}}_{j,p\text{LC}} = \sum_{k \in r_y} \frac{y_{kj}}{\pi_k \widehat{p}_k^{LC}} + \sum_{k \in r \setminus r_y} \frac{y^*_{kj}}{\pi_k \widehat{p}_k^{LC}}, \tag{19}$$

  where $\widehat{p}_k^{LC}$ is an estimator of the response probability defined as $p_k^{LC} = P(R_k = 1 | y^*_{kj})$ that accounts for non ignorable nonresponse using imputed values. Note that it takes the same form of estimator $\widehat{Y}^{\text{imp}}_{j,p}$ in (17), the only difference lays in the way in which the response probabilities p_k are estimated.

The simulations are carried out in R version 2.15, using the R package `ltm` (Rizopoulos, 2006) to fit the latent models. The following performance measures have been computed for each estimator, generically denoted below by $\widehat{Y}$ where suffix j is dropped for ease of notation:

- the Monte Carlo Absolute Bias

$$\text{AB} = E_{sim}(\widehat{Y}) - Y,$$

  where

$$E_{sim}(\widehat{Y}) = \sum_{i=1}^{M} \widehat{Y}_i / M,$$

  $\widehat{Y}_i$ is the value of the estimator $\widehat{Y}$ at the i-th simulation run and M is total number of simulation runs and

- the Relative Bias

$$\text{RB} = \frac{\text{AB}}{Y};$$

- the Monte Carlo Standard Deviation

$$\sqrt{\text{VAR}} = \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} \left(\widehat{Y}_i - E_{sim}(\widehat{Y}) \right)^2};$$

- the Monte Carlo Mean Squared Error

$$\text{MSE} = \text{AB}^2 + \text{VAR};$$

- the Bias-to-Standard error Ratio

$$\text{BSR} = \frac{\text{AB}}{\sqrt{\text{VAR}}}.$$

4.1 Simulation setting 1

We consider the Abortion data set formed by four binary variables extracted from the 1986 British Social Attitudes Survey and concerning the attitude to abortion. $N = 379$ individuals answered to the following questions after being asked if the law should allow abortion under the circumstances presented under each item:

1. The woman decides on her own that she does not wish to keep the baby.

2. The couple agree that they do not wish to have a child.

3. The woman is not married and does not wish to marry the man.

4. The couple cannot afford any more children.

The variable of interest y_j is selected to be the second one ($j = 2$) with a total $Y_j = 225$ in the population, that is taken to be our parameter of interest.

The data is analyzed by Bartholomew et al. (2002) as an example in which a latent variable can be found that measures the attitude to abortion. At the population level we have computed on the $\{y_{k\ell}\}_{k=1,\dots,N;\ell=1,\dots,4}$ data such latent variable (denoted here by θ_k^a) using the 2PL model in (9). The correlation between the values $y_{k\ell}$ and θ_k^a is approximatively equal to 0.85, for $\ell = 1, \dots, 4$. Afterwards, we have set $\theta_k = \widehat{\theta}_k^a$, for all $k = 1, \dots, N$.

At the population level, the unit response probabilities are generated using the following response model

$$p_k = 1/(1 + \exp(-(0.7 + y_{k2} + \theta_k + 0.2\varepsilon_k))), \tag{20}$$

with $\varepsilon_k \sim U(0, 1)$, to simulate non ignorable nonresponse. The population mean of p_k is approximately 0.74.

To generate item response probabilities at the population level the following model is used

$$q_{k\ell} = 1/(1 + \exp(-(b_\ell\theta_k + a_\ell + y_{k\ell}))), \quad \text{for } \ell = 1, \dots, 4 \tag{21}$$

where $b_\ell = 3$, for $\ell = 1, \dots, 4$, while a_ℓ takes different values according to ℓ; in particular, $a_1 = 1$, $a_2 = 0$, $a_3 = -0.5$ and $a_4 = 1$. The nominal item nonresponse rate for the four items in the population is 35%, 42%, 47%, 31%, respectively.

We draw $M = 10'000$ simple random samples without replacement ($\pi_k = n/N$) from the population using two sample sizes: $n = 50$ and $n = 100$. In each sample s the units have been classified as respondents according to Poisson sampling, using the probabilities p_k computed as in equation (20) and resulting in the set r. Then, given r, the matrix $\{x_{k\ell}\}_{k\in r;\ell=1,\dots,4}$ is constructed where the values $x_{k\ell}$ are drawn using Poisson sampling with probabilities $q_{k\ell}$ defined in (21).

In each simulation run, Model (9) and the respondents set r are used to compute the variable $\widehat{\theta}_k$ for all $k \in s$. Model (12) is fitted to obtain $\widehat{p}_k$. To avoid the separation problem in fitting Model (12) we iteratively add to each sample s new observations (R_k, θ_k) until convergence, with $R_k = 0$ and $\theta_k \sim N(0, 1)$. These observations are afterwards deleted from the data set. The maximum number of added observations in simulations is 10 for $n = 50$ and 7 for $n = 100$. The average item nonresponse rate over simulations for the four items is found to be 26%, 33%, 38% and 23%.

Table 1: Simulation results for setting 1 − Abortion data set − $n = 50$

Estimator	AB	$\sqrt{\text{VAR}}$	MSE	% RB	% BSR
HT	0.6	24.6	603.3	0.3	2.5
$\widehat{Y}_{j,\text{naive}}$	127.1	19.3	16531.1	56.5	658.3
$\widehat{Y}_{j,\text{naive}}^{\text{imp}}$	34.4	33.6	2311.4	15.3	102.4
$\widehat{Y}_{j,pq}$	-4.5	53.9	2921.9	-2.0	-8.3
$\widehat{Y}_{j,pq,\text{true}}$	0.5	35.0	1226.5	0.2	1.5
$\widehat{Y}_{j,p}^{\text{imp}}$	-12.7	48.3	2491.4	-5.6	-26.3
$\widehat{Y}_{j,pq}^{\text{imp}}$	-0.7	59.0	3479.4	-0.3	-1.2
$\widehat{Y}_{j,p\text{LC}}^{\text{imp}}$	-27.8	33.1	1866.0	-12.3	-84.0

Table 2: Simulation results for setting 1 − Abortion data set − $n = 100$

Estimator	AB	$\sqrt{\text{VAR}}$	MSE	% RB	% BSR
HT	0.1	16.0	254.5	0.1	0.3
$\widehat{Y}_{j,\text{naive}}$	126.9	13.4	16284.9	56.4	945.0
$\widehat{Y}_{j,\text{naive}}^{\text{imp}}$	36.1	23.5	1855.0	16.0	153.8
$\widehat{Y}_{j,pq}$	-7.9	67.3	4588.1	-3.5	-11.8
$\widehat{Y}_{j,pq,\text{true}}$	0.2	23.7	563.4	0.1	0.8
$\widehat{Y}_{j,p}^{\text{imp}}$	-17.9	33.6	1449.5	-7.9	-53.2
$\widehat{Y}_{j,pq}^{\text{imp}}$	-5.5	44.4	2000.9	-2.5	-12.5
$\widehat{Y}_{j,p\text{LC}}^{\text{imp}}$	-27.4	22.5	1255.6	-12.2	-122.0

The imputed values y_{k2}^* (binary values) used by estimators $\widehat{Y}_{j,\text{naive}}^{\text{imp}}$, $\widehat{Y}_{j,q}^{imp}$, $\widehat{Y}_{j,pq}^{\text{imp}}$ and $\widehat{Y}_{j,p\text{LC}}^{\text{imp}}$ are computed according to Poisson sampling with predicted probabilities computed by Model (9), for all $k \in r_y$. Tables 1 and 2 give the results for $n = 50$ and $n = 100$, respectively.

As expected, HT and $\widehat{Y}_{j,pq,\text{true}}$ have almost zero bias, with the second one showing a relatively larger MSE that is due uniquely to the smaller sample size. The naive estimator shows a very large positive bias. This is due to the fact that units with a zero value of y_j are less likely to respond and the total is clearly overestimated. This bias is reduced by $\widehat{Y}_{j,\text{naive}}^{\text{imp}}$, by meaning that the imputation model employed is relatively good in describing item nonresponse. The proposed estimators all reduce significantly the bias compared to the two aforementioned estimators, with $\widehat{Y}_{j,pq}^{\text{imp}}$ providing the best performance, meaning that the unit and item response probability models are able to capture the attitude to respond to the questionnaire that is linked with the attitude towards abortion. In particular, $\widehat{Y}_{j,pq}$ and $\widehat{Y}^{\text{imp}}_{j,pq}$ show a very good coverage measured by BSR in the last column, meaning that the model for the item response probabilities is better specified than the imputation model. This fact also affects the not very good performance of $\widehat{Y}^{\text{imp}}_{j,p\text{LC}}$ that bases the model for the response probabilities p_k's on the imputation model. In general, it can be noted that estimators that use imputation tend to have a smaller variance (like a shrinkage effect) that influences coverage. Note that the performance of all estimators but HT and $\widehat{Y}_{j,pq,\text{true}}$, is mostly driven by the absolute bias, so that the performance is not relatively different when increasing the sample size, apart from a reduction in coverage given by the decrease in variance.

To study the performance of the latent model on this data set and the correlation between the variable of interest and the estimated latent variable we apply the procedure described before to construct the matrix $\{x_{k\ell}\}_{k=1,\ldots,N;\ell=1,\ldots,4}$ for all population units. On this dataset, we fit the 2PL model in (9) and compute the variable θ_k for all $k = 1, \ldots, N$. The Cronbach's alpha measure takes value 0.83 showing a good internal consistency of the items. The χ^2 p-values for pairwise associations between the four items are smaller than 0.01. These p-values correspond to the 2×2 contingency tables for all possible pairs of items. Inspection of non significant results can be used to reveal 'problematic' items. Latent variable models assume that high association between items can be explained by a set of latent variables. Thus, for pairs of items that do not reject independence we could say that they violate this assumption.

The correlation coefficient between the variable of interest and the estimated latent variable takes value 0.76, indicating that the latent auxiliary information has a powerful power of predicting y_{k2}, as advocated in the model of Cassel et al. (1983). The inspection of the two-way margins for the matrix $\{x_{k\ell}\}$ gives the residuals $(O-E)^2/E$ between 0.03 and 0.23. Similarly the three-way margins for the matrix $\{x_{k\ell}\}$ give residuals between 0 and 1.19. This indicates that we have no reason to reject here the one-factor latent Model (9). This decision is based on the fact that if we consider the residual in each cell as having a χ^2 distribution with one degree of freedom, then a value of the residual greater than 4 is indicative of poor fit at the 5% significance level (see Bartholomew et al., 2002, p. 186).

4.2 Simulation setting 2

We generate $\{y_{k1}, \ldots, y_{k6}, \theta_k\}$ for $k = 1, \ldots, N = 2'000$ using a multivariate normal distribution. The degree of correlation between y_ℓ and $y_{\ell'}$ is 0.8, with $\ell, \ell' = 1, \ldots, 6, \ell \neq \ell'$. We set the variable of interest to be y_1 and consider different degrees of correlation between the values y_{k1} and θ_k: namely, 0.3, 0.5, 0.8. The values of θ_k are afterwards standardized to have mean 0 and variance 1.

The response probabilities are obtained by first computing

$$p_k^\circ = 1/(1 + \exp(-(0.5 + y_{k1} + \theta_k))), \quad \text{for } k = 1, \ldots, N, \tag{22}$$

and then rescaling them to take values between 0.1 and 0.9 using the transformation

$$p_k = (p_k^\circ - \min_k p_k^\circ)/(\max_k p_k^\circ - \min_k p_k^\circ) \times 0.8 + 0.1, \tag{23}$$

with a population mean approximatively equal to 0.7.

The item response probabilities are generated by first computing

$$q_{k\ell}^\circ = 1/(1 + \exp(-(b_\ell \theta_k + a_\ell + y_{k\ell}))), \quad \text{for } k = 1, \ldots, N \text{ and } \ell = 1, \ldots, 6, \tag{24}$$

where $\{a_\ell\}_{\ell=1,\ldots,6} = \{1, 0, -0.5, 1, 0, -0.5\}$ and $\{b_\ell\}_{\ell=1,\ldots,6} = \{1, 1, 1, 1.5, 1.5, 1.5\}$, and then rescaling the values to be between 0.1 and 0.95 using the transformation

$$q_{k\ell} = (q_{k\ell}^\circ - \min_k q_{k\ell}^\circ)/(\max_k q_{k\ell}^\circ - \min_k q_{k\ell}^\circ) \times 0.85 + 0.1. \tag{25}$$

We draw $M = 10'000$ samples by simple random simple without replacement of dimension $n = 200$. For each sample s, a response set r is created by carrying out Poisson sampling with parameter p_k. Each element of the matrix $\{x_{k\ell}\}_{k \in r, \ell=1,\ldots,6}$ is generated using Poisson sampling with parameter $q_{k\ell}$. Item nonresponse rates over simulations take approximately value 12%, 20%, 25%, 14%, 22%, 26%, for $\ell = 1, \ldots, 6$, respectively.

For each simulation run, the 2PL model in (9) is used to compute the variable $\widehat{\theta}_k$ for all $k \in s$. Model (12) is then fitted to obtain $\widehat{p}_k$. In this simulation setting there has been no need to add randomly generated values to avoid the separation problem in fitting Model (12).

The imputed values y_{k1}^* (continuous in this case) to be used by estimators $\widehat{Y}_{j,\text{naive}}^{\text{imp}}$, $\widehat{Y}_{j,q}^{imp}$, $\widehat{Y}_{j,pq}^{\text{imp}}$ and $\widehat{Y}_{j,p\text{LC}}^{\text{imp}}$ have been computed using the regression imputation method. In particular, the following linear regression model

$$y_{k1} = \alpha_0 + \alpha_1 \widehat{\theta}_k + \varepsilon_k, \quad \varepsilon_k \sim N(0, \sigma^2) \tag{26}$$

Table 3: Simulation results for setting 2 – Simulated continuous data – correlation coefficient 0.3

Estimator	AB	$\sqrt{\text{VAR}}$	MSE	% RB	% BSR
HT	0.4	131.8	17360.33	<0.0	0.3
$\widehat{Y}_{j,\text{naive}}$	392.2	149.4	176159.8	9.9	262.5
$\widehat{Y}^{\text{imp}}_{j,\text{naive}}$	377.7	151.7	165649.9	9.5	248.9
$\widehat{Y}_{j,pq}$	-27.3	472.7	224187.0	0.7	-5.8
$\widehat{Y}_{j,pq,\text{true}}$	0.2	222.1	49323.2	<0.0	0.1
$\widehat{Y}^{\text{imp}}_{j,p}$	89.3	398.8	167027.6	2.2	22.4
$\widehat{Y}^{\text{imp}}_{j,pq}$	82.0	393.8	161781.3	2.1	20.8
$\widehat{Y}^{\text{imp}}_{j,p\text{LC}}$	296.0	219.7	135877.58	7.4	134.7

Table 4: Simulation results for setting 2 – Simulated continuous data – correlation coefficient 0.5

Estimator	AB	$\sqrt{\text{VAR}}$	MSE	% RB	% BSR
HT	0.8	136.1	18514.7	<0.0	0.6
$\widehat{Y}_{j,\text{naive}}$	483.5	152.2	256972.7	12.1	317.6
$\widehat{Y}^{\text{imp}}_{j,\text{naive}}$	439.8	160.0	219048.5	11.0	274.9
$\widehat{Y}_{j,pq}$	-116.3	520.4	284354.1	-2.9	-22.4
$\widehat{Y}_{j,pq,\text{true}}$	6.5	223.7	50080.7	0.2	2.9
$\widehat{Y}^{\text{imp}}_{j,p}$	14.3	475.1	225957.74	0.4	3.0
$\widehat{Y}^{\text{imp}}_{j,pq}$	0.6	463.0	214326.4	<0.0	0.1
$\widehat{Y}^{\text{imp}}_{j,p\text{LC}}$	229.2	328.2	160281.0	5.7	69.8

is fitted on the set r_y. The imputed values y^*_{k1} are given by the predicted values of Model (26) for given $\widehat{\theta}_k, k \in r\backslash r_y$. Tables 3, 4, and 5 give the performance of the estimators for the three values taken by the nominal correlation coefficient between y_{k1} and θ_k – 0.3, 0.5, and 0.8, respectively.

The proposed estimators are always able to reduce the bias over the naive estimators, even when the correlation between the variable of interest and the latent variable gets smaller. The relative bias takes acceptable values in most cases, while the ratio between the bias and the standard error gets in some instances larger than 20%, providing some concern on coverage. The absolute bias deserves a closer look. The naive estimators in all cases largely overestimate the total. This is expected, because the values p_k, q_{kj}, θ_k and y_{k1} all go in the same direction. Therefore we have in our respondents sample more likely relative larger values for y_1 by this providing overestimation for the naive estimators. The imputation model uses the relationship between y_{k1} and θ_k and, therefore,

Table 5: Simulation results for setting 2 – Simulated continuous data – correlation coefficient 0.8

Estimator	AB	$\sqrt{\text{VAR}}$	MSE	% RB	% BSR
HT	-2.0	136.3	18578.1	<0.0	-1.5
$\widehat{Y}_{j,\text{naive}}$	535.5	147.8	308568.0	13.2	362.2
$\widehat{Y}^{\text{imp}}_{j,\text{naive}}$	447.6	166.8	228191.4	11.1	268.3
$\widehat{Y}_{j,pq}$	-220.4	531.8	331350.2	-5.5	-41.4
$\widehat{Y}_{j,pq,\text{true}}$	8.3	220.6	48738.8	0.2	3.8
$\widehat{Y}^{\text{imp}}_{j,p}$	-111.5	532.9	296441.1	-2.8	-20.9
$\widehat{Y}^{\text{imp}}_{j,pq}$	-132.4	510.2	277853.3	-3.3	-25.9
$\widehat{Y}^{\text{imp}}_{j,p\text{LC}}$	23.6	480.1	231081.5	0.6	4.9

helps $\widehat{Y}^{\text{imp}}_{j,\text{naive}}$ to decrease the AB, and this gets relatively larger as the level of the correlation between y_{k1} and θ_k increases. On the other hand, $\widehat{Y}_{j,pq}$ underestimates the total because it is based only on the observed units of r_y that do have relatively large values for y_1, but also relatively large values for p and q_1 and, therefore, end up having a small weight. This underestimation gets larger and larger as the correlation between θ_k and y_{k1} (and therefore, between θ_k and the response probabilities) increases. The use of the imputation model in estimators $\widehat{Y}^{\text{imp}}_{j,p}$ and $\widehat{Y}^{\text{imp}}_{j,pq}$ helps in adjusting by excess this estimate, with the latter one being a good compromise. In fact, $\widehat{Y}^{\text{imp}}_{j,pq}$ provides more often a better performance among the three. The $\widehat{Y}^{\text{imp}}_{j,p\text{LC}}$ also shows a good behavior in reducing the bias, although its performance is very much linked to the degree of the correlation between the variable of interest and the latent variable (therefore it behaves worse as it decreases).

The matrix of population values $\{x_{k\ell}\}_{k=1,\ldots,2000,\ell=1,\ldots,6}$ has been constructed in the same way as in Section 4.1 to validate the assumptions behind the 2PL model. The Cronbach's alpha takes approximately value 0.5 for the correlation coefficient equal to 0.3, 0.6 for 0.5, and 0.7 for 0.8; the pairwise association between the six items reveales p-values smaller than 0.01. The inspection of the two-way and three-way margins of the matrix $\{x_{k\ell}\}$ gives residuals $(O-E)^2/E$ that all take values smaller than 4. Therefore, the one factor latent model can be accepted and items seem to be all measuring the same latent trait.

To test what happens to estimators $\widehat{Y}^{\text{imp}}_{j,\text{naive}}$, $\widehat{Y}^{imp}_{j,q}$, $\widehat{Y}^{\text{imp}}_{j,pq}$ and $\widehat{Y}^{\text{imp}}_{j,p\text{LC}}$ if we use a different imputation model, we run another simulation. To this end, we use the simulated population data where the correlation coefficient between y_{k1} and θ_k is 0.8. The imputed values y^*_{k1} are now computed

Table 6: Simulation results for setting 2 – Simulated continuous data – correlation coefficient 0.8

Estimator	AB	$\sqrt{\text{VAR}}$	MSE	% RB	% BSR
HT	0.1	137.9	19016.2	<0.0	< 0.0
$\widehat{Y}_{j,\text{naive}}$	558.8	149.3	334524.8	14.0	374.3
$\widehat{Y}^{\text{imp}}_{j,\text{naive}}$ $\star$	558.8	149.3	334524.8	14.0	374.3
$\widehat{Y}_{j,pq}$	-237.8	534.2	341952.7	6.0	-44.5
$\widehat{Y}_{j,pq,\text{true}}$	5.4	220.0	48429.1	0.1	2.4
$\widehat{Y}^{\text{imp}}_{j,p}$ $\star$	80.6	405.4	170841.8	2.0	19.9
$\widehat{Y}^{\text{imp}}_{j,pq}$ $\star$	27.2	404.5	164319.4	0.7	6.7
$\widehat{Y}^{\text{imp}}_{j,p\text{LC}}$ $\star$	558.8	149.3	334524.8	14.0	374.3

* Mean imputation model used to obtain y^* values.

using the mean of all respondents of the set r_y, i.e. a neutral imputation model. Table 6 gives the results of this simulation. It is expected that the naive estimators and $\widehat{Y}^{\text{imp}}_{j,p\text{LC}}$ take the same value in this case. Note also that such values are similar to that of the naive estimator in Table 5; this is true also for $\widehat{Y}_{j,pq}$. In fact, differences in these cases stem only from simulation error. Estimators $\widehat{Y}^{\text{imp}}_{j,p}$ and $\widehat{Y}^{\text{imp}}_{j,pq}$ improve their performance over that of Table 5, because they use an imputation model that predicts y^* values that are relatively larger (recall that units in the r_y set have relatively larger values of y) and, therefore, help reducing the underestimation of estimator $\widehat{Y}_{j,pq}$ in this case. It is therefore, very important, to take into consideration the correlation structure existing among p, q, θ and y to try to take the best out of it.

Finally, we explore in another simulation the robustness of the proposed method to the case in which θ_k does not follow a normal distribution. Population data is generated like in the last simulation. Then, the values θ_k are transformed to follow a skew distribution by adding a χ^2 random variable with one degree of freedom. This provides a new variable θ_k with skewness equal to 1.54. The correlation coefficient between the variable of interest and the obtained latent variable is about 0.5.

At the population level, the unit response patterns are generated by the following equation

$$p_k = 1/(1 + \exp(-(0.7 + y_{k1} - 0.3\theta_k))), \tag{27}$$

to achieve a population mean approximatively equal to 0.82. The item response probabilities are generated as in (24) and then rescaled to the range 0.1 and 0.75. The items' nonresponse rate

Table 7: Simulation results for setting 2 – Simulated continuous data – skew distribution for θ_k, correlation coefficient 0.5

Estimator	AB	$\sqrt{\text{VAR}}$	MSE	RB (in %)	BSR (in %)
HT	-13.3	181.8	33227.0	-0.3	-7.3
$\widehat{Y}_{j,\text{naive}}$	266.6	248.9	133017.0	6.7	107.1
$\widehat{Y}^{\text{imp}}_{j,\text{naive}}$	259.2	248.2	128796.9	6.5	104.4
$\widehat{Y}_{j,pq}$	-104.3	416.3	184162.8	-2.6	-25.1
$\widehat{Y}_{j,pq,\text{true}}$	12.0	315.0	99370.5	0.3	3.8
$\widehat{Y}^{\text{imp}}_{j,p}$	142.6	256.2	85986.3	3.6	55.7
$\widehat{Y}^{\text{imp}}_{j,pq}$	144.0	255.1	85829.5	3.6	56.4
$\widehat{Y}^{\text{imp}}_{j,p\text{LC}}$	242.9	249.1	121078.4	6.1	97.5

means in simulations take values between 25% and 29%. Imputed values y^*_{k1} for estimators $\widehat{Y}^{\text{imp}}_{j,\text{naive}}$, $\widehat{Y}^{imp}_{j,q}$, $\widehat{Y}^{\text{imp}}_{j,pq}$ and $\widehat{Y}^{\text{imp}}_{j,p\text{LC}}$ are computed using the model of equation (26).

Table 7 gives the performance estimators for this simulation setting. As in the previous simulations the proposed estimators perform better than the naive estimators in terms of bias. It seems that the skewness of the distribution of θ_k does not have a negative influence on the behavior of the proposed estimators. Estimator $\widehat{Y}_{j,pq}$ shows the best performance with the lowest bias and a good coverage. On the other hand, there is some concern on coverage for those estimators based on imputation, for which the standard errors are relatively lower, but bias, although smaller than for the naive estimators, is still quite high.

5 Conclusions

We have proposed a reweighting system to compensate for unit non-ignorable nonresponse based on a latent auxiliary variable. This variable is computed for each unit in the sample using a latent model assuming the existence of item nonresponse and that the same latent structure is hidden behind item and unit nonresponse. Unit response probabilities are then estimated by a logistic model that uses as covariate the latent trait extracted by the response patterns using Item response theory models. The proposed reweighting system is then used in different estimators, assuming imputation, reweighting or a combination of the two to handle item nonresponse. The main goal is to reduce the nonresponse bias in the estimation of the population total. The proposed estimators perform well in our simulation studies comparing to the naive estimators, and the gain in efficiency is substantial in certain cases. Reductions in bias are displayed also when the correlation between the latent trait and the variable of interest is modest.

It is also interesting to study the trade-off between the reduction of nonresponse bias and the increase in variance. By construction the estimated latent variable $\widehat{\theta}_k$ is related to the response indicators x_{kj} for the variable of interest y_j; since nonresponse is assumed to be non-ignorable, y_{kj} and x_{kj} are related as well. If the following condition holds,

$$\rho^2_{y_j,x_j} + \rho^2_{\widehat{\theta},x_j} > 1,$$

where the correlation coefficients $\rho_{y_j,x_j}, \rho_{\widehat{\theta},x_j} > 0$, then y_j and $\widehat{\theta}$ are positively correlated (see Langford et al., 2001). Following Little and Vartivarian (2005), 'a covariate for a weighting adjustment must have two characteristics to reduce nonresponse bias - it needs to be related to the probability of response, and it needs to be related to the survey outcome. If the latter is true, then weighting can reduce, not increase, sampling variance.' A positive correlation between y_j and $\widehat{\theta}$ can thus provide also reduction of the variance of the proposed estimators.

We have considered the case in which no auxiliary information is available at the sample or population level to reduce nonresponse bias. Observed covariates (if available) and the latent variable $\widehat{\theta}_k$ can be, however, used together in the estimation of the response probabilities. Moreover, latent trait models can be fitted with covariates. The introduction of covariates in these models should be carried out with the prudence of variance increasing.

References

Bartholomew, D. J. and Knott, M. (1999). *Latent Variable Models and Factor Analysis.* Arnold, London.

Bartholomew, D. J., Steele, F., Moustaki, I., and Galbraith, J. I. (2002). *The Analysis and Interpretation of Multivariate Data for Social Scientists.* Chapman and Hall/CRC.

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

Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4(3):251–260.

Biemer, P. P. and Link, M. W. (2007). *Evaluating and modeling early cooperator effects in RDD surveys*. New York: Wiley.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1983). Some uses of statistical models in connection with the nonresponse problem. In Madow, W. G. and Olkin, I., editors, *Incomplete Data in Sample Surveys*, volume 3, pages 143–160. New York: Academic Press.

Chambers, R. L. and Skinner, C. (2003). *Analysis of Survey Data*. Wiley, New York.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public use samples using bayesian logistic regression. *Journal of the American Statistical Association*, 86:68–78.

Copas, A. J. and Farewell, V. T. (1998). Dealing with non-ignorable non-response by using an 'enthusiasm-to-respond' variable. *Journal Of The Royal Statistical Society, Series A*, 161:385–396.

De Leeuw, E. D. (1992). *Data Quality in Mail, Telephone and Face to Face Surveys*. T.T. Publikaties, Amsterdam.

De Menezes, L. M. and Bartholomew, D. J. (1996). New developments in latent structure analysis applied to social attitudes. *Journal of Royal Statistical Society A*, 159:213–224.

Drew, J. H. and Fuller, W. A. (1980). Modeling nonresponse in surveys with callbacks. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*.

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.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70 (5):646–675.

Groves, R. M., Couper, M., Presser, S., Singer, E., Tourangeau, R., Acosta, G. P., and Nelson, L. (2006). Experiments in producing nonresponse bias. *Public Opinion Quarterly*, 70(5):720–736.

Hansen, M. and Hurwitz, W. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41:517–529.

Juster, T. and Smith, J. P. (1997). Improving the quality of economic data: Lessons from the HRS and AHEAD. *Journal of the American Statistical Association*, 92:1268–1278.

Kersten, H. M. P. and Bethlehem, J. G. (1984). Exploring and reducing the nonresponse bias by asking the basic question. *The Statistical Journal of the United Nations Commission for Europe*, 2:369–380.

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, 35:501–514.

Kim, J. K. and Park, H. (2006). Imputation using response probability. *Canadian Journal of Statistics*, 34:171–182.

Laaksonen, S. and Chambers, R. (2006). Survey estimation under informative nonresponse with follow-up. *Journal of Official Statistics*, 22(1):81–95.

Langford, E., Schwertman, N., and Owens, M. (2001). Is the property of being positively correlated transitive? *The American Statistician*, 55 (4):322–325.

Little, R. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31:161–168.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Matei, A. (2010). Dealing with nonresponse in survey sampling: an item response modeling approach. In Lechevallier, Y. and Saporta, G., editors, *Proceedings of COMPSTAT'2010*, pages 1353–1358. Springer-Verlag.

Moore, J. C., Stinson, L. L., and Welniak, Jr., E. J. (1999). Income reporting in surveys: Cognitive issues and measurement error. In Sirken, M. G., Herrmann, D. J., Schechter, S., Schwarz, N., Tanur, J. M., and Tourangeau, R., editors, *In Cognition and Survey Research*. New York: Wiley.

Moustaki, I. and Knott, M. (2000). Weighting for item non-response in attitude scales using latent variable models with covariates. *Journal of Royal Statistical Society, Series A*, 163:445–459.

Oh, H. L. and Scheuren, F. J. (1983). Weighting adjustments for unit non-response. In Madow, W. G., Olkin, I., and Rubin, D. B., editors, *Incomplete Data in Sample Surveys*, volume 2, pages 143–184. New York: Academic Press.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61:317–337.

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.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. The Danish Institute of Educational Research, Copenhagen.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17 (5):1–25.

Rizopoulos, D. and Moustaki, I. (2008). Generalized latent variable models with non-linear effects. *British Journal of Mathematical and Statistical Psychology*, 61:415–438.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

Skrondal, A. and Rabe-Hesketh, S. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, 34:712–745.

Särndal, C. E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons, New York.

Zhang, L. C. (2002). A method of weighting adjustment for survey data subject to nonignorable nonresponse. DACSEIS research paper no. 2, http://w210.ub.uni-tuebingen.de/dbt/volltexte/2002/451.