Adjusting for nonignorable nonresponse using a latent variable modeling approach

Matei, Alina
Institute of Statistics, University of Neuchâtel
and Institute of Pedagogical Research and Documentation, Neuchâtel
Rue Pierre à Mazel 7
2000 Neuchâtel, Switzerland
E-mail: alina.matei@unine.ch

Ranalli, M. Giovanna
Dipartimento di Economia, Finanza e Statistica, Università degli Studi di Perugia
Via Pascoli
06123 Perugia, Italy
E-mail: giovanna@stat.unipg.it

Introduction

In survey sampling, a random sample is drawn from a finite population in order to perform inference on descriptive characteristics of some variables of interest. Usually, nonresponse occurs and, as a consequence, the variables of interest are not observed for the entire selected sample by causing missing data. We distinguish between two types of missing data: *unit nonresponse*, when a selected sample unit is not observed at all – reasons for this may be that the unit is not found at home, or he/she is not in the condition of providing the information required because ill or not informed, or simply because he/she refuses to collaborate – and *item nonresponse*, when an interviewed unit does not respond to all of the questions in the questionnaire. Addressing the issue of nonresponse is very important, since nonresponse is present in almost all surveys and, above all, can highly bias estimates if the responding units are systematically different from the non responding ones.

Several techniques have been proposed in the literature to deal with nonresponse at the estimation stage. Typically, unit and item nonresponse are treated separately: unit nonresponse adjustments use methods based on response modeling or on calibration (see e.g. Cassel et al., 1983; Kim and Kim, 2007; Särndal and Lundström, 2005), while item nonresponse is usually addressed via imputation (single or multiple, see e.g. Rubin, 1987). Usually, unit nonresponse is treated in a two-phase framework, in which the selected sample is the first phase sample, while the set of respondents is considered as a second phase sample with unknown probabilities of inclusion. The latter are unknown individual characteristics defined for all units in the population and measure the probability that a unit responds given that it was included in the sample. When auxiliary information is available for all units in the original sample, these probabilities can be estimated. A common approach is to use a logistic model for the response indicator (see e.g. Kim and Kim, 2007).

Note that the response probability is a measure of the propensity of a unit to participate in the survey and that, therefore, it can also be considered as a latent variable. The use of latent variable models with covariates was proposed by Moustaki and Knott (2000) for weighting in the presence of item non-response. 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. Non-ignorable non-response is typical of surveys with sensitive questions (concerning drug abuse, sexual attitudes, politics, income, etc). The proposed method develops weights for the respondents by first linking unit non-response to item non-response via a continuous latent variable. This latent variable will be then used as a covariate for response probability estimation. Following Moustaki and Knott (2000), 'weight-

ing through latent variable modelling 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 will perform better in reducing bias than the naive estimator computed without this adjustment. The paper is organized as follows. After a short introduction to latent variable models, the proposed methodology is illustrated. The properties of the proposed estimators are sketched and some results from a simulation study are presented, together with some concluding remarks.

Latent trait models

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 categorical data (see Bartholomew et al., 2002). We focus here on binary data and continuous latent variables. Let $\mathbf{x}_k = (x_{k1}, \ldots, x_{k\ell}, \ldots, x_{km})'$ be the vector of binary indicator variables observed for unit $k = 1, \ldots, n$. For example, in the psychometric literature, $x_{k\ell} = 1$ if student k provides a correct answer to test item ℓ and $x_{k\ell} = 0$ otherwise. Denote by $q_{k\ell} = Pr(x_{k\ell} = 1|\boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k = (\theta_{k,1}, \ldots, \theta_{k,j}, \ldots, \theta_{k,J})'$ is the value taken on unit k for the vector of $J < m$ latent variables. The latent trait model is defined as

$$\ln\left(\frac{q_{k\ell}}{1-q_{k\ell}}\right) = \beta_{\ell 0} + \sum_{j=1}^{J} \beta_{\ell j}\theta_{k,j},$$

where $\beta_{\ell 0}, \ldots, \beta_{\ell J}$ are the model parameters, $\ell = 1, \ldots, m$, and $k = 1, \ldots, n$.

An important special case is obtained by taking $J = 1$, i.e. a unidimensional latent trait model

$$\ln\left(\frac{q_{k\ell}}{1-q_{k\ell}}\right) = \beta_{\ell 0} + \beta_{\ell 1}\theta_{k,1}. \tag{1}$$

For simplicity denote $\theta_{k,1}$ by θ_k. Usually it is assumed that $\theta_k \sim N(0,1)$. Model (1) is also referred to as a two parameter logistic Rasch model, and is essentially a logistic regression except that the θ_k's are not observed. In the psychometric example, the latent variable θ_k can be interpreted as the ability of student k to solve the test. The probability of success $q_{k\ell}$ is assumed to be a monotonic increasing function of the ability θ (*monotonicity* assumption). Under the assumption of *local independence*, i.e. that responses to the m items are independent for each k given θ_k, the parameters $\beta_{\ell 0}$ and $\beta_{\ell 1}$ are estimated for each item ℓ: $\beta_{\ell 0}$ reflects the extremeness of item ℓ and is also a measure of easiness (larger values are connected with larger values of a positive response at all points in the latent space); $\beta_{\ell 1}$ is known as the 'discrimination' parameter, and is a measure of how much information the item ℓ provides about the latent variable θ_k. Different goodness-of-fit measures of Model (1) are available in the literature to check whether the assumptions of unidimensionality, monotonicity and local independence hold (see e.g. Bartholomew et al., 2002).

Latent trait modelling for response propensities

Let U be a finite population of size N, indexed from 1 to N. Let s denote the set of sample labels, so that $s \subset U$. The sample s is drawn using a probabilistic sampling design $p(s)$, which induces first order inclusion probabilities π_k. Each unit k has also associated a response probability p_k. Denote by $r \subset s$ the set of respondents, and by $\bar{r} = s \backslash r$ the set of nonrespondents. The response mechanism is given by the distribution $q(r|s)$ such that for every s we have $q(r|s) \geq 0$, for all $r \in \mathcal{R}_s$ and $\sum_{s\in\mathcal{R}_s} q(r|s) = 1$, where $\mathcal{R}_s = \{r|r \subset s\}$. Under unit nonresponse, we define the response indicator $R_k = 1$ if $k \in r$ and 0 otherwise. Thus $r = \{k \in s|R_k = 1\}$ and $p_k = P(k \in r|k \in s) = P(R_k = 1|k \in s)$. We assume that units respond independently of each other and of s, and so $q(r|s) = \prod_{k\in r} p_k \prod_{k\in\bar{r}}(1-p_k)$.

Let y_j be a particular variable of interest (y_{kj} is its value for unit k), and let the response mechanism depend on it (non-ignorable nonresponse). For such a case, the response probability p_k should be modelled using a logistic regression as follows

$$p_k = P(R_k = 1|y_{kj}) = \frac{1}{1+\exp(-(a_0 + a_1 y_{kj}))}, \text{ for } k \in s, \text{ or as follows} \tag{2}$$

$$p_k = P(R_k = 1|y_{kj}, \mathbf{z}_k) = \frac{1}{1+\exp(-(a_0 + a_1 y_{kj} + \mathbf{z}_k'\boldsymbol{\alpha}))}, \text{ for } k \in s, \tag{3}$$

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. Since y_{kj} is only observed on the respondents, Models (2) and (3) cannot be estimated. Cassel et al. (1983) suggest using 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' in the following model

$$p_k = \frac{1}{1+\exp(-\mathbf{z}_k'\boldsymbol{\alpha})}. \tag{4}$$

Then, maximum likelihood can be used to fit Model (4) using the data $(R_k, \mathbf{z}_k)$ for $k \in s$. This leads to an estimate $\widehat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ and to the estimated response probabilities $\hat{p}_k = 1/(1+\exp(-\mathbf{z}_k'\hat{\boldsymbol{\alpha}}))$. A two-phase type estimator that uses weights $1/(\pi_k \hat{p}_k)$ 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).

Now, let M be the total number of variables (or items) in the questionnaire. Suppose that item nonresponse is also present for variable $\ell = 1, \ldots, m \leq M$ in the survey. For each item ℓ another response indicator is introduced: $x_{k\ell}$ is a binary variable 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 now the vector of response indicators of unit k to the m items. Let also $y_{k\ell}$ be the response value of unit k to item ℓ, and let $\mathbf{y}_k = (y_{k1}, \ldots y_{k\ell}, \ldots, y_{km})'$ be the vector of values taken by m survey variables $\mathbf{y}$ on unit k. Item nonresponse can, for example, be due to the lack of a strong opinion or interest on the topics of the survey, or to a refusal to answer a sensitive question. Suppose the $x_{k\ell}$'s are related to an assumed underlying latent continuous variable; they are the indicators of a latent trait denoted by θ_k. If the m items are selected so that they are particularly relevant for the survey and include the survey variable upon which the response mechanism is believed to depend, then we may interpret θ_k as the *will* or the *propensity to respond to the survey* of unit k. We can obtain estimates of θ_k using Model (1) on $\mathbf{x}_k$ for $k \in r$. If we knew θ_k for all units in s, we could use it as another covariate in Model (4). In the case in which no covariates $\mathbf{z}$ were available, Model (4) could be rewritten simply as

$$p_k = P(R_k = 1|\theta_k) = \frac{1}{1+\exp(-(\alpha_0 + \alpha_1 \theta_k))}. \tag{5}$$

To compute a value θ_k for a nonrespondent $k \in \bar{r}$, we consider that unit nonresponse generalizes item nonresponse. In fact, 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.' Thus, a nonrespondent can be considered as a unit that does not answer any item ℓ and thus $x_{k\ell} = 0$, for all $\ell = 1, \ldots, m$, and $k \in \bar{r}$. Therefore, if we assume that the causes of item nonresponse and unit nonresponse are related, Model (1) allows the computation of θ_k also for all $k \in \bar{r}$.

Model (5) cannot be directly fitted, since the problem of separation is present. This problem is observed in the fitting process of a logistic regression model if 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 $k \in \bar{r}$) have the same value of the covariate θ_k. To overcome this problem two practical solutions can be used here: (i) add a jitter to the value of θ_k (some noise to the θ_k value

of the cases with $R_k = 0$), or (ii) 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.

Adjustment for unit and item nonresponse: the proposed 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 mean $\bar{Y}_j = \sum_{k=1}^{N} y_{kj}/N$ of y_j, then a Hájek type naive estimator that does not correct neither for unit nor for item nonresponse is given by

$$\widehat{\bar{Y}}_{j,\text{naive}} = \sum_{k\in r_y} \frac{y_{kj}}{\pi_k} / \sum_{k\in r_y} \frac{1}{\pi_k}. \tag{6}$$

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

$$\widehat{\bar{Y}}^{imp}_{j,\text{naive}} = \Big(\sum_{k\in r_y} \frac{y_{kj}}{\pi_k} + \sum_{k\in r\setminus r_y} \frac{y^*_{kj}}{\pi_k} \Big) / \Big(\sum_{k\in r} \frac{1}{\pi_k} \Big). \tag{7}$$

Reweighting item responders is an 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 $q_{k\ell} > 0$. Therefore, a possible adjustment weight for item and unit nonresponse associated to unit $k \in r_y$ is $1/(\widehat{p}_k \widehat{q}_{kj})$. By considering different combinations of the aforementioned methods we propose the following set of estimators:

- a Horvitz-Thompson estimator adjusted for item and unit nonresponse via reweighting

$$\widehat{\bar{Y}}_{j,q\_\text{HT}} = \sum_{k\in r_y} \frac{y_{kj}}{\pi_k \widehat{p}_k \widehat{q}_{kj}} / N; \tag{8}$$

- a Hájek type estimator adjusted for item and unit nonresponse via reweighting

$$\widehat{\bar{Y}}_{j,q\_\text{Hajek}} = \sum_{k\in r_y} \frac{y_{kj}}{\pi_k \widehat{p}_k \widehat{q}_{kj}} / \sum_{k\in r_y} \frac{1}{\pi_k \widehat{p}_k \widehat{q}_{kj}}; \tag{9}$$

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

$$\widehat{\bar{Y}}^{imp}_{j,q\_\text{HT}} = \Big(\sum_{k\in r_y} \frac{y_{kj}}{\pi_k \widehat{p}_k} + \sum_{k\in r\setminus r_y} \frac{y^*_{kj}}{\pi_k \widehat{p}_k} \Big) / N; \tag{10}$$

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

$$\widehat{\bar{Y}}^{imp}_{j,q\_\text{Hajek}} = \Big(\sum_{k\in r_y} \frac{y_{kj}}{\pi_k \widehat{p}_k} + \sum_{k\in r\setminus r_y} \frac{y^*_{kj}}{\pi_k \widehat{p}_k} \Big) / \sum_{k\in r} \frac{1}{\pi_k \widehat{p}_k}. \tag{11}$$

The asymptotic properties of estimators (8)-(9) and of estimators (10)-(11) depend on the assumptions about the response and imputation mechanism. In particular, all estimators assume a second

phase of sampling with unknown response probabilities. If we ignore estimation of θ_k in Model (5), the results in Kim and Kim (2007) hold here as well. As of the imputation process, Estimators (8)-(9) follow a fully weighted approach, while Estimators (10)-(11) a full imputation approach (e.g. Särndal and Lundström, 2005, Chap. 12). The former assume a third phase of sampling with response probabilities from Model (1): assumptions similar to those in Kim and Kim (2007) can be established when considering conditional maximum likelihood estimates for the parameters β_{j0} and β_{j1} in $\hat{q}_{kj}$. For the latter, properties depend on the validity of the imputation model considered to obtain y^*_{kj}.

Simulation study and concluding remarks

Due to the lack of space, only results from one of a set of simulation studies are presented, to evaluate the finite sample performance of the estimators proposed in the previous section. We consider the Abortion data set analyzed by Bartholomew et al. (2002) and formed by four binary variables extracted from the 1986 British Social Attitudes Survey and concerning attitude to abortion. $N = 379$ individuals answered to the following questions after being asked if they agreed that the law should allow abortion under the circumstances presented under each item: (1) the woman decides on her own, (2) the couple agrees that they do not wish to have a child, (3) the woman is not married and does not wish to marry the man, and (4) the couple cannot afford any more children. We consider item (2) as our variable of interest y_j, with a population mean given by 59.4%. On the population level, the unit response probabilities were generated under the response function

$$p_k = \frac{\exp(0.7 + y_{k2} + 0.2\varepsilon)}{1 + \exp(0.7 + y_{k2} + 0.2\varepsilon)}, \tag{12}$$

with ε randomly generated from $U(0,1)$ distribution. We have added the variable ε to create several values for p_k since y_{k2} is binary. The population mean of p_k approximately equals 0.7.

To generate item response probabilities $q_{k\ell}$, a proportion of 'no' responses on each item, but none of the 'yes' responses, was changed to missing. Thus, we have forced the nonresponse to depend on the variable of interest. The change to missing was doing with a probability $q_{k\ell}$ randomly generated from $U(0,1)$ distribution. The values $y_{kj} = 1$ corresponded to the 'yes' responses in the initial data set. Thus, for the values $y_{kj} = 1$ we have considered a probability $q_{k\ell}$ of 1. We have drawn $10,000$ simple random samples without replacement from the Abortion data with $n = 50$. In each sample s, units have been classified as respondents according to Poisson sampling, using the probabilities p_k computed in (12) and resulting in the set r. Then, the matrix with entries $\{x_{k\ell}\}_{k\in r,\ell=1,\ldots,4}$ is constructed for each set r. Missing values ($x_{k\ell} = 0$) have been created according to Poisson sampling with the probabilities $q_{k\ell}$ given before. Units with all zero entries have been deleted from the set r.

Estimators (6)-(11) have computed for each replicate. The Horvitz-Thompson estimator, i.e. the sample mean, is also computed as a benchmark in a condition of no item and unit nonresponse and will be denoted by $\widehat{\bar{Y}}_n$. In addition, estimators using the true values for p_k and q_{kj} are also computed. They will be denoted by $\hat{\bar{Y}}^{true}_{j,q\_\text{HT}}$ and $\hat{\bar{Y}}^{true}_{j,q\_\text{Hajek}}$ corresponding to Estimators (8) and (9), respectively. For Estimators (7), (10) and (11), imputed binary values y^*_{k2} have been computed according to Poisson sampling, using each time a probability of 0.25 (1/ number of items). The simulations were carried out in R version 2.12.1, and using the R package 'ltm' (Rizopoulos, 2006) to fit the latent trait models.

The performance of the different estimators is given in Table 1 in terms of the Monte Carlo estimate of the (i) bias, $\hat{E}(\hat{\bar{Y}}) - \bar{Y}$ with $\hat{E}(\hat{\bar{Y}}) = \sum_{i=1}^{10,000} \hat{\bar{Y}}_i/10,000$, (ii) square root of variance, where the latter is $\sum_{i=1}^{10,000}(\hat{\bar{Y}}_i - \hat{E}(\hat{\bar{Y}}))^2/9,999$, (iii) mean squared error and (iv) relative bias, $(\hat{E}(\hat{\bar{Y}}) - \bar{Y})/\bar{Y}$. Estimators (8) and (9) perform better than Estimator (6) in terms of bias. The same conclusion is available for estimators (10) and (11) comparing to Estimator (7). Note that Estimators (8) and (10) reduce the bias more rapidly than Estimators (9) and (11), respectively. Note also that all the

estimators using $\widehat{p}_k$ have a standard deviation slightly larger than the naive estimators, but perform better in term of MSE.

Table 1: Results for the simulation study on the Abortion data: Monte Carlo bias, square root of variance, mean squared error, relative bias of the estimators

Estimator	Equation	bias (%)	$\sqrt{\text{var}}$ (%)	mse	rel bias(%)
$\widehat{\widehat{Y}_n}$		0.1	6.6	0.004	0.0
$\widehat{\widehat{Y}}_{j,\text{naive}}$	(6)	28.5	6.0	0.085	48.0
$\widehat{\widehat{Y}}^{imp}_{j,\text{naive}}$	(7)	15.3	7.0	0.028	25.8
$\widehat{\widehat{Y}}_{j,q\_\text{HT}}$	(8)	3.3	11.8	0.015	5.6
$\widehat{\widehat{Y}}_{j,q\_\text{Hajek}}$	(9)	20.1	11.3	0.053	33.8
$\widehat{\widehat{Y}}^{true}_{j,q\_\text{HT}}$		0.1	7.8	0.006	0.2
$\widehat{\widehat{Y}}^{true}_{j,q\_\text{Hajek}}$		4.2	15.7	0.026	7.0
$\widehat{\widehat{Y}}^{imp}_{j,q\_\text{HT}}$	(10)	-0.3	8.1	0.007	-0.5
$\widehat{\widehat{Y}}^{imp}_{j,q\_\text{Hajek}}$	(11)	10.8	7.6	0.017	18.2

The proposed reweighting system was used in two estimators, a Hájek type estimator and a Horvitz-Thompson type estimator, using both imputation and reweighting methods to deal with item nonresponse. The main goal was to reduce non-response bias in the estimation of the population mean. The previous estimators performed well in our simulation study compared to the naive estimator, and the gain in efficiency was substantial in certain cases.

REFERENCES

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.

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.

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

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.

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

Rubin, D. (1987). *Multiple Imputation for Nonreponse in Surveys*. Wiley, New York.

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