On Auxiliary Variables and Models in Estimation in Surveys with Nonresponse

June 10, 2016

Bernardo João Rota[1,3)] and Thomas Laitila[1,2)]

[1)] Department of Statistics, Örebro University, 701 82 Örebro, Sweden

[2)] Department of Research and Development, Statistics Sweden, 701 89 Örebro, Sweden

[3)] Statistics, Department of Mathematics and Informatics, Eduardo Mondlane University, Maputo, Mozambique.

Abstract

This paper gives a brief discussion on two alternative weighting procedures. Weighting with and without explicitly modeling for the response mechanism, which are known as the *direct weighting* and the *weighting* approaches. The generalized regression estimator benchmarks the weighting methods while a general double weighted Horvitz-Thompsom estimator represents the direct weighting approach. A general reliance on the strenght of the correlation between the auxiliary variables, the response behavior and the study variables prevailing mostly on weighting approaches, is shown to be inappropriate in some cases, that is, it increases the bias of the resulting estimator. On the other hand, the traditional use of simple models in representation of the true response behavior is addressed through an example in which it is shown to be adequate under very specific assumptions.

1 Introduction

In adjusting for nonresponse, weighting is a commonly used approach by survey methodologists. Weighting relies on auxiliary variables, which can be defined as variables on

which information is available for respondents and nonrespondents. In weighting for nonresponse adjustment the role of auxiliary variables is crucial in reducing the nonresponse errors. Rizzo, Kalton, and Brick (1996), pointed out that the selection of auxiliary variables could be more important than the weighting scheme. Furthermore, Särndal (2011) also claims that in case of bias inflation by nonresponse, access to powerful auxiliary variables becomes the key in minimizing the problem. These auxiliary variables are demanded to predict: (a) the propensities to respond and (b) the key survey variables to effectively adjust for nonresponse (West and Little, 2012).

Use of auxiliary variables in estimation can be found in for example, Bethlehem (1988), Estevão and Särndal (2000), Kalton and Flores-Cervantes (2003), Särndal and Lundström (2005), Särndal (2007). In practice there is a wide choice of variables (Särndal and Lundström, 2008), and one needs to decide on their selection for effective adjustment. The literature provides some suggestions to guide in selection of auxiliary variables. Särndal and Lundström (2008) propose a selection device based on the variability of the reciprocals of estimated propensities. The propensities are determined under the assumption that the auxiliary variables satisfy some pre-specified condition. Geuzinge, Rooijen and Bakker (2000) propose a selection indicator based on a product of correlations arising from (a) and (b). In adjusting for nonresponse using the regression estimator, Bethlehem and Schouten (2004) and Schouten (2007) propose a selection based on minimizing a maximal absolute bias of the estimator. The method relies on computing an interval for the maximal absolute bias and selecting those variables that minimize its width.

Searching for auxiliary variables satisfying the requirements (a) and (b) simultaneously may be a difficulty task. Survey practices involve many variables of interest, as Kott (2013) comments on Brick's (2013) discussion paper, one can seldom encounter auxiliary variables fulfilling (b) for every variable of interest in a multipurpose survey. Kreuter and Olson (2011) also noted the same difficulty. Furthermore, as we illustrate with a simple example in Section 3, fulfilling requirements (a) and (b) simultaneously does not generally guaranty effectiveness in bias protection for target estimates, it may even introduce rather than remove the bias. Perhaps it is on the nonresponse adjustment methods that do not explicitly model the response behavior that prevail the requirement (a) and (b) to effective adjust. We illustrate through an example that it may not be appropriate to entirely rely on correlation relations between the variables involved in the study. Adjustment methods where the response behavior is explicitly modeled the primary goal is in observing those variables that are linked to response pattern, thus, the estimation of targets is viewed as second objective after the estimation of the response model. However, the approach is also challenging in the sense that it is hard or even impossible to guess the appropriate response behavior in which some may have simple forms while others complex. Simple models like the logit and probit models are usually used in representation of the true response model. We use a telephone survey case to show that such simple models are

adequate under very specific assumptions.

2 Weighting for nonresponse adjustment

Suppose a sample $s = \{1, 2, ..., k, ..., n\}$ of size n is drawn from a population $U = \{1, 2, ..., k, ..., N\}$ of size N with a probability sampling design $p(s)$, yielding sample inclusion probabilities $\pi_k = \Pr(k \epsilon s) > 0$ and corresponding design weights $d_k = 1/\pi_k$ for all $k \in U$. Let y and $\mathbf{x}$ be the survey variable of interest and an L-dimensional column vector of auxiliary variables, respectively. We want to estimate $Y = \sum_U y_k$.

In the prospect of producing good adjustment weights, the weighting methods relie on the proper use of available auxiliary information (e.g. Falk, 2012; Brick, 2013). We can emphasize here weighting in two directions, that is, with and without an explicit modeling of response function. The paper by Kim and Kim (2007) points out that the weighting procedures for nonresponse adjustment are mainly made by applying one of the two approaches: the *weighting adjustment* or the *direct weighting adjustment*.

2.1 The weighting adjustment

In the *weighting adjustment* the auxiliary information is embedded into the estimation of targets in which case it improves the efficiency of the resulting estimators. The generalized regression (GREG) estimator is an example of this kind of adjustment.

Assume the following relation between y and $\mathbf{x}$ described through the model:

$$\zeta : y_k = \boldsymbol{\beta}^t \mathbf{x}_k + \varepsilon_k, k = 1, .., N \tag{1}$$

where $\boldsymbol{\beta}$ is an L-dimensional column vector of model parameters and ε_k is a zero-mean random variable with $V_\zeta(\varepsilon_k) = \sigma_k^2$.

The generalized regression (GREG) estimators for Y based on the relation between y and $\mathbf{x}$ given by equation (1), are class of estimators of the form

$$\hat{Y}_{reg} = \left(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k \right)^t \hat{B}_s + \sum_s d_k y_k \tag{2}$$

where $\hat{B}_s = \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k^t\right)^{-1} \sum_s d_k \mathbf{x}_k y_k$.

According to Cobben (2009), the GREG estimator was introduced by Särndal (1980) and Bethlehem and Keller (1987). The GREG estimator (2) is extensively studied in Särndal, et. al (1992), its properties rely on the sampling design and close linear fit between y and $\mathbf{x}$, without explicitly depending on whether (1) is true or not. In this setting, the regression estimator (2) is deemed model assisted rather than dependent (Särndal, 2007). The model dependent regression estimator is extensively reviewed in Fuller (2002). The regression estimator can be written in a simpler form as a weighted sum of the values of the survey variable. This is done by writing $\left(\sum_U \mathbf{x}_k - \sum_s d_k\mathbf{x}_k\right)^t \left(\sum_s d_k\mathbf{x}_k\mathbf{x}_k^t\right)^{-1} \sum_s d_k\mathbf{x}_k y_k$ as $\sum_s d_k M_k y_k$, where $M_k = \left(\sum_U \mathbf{x}_k - \sum_s d_k\mathbf{x}_k\right)^t \left(\sum_s d_k\mathbf{x}_k\mathbf{x}_k^t\right)^{-1} \mathbf{x}_k$. Thus, equation (2) becomes

$$\hat{Y}_{reg} = \sum_s w_k y_k \tag{3}$$

with $w_k = d_k(1 + M_k)$.

This particular form of the regression estimator is advantageous in that the weights w_k can be applied to any survey variable and have the following property:

$$\sum_s w_k\mathbf{x}_k = \sum_U \mathbf{x}_k \tag{4}$$

Furthermore, when $\sum_U \mathbf{x}_k$ can be constructed by just adding up $\mathbf{x}_k$ in the sampling frame, a number of regression estimators can be constructed. However, observing $\mathbf{x}_k, k = 1, .., N$ is not a requirement for the regression estimator based on (1), it suffices to know only $\sum_U \mathbf{x}_k$, which can be information obtained from others sources.

Letting $M_k^* = \left(\sum_U \mathbf{x}_k - \sum_s d_k\mathbf{x}_k\right)^t \left(\sum_s d_k\mathbf{z}_k\mathbf{x}_k^t\right)^{-1} \mathbf{z}_k$, where $\mathbf{z}_k$ is a vector of auxiliary variables conceptually different, but of the same dimension as $\mathbf{x}_k$, we obtain a more general regression estimator given in (5). In this case the regression estimator resembles the instrumental variable regression estimator learned from econometric theory.

$$\hat{Y}_{IVreg} = \left(\sum_U \mathbf{x}_k - \sum_s d_k\mathbf{x}_k\right)^t \hat{B}_{IVs} + \sum_s d_k y_k. \tag{5}$$

where $\hat{B}_{IVs} = \left(\sum_s d_k\mathbf{z}_k\mathbf{x}_k^t\right)^{-1} \sum_s d_k\mathbf{z}_k y_k$.

In our case where sampling is followed by nonresponse (e.g. An, 1996; Fuller and An, 1998; Singh and Kumar, 2011), the GREG estimator (5) is given by

$$\hat{Y}_{IVreg^*} = \left(\sum_U \mathbf{x}_k - \sum_r d_k\mathbf{x}_k\right)^t \hat{B}_{IVr} + \sum_r d_k y_k \tag{6}$$

where $\hat{B}_{IVr} = \left(\sum_r d_k\mathbf{z}_k\mathbf{x}_k^t\right)^{-1} \sum_r d_k\mathbf{z}_k y_k$ and r is the set of respondents.

Let us assume that the condition $\boldsymbol{\lambda}^t\mathbf{z}_k = 1$ holds for all k, where $\boldsymbol{\lambda}$ is independent of k.

The equation (6) becomes

$$\hat{Y}_{IVreg^*} = \sum_U \mathbf{x}_k^t \hat{B}_{IVr} \tag{7}$$

The expected mean of $\hat{Y}_{IVreg^*}$ is

$$E\left(\hat{Y}_{IVreg^*}\right) \approx \sum_U \mathbf{x}_k^t B_{IV\theta} \tag{8}$$

where $B_{IV\theta} = \left(\sum_U \theta_k \mathbf{x}_k \mathbf{z}_k^t\right)^{-1} \sum_U \theta_k \mathbf{z}_k y_k$ and $\theta_k = \Pr(k \epsilon r | k \epsilon s)$.
Equation (8) says that the bias of the regression estimator is almost entirely depending on the properties of the response based regression coefficients $\hat{B}_{IVr}$. If all $\theta_k = 1$, then the regression estimator is approximately unbiased for $\sum_U y_k$. The following interesting statement is given by Cobben (2009): "*Practical experience (at least in the Netherlands) shows that nonresponse often seriously affects estimators like means and totals, but less often causes estimates of relationships to be biased. Particularly if relationships are strong, i.e. the regression line fits the data well, the risk of finding wrong relationships is small*". Furthermore, Bethlehem (1988) show to hold the following relation between the $B_{IV\theta}$ and $B = \left(\sum_U \mathbf{z}_k \mathbf{x}_k^t\right)^{-1} \sum_U \mathbf{z}_k y_k$:

$$B_{IV\theta} - B = \left(\sum_U \theta_k \mathbf{z}_k^t \mathbf{x}_k\right)^{-1} \sum_U \theta_k e_k \tag{9}$$

where $e_k = y_k - \mathbf{x}_k^t B$.
Equation (9) says that the estimator (7) is approximately unbiased for Y if the linear fit between y_k and $\mathbf{x}_k^t$ is strong or the regression errors are uncorrelated with the response probabilities. Thus, the need for strong relatioships between the explanatory variables, the response probabilities and the variables of interest.

- **The auxiliary variable may introduce bias**

Fuller and An (1998) enphasize that the level of bias reduction depends on the relations between the auxiliary variable, the variable of interest, and the response probability. In this section we provide a simple example in which a candidate auxiliary variable satisfying the requirements of being correlated with both the variable of interest and the probability of response turns the resulting GREG estimator biased while a simple expanded HT estimator provides with approximately unbiased estimation.

Let us assume the following relationship:

$$\begin{aligned} y_k &= \beta_0 + \beta_1 x_k + e_k \\ \theta_k &= \Pr(k \epsilon r | k \epsilon s) \end{aligned} \tag{10}$$

Assume the following conditions in (10):

1. $\sum_U e_k = 0$ and $\sum_U x_k e_k = 0$
2. $\sum_U \theta_k y_k = \sum_U \theta_k \sum_U y_k / N$
3. $\sum_U \theta_k x_k \neq 0$ and $\sum_U \theta_k e_k \neq 0$

Then, the expanded Horvitz-Thompson estimator for the total of y is approximately unbiased. To obtain this result, let

$$\hat{Y}_{\text{exp}} = \frac{N}{\sum_r d_k} \sum_r d_k y_k$$

Then

$$NearBias(\hat{Y}_{exp}) = N \frac{E\left(\sum_r d_k y_k\right)}{E\left(\sum_r d_k\right)} - \sum_U y_k = N \frac{\sum_U \theta_k y_k}{\sum_U \theta_k} - \sum_U y_k$$

$$= N \frac{\sum_U \theta_k y_k - \sum_U \theta_k \sum_U y_k}{\sum_U \theta_k} = \frac{N^2 cov(\theta_k, y_k)}{N\bar{\theta}} = N \frac{cov(\theta_k, y_k)}{\bar{\theta}} = 0$$

because $cov(\theta, y) = 0$.

One special case of the regression estimator is the ratio estimator. It is obtained from (6) by setting $\mathbf{x}_k = x_k$ and $\mathbf{z}_k = 1$. The ratio estimator is in the literature suggested to have smaller bias due to nonresponse than the expansion estimator. Then the approximate bias for the GREG estimator is given by

$$NearBias(\hat{Y}_{RA}) = \frac{N^2 \bar{x} \sigma_{\theta e}}{\sum_U \theta_k x_k} \tag{11}$$

where $\bar{x} = N^{-1} \sum_U x_k$, $\sigma_{\theta e} = cov(\theta, e)$ and in the model (10) we assume $\beta_0 = 0$. Thus, generally the ratio estimator has a nonzero approximate bias resulting from the choice of auxiliary variable.

Defining $\mathbf{x}_k = \mathbf{z}_k = \begin{pmatrix} 1 & x_k \end{pmatrix}^t$ equation (7) gives another well known estimator, the simple linear regression, which is also suggested to be more efficient than the expansion estimator. In this case the bias of the regression estimator is given by:

$$NearBias(\hat{Y}_{reg}) = \frac{\sigma_{xF} \sum_U \theta_k e_k - \sigma_{x\theta} \sum_U F_k e_k}{\bar{\theta} \bar{F} x - \bar{F}^2} \tag{12}$$

where $F_k = \theta_k x_k$, $\sigma_{xF} = cov(x, F)$, $\sigma_{x\theta} = cov(x, \theta)$, $\bar{F} = N^{-1} \sum_U F_k$, $\bar{F}x = N^{-1} \sum_U F_k x_k$ and in (10) we longer assume $\beta_0 = 0$.

Again the bias of this estimator is generaly nonzero. Simple examples considered here show that the recommendation of selecting "powerful" auxiliary variables in the sense of being correlated with variables of interest and response probability may introduce bias due to nonresponse instead of reducing it.

2.2 The direct weighting adjustment

In the *direct weighting adjustment* it is assumed that the functional form of the response probability is known and given by $\theta_k = p(\cdot \mathbf{z}_k)$, where $\mathbf{z}_k$ is a vector of model variables. The primary goal is to estimate this function so that the observed values of the target variable are double weighted, that is, each y_k is multiplied by $d_k \hat{\theta}_k^{-1}$. The target population Y can then be estimated by

$$\hat{Y}_{nr} = \sum_r d_k \hat{\theta}_k^{-1} y_k \tag{13}$$

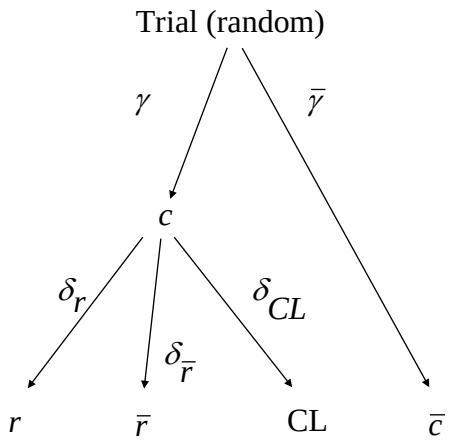
The estimator (13) is widely suggested in the literature of nonresponse adjustment (see e.g. Chang and Kott, 2008; Kim and Park, 2010; Kim and Riddles, 2012). The properties of $\hat{Y}_{nr}$ are conditioned on the properties of $\hat{\theta}$. For example, consistency of $\hat{Y}_{nr}$ may depend on the correct specification of the function θ. Thus wrongly specified θ may lead into an inconsistency of $\hat{Y}_{nr}$. Given the limitation on knowlegde of the response mechanism (Särndal and Lundström, 2005) it is difficult to suggest whether a proposed response mechanism is the appropriate or no. Simple models as logit and probit have been widely used in application (e.g. Chang and Kott, 2008). An immediate question raised is when can these simple models be used? We discuss this issue in the following example of nonresponse in a telephone survey.

- **Response probability modeling**

An attempt to make contact with a unit in the sample and collect a response can be seen as a random trial. The possible outcomes in one attempt in a telephone survey are illustrated in Figure 1. When calling it may result in a contact (c) with probability γ, or a fail in making a contact ($\bar{c}$) with probability $(1 - \gamma)$.

Given a contact is made it may result in a response (r), a refusal to participate ($\bar{r}$), or an agreement to call back later (CL). Conditionally on c, let the probabilities of these outcomes be denoted by δ_r, $\delta_{\bar{r}}$, and δ_{CL}, respectively.

Factors affecting the probability of a contact include the telephone number is not correct, the unit cannot at the time respond to a telephone call and, the respondent is not willing to respond to an unknown incoming call number. If a contact is made, other factors are involved in a decision to respond, refuse or agree to a contact later. The presentation of the survey, the topic of the survey and the time required to respond come into play.



r = response, $\bar{r}$ = refusal, CL = call later,

$\bar{c}$ = no contact, γ = probability of contact (c),

δ_a = probability of outcome a given c

Figure 1: Tree diagram of potential outcome of a contact trial in a telephone survey.

From the tree diagram, the probabilities of the outcomes of the trial are $Pr(r) = \theta_1 = \gamma\delta_r$, $Pr(\bar{r}) = \theta_2 = \gamma\delta_{\bar{r}}$, $Pr(CL) = \theta_3 = \gamma\delta_{CL}$ and $Pr(\bar{c}) = (1-\gamma) = \bar{\gamma}$. If the outcome of the trial is either a fail to make contact or an agreement to call back, a second trial to get a response from the unit can be made. The same potential outcomes are possible. However, it is here assumed that given an agreement of calling back later, contact is made and the outcome is either a response (r) or a nonresponse ($\bar{r}$).

Consider a sequence of contact trials and let $\mathbf{P}_t$ denote a column vector of probabilities of the outcomes (r,$\bar{r}$, CL, $\bar{c}$) at trial t. The sequence of trials can be modeled as a stochastic process with a transition matrix $\mathbf{\Gamma}_t$. For $t \geq 2$ this matrix contains probabilities conditionally on the outcome of the $(t-1)$th trial. Thus, $P_t = \mathbf{\Gamma}_t\mathbf{P}_{t-1}(t \geq 2)$ with $\boldsymbol{P}_1 = \begin{pmatrix} \theta_1 & \theta_2 & \theta_3 & \bar{\gamma} \end{pmatrix}^t$.

There are some information on the transition matrix. First, the response and nonresponse outcomes are absorbing, and above the outcome CL is assumed to yield either a response or a nonresponse in the following trial. If $\bar{\gamma} < 1$ these assumptions will eventually yield either a response or a nonresponse in a sequence of trials.

Suppose $\mathbf{\Gamma}_t = \mathbf{\Gamma}_2$ for all $t \geq 2$, and consider

$$\mathbf{\Gamma}_2 = \begin{pmatrix} 1 & 0 & \theta_{31} & \theta_1 \\ 0 & 1 & \theta_{32} & \theta_2 \\ 0 & 0 & 0 & \theta_3 \\ 0 & 0 & 0 & \bar{\gamma} \end{pmatrix}$$

The fourth column in the matrix equals $\boldsymbol{P}_1$, meaning the conditional probabilities in a trial following upon a series of no contacts are the same, they do not change with the number of trials earlier made. With this model the probabilities of the different outcomes can be expressed as $\boldsymbol{P}_t = \mathbf{\Gamma}_2^{t-1}\mathbf{P}_1$ and letting the number of trials converge to infinity yields the probability vector

$$\boldsymbol{P}_\infty = \begin{pmatrix} \lambda_r + \theta_{31}\lambda_{CL} \\ \lambda_{\bar{r}} + \theta_{32}\lambda_{CL} \\ 0 \\ 0 \end{pmatrix}$$

using the definitions of $\theta_j (j = 1, 2, 3)$. This model then shows the probability of a response from a unit being made up of three unknown probabilities, i.e. $Pr(r) = \lambda_r + \theta_{31}\lambda_{CL}$.

Adding the assumption $\lambda_{CL} = 0$ yields the traditional dichotomy of response/nonresponse suggesting modeling λ_r with e.g. a normal or a logistic distribution function. The same modeling approach can also be motivated if $\theta_{31} = 0$.

A different case is obtained by setting $\theta_{31} = 1$ whereby $Pr(r) = \lambda_r + \lambda_{CL}$, and the modeling of $Pr(r)$ using probit or logit models is less appropriate. Rather, these models imply modeling of nonresponse, i.e. $Pr(\bar{r})$, due to their symmetry. With a distribution function F having a symmetric density and $Pr(\bar{r}) = F(-\mathbf{x}^t\alpha)$,then $Pr(r) = 1 - Pr(\bar{r}) = F(\mathbf{x}^t\alpha)$. If the distribution is asymmetric it's appropriate to model nonresponse instead of response.

A final special case of interest is obtained with $\theta_{31} = \lambda_r/(\lambda_r + \lambda_{\bar{r}})$, which corresponds to the independence of irrelevant alternative (IIA) assumption underlying the conditional logit model (Luce, 1959). Under the IIA assumption $Pr(r) = \lambda_r/(\lambda_r + \lambda_{\bar{r}})$. Now suppose $\lambda_a = e^{V_a}/(e^{V_r} + e^{V_{\bar{r}}} + e^{V_{CL}})$ $(a\epsilon\{r, \bar{r}\})$ where $V_\cdot$ are nonrandom scalars. Then $Pr(r) = e^{V_r}/(e^{V_r} + e^{V_{\bar{r}}}) = e^{V_D}/(1 + e^{V_D})$, where $D = V_r - V_{\bar{r}}$, and the logit model is obtained.

In the discrete choice literature (e.g. McFadden, 1974) V_r, $V_{\bar{r}}$ and V_{CL} represent systematic parts of the utilities of choosing the alternatives r, $\bar{r}$ and CL, respectively. The utilities for the units are obtained by adding individual specific components ϵ_a $(a\epsilon\{r, \bar{r}, CL\})$ yielding $U_a = V_a + \epsilon_a$. Under the maximum utility paradigm, the unit selects the alternative yielding maximum utility, that is a unit responds if $U_r > max(U_{\bar{r}}, U_{CL})$.

Let $\mathbf{x}$ denote a vector characterizing the respondent and $V_a = \mathbf{x}^t\alpha_a$ such that $U_a = \mathbf{x}^t\alpha_a + \epsilon_a$ $(a\epsilon\{r, \bar{r}, CL\})$. Suppose ϵ_a $(a\epsilon\{r, \bar{r}, CL\})$ are independent and identically Gumbel

distributed, then $Pr(r) = e^{\mathbf{x}^t\alpha_D}/(1+e^{\mathbf{x}^t\alpha_D})$ where $\alpha_D = \alpha_r - \alpha_{\bar{r}}$ (e.g. McFadden, 1974). Again the logit model is obtained.

3 Discussion

Sampling theory shows how to utilize randomization for valid, objective inference from empirical observations. Its application in the social sciences and for official statistics production is, however, hampered by nonresponse, because the theory assumes observations are obtained for all units in the sample. Thus, this excellent theory cannot be applied as is in practice.

There are early suggestions on how to correct for nonresponse where the theory is applied in two or more steps. One example is the Hansen and Hurwitz (1946) method, where a subset of the set of nonrespondents are sampled and measured. A similar idea is given by Bartholomew (1961). Again, however, for these theories to work in practice, full response is required when sampling from the subset of nonrespondents. Another problem is the extra time required for completion of the study.

Later the view of response being an outcome of a random trial was adopted. Oh and Scheuren (1983) consider this interpretation as a quasi randomization approach when treating the response set generated as a second sampling phase with an unknown second phase sampling design. However, the idea makes standard theory on estimation applicable by using estimated response probabilities.

Using a model means by its definition use of approximations. A model cannot be assumed correct and valid inference cannot be guaranteed from its application. Estimators based on estimation of response probability functions are therefore biased and inconsistent and the size of bias is unknown. An essential part here is how well the model approximates the true response probabilities. One popular alternative is the simple binary logit model.

This paper contributes with conditions under which the logit model is correct and, it is interesting to note how contributions in the discrete choice literature can be adapted in modeling response probabilities. The results presented here are based on a proposed model of the possible outcomes of a contact trial, and the logit model is obtained under very restrictive assumptions on this model. In particular the response probability is obtained as a function of several probabilities of different events. An approximation with the cdf of the logistic distribution is therefore too simple.

Another approach for nonresponse adjustments based on the random response interpretation is weighting, where auxiliary information are used to adjust design weights to capture response probability patterns. Here the method does not require a known form of or the variables in the response probability function. There are plenty results in the literature showing this approach to be successful in reducing bias due to nonresponse.

In this paper a question rarely raised in the literature is considered and can be formulated as: how does weighting affect estimates if the response set mean is unbiased? One potential reason for this problem not being addressed is, the adaptation of concepts on the relation between the study variable and the generation of the response set from the model based inference literature, e.g. MAR (missing at random) and MCAR (missing completely at random), and ignorable and nonignorable nonresponse.

For estimation of population means or totals in the finite population framework such concepts may be misleading. MCAR is a stronger concept than MAR, usually meaning that if MCAR holds, so does MAR. Methods derived to handle MAR cases then also encompasses the MCAR cases. However, this may not be true in the finite population context for similar concepts; MCAR might hold but not MAR.

If MCAR is defined as $\sum_U \theta_k y_k = \bar{\theta} \sum_U y_k$, and $U_x = \{x : x_k = x, k \epsilon U\}$ then MCAR does not imply $\sum_{U_x} \theta_k y_k = \bar{\theta} \sum_{U_x} y_k$. The same argument can be derived by considering a random draw from the population and observing y and R. Then MCAR defined as $F(y|R = 1) = F(y|R = 0)$ does not imply $F(y|R = 1, x) = F(y|R = 0, x)$, where $F(\cdot)$ denotes the cdf.

4 Conclusions

Results show weighting adjustments for nonresponse may yield biased estimators while the simple expansion estimator is approximately unbiased. This issue has to be considered when choosing auxiliary variables and new indicators or tests must be developed.

Simple models of response probabilities do not capture the complex process of attempts for contacts and choice of the unit to respond or not. This discrepancy is a source of bias and models capturing the characteristics of the data collection process have to be developed. Graph models in combination with models for discrete choice data may here provide new tools for modeling response probabilities.

5 References

An, A.B. (1996) Regression estimation for finite population means in the presence of nonresponse. Retrospective Theses and Dissertations. Paper 11357.

Bartholomew, B.J. (1961). A method for allowing for 'not-at-home' bias in sample surveys, *Journal of the Royal Statistical Society, Series C*, 10:1, 52-59.

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

Bethlehem, J. and B. Schouten (2004). Nonresponse adjustment in household surveys, *Discussion paper 04007. Statistics Netherlands, Voorburg/Heerlen, The Netherlands.*

Brick, J.M. (2013). Unit Nonresponse and Weighting Adjustments: A Critical Review, *Journal of Official Statistics*, **29:3**, 329–353.

Chang, T. and P.S. Kott (2008). Using calibration weighting to adjust for nonresponse under a plausible model, *Biometrika*, **95:3**, 555–571.

Cobben, F. (2009). Nonresponse in sample surveys : methods for analysis and adjustment, *Statistics Netherlands*, *Downloaded from UvA-DARE, the institutional repository of the University of Amsterdam (UvA) http://hdl.handle.net/11245/2.69219*

Estevão, V.M. and C.E. Särndal (2000). A functional form approach to calibration, *Journal of Official Statistics*, **16:4**, 379–399.

Falk. G. (2012). Calibration adjustment for nonresponse in cross-classified data, *Section on Survey Research-JSM 201.*

Fuller, W.A. (2002). Regression Estimation for Survey Samples, *Survey Methodology*, **28:1** 5–23.

Fuller, W. A. and An, A.B. (1998). Regression Adjustment for Nonresponse, *Jour. Ind. Soc. Ag. Statistics,* **51**, 331–342.

Hansen, M.H. and W.N. Hurwitz (1946). The problem of non-response in sample surveys, *Journal of the American Statistical Association*, 41:236, 517-529.

Geuzinge, L., Rooijen, J. van and B.F.M. Bakker (2000), The use of administrative registers to reduce non-response bias in household surveys, *Netherlands Official Statistics* **2000:2**, 32–39.

Hansen, M.H. and Hurwitz, W.N. (1946) The Problem of Non-Response in Sample Surveys. *Journal of the American Statistical Association*, **41:236**, 517–529.

Holt, D. and D. Elliot (1991). Methods of weighting for unit non-response, *Journal of the Royal Statistical Society. Series D (The Statistician),* **40:3**,333-342.

Kalton, G. and I. Flores-Cervantes (2003). Weighting methods, *Journal of Official Statistics,* **19:2**, 81–97.

Kim, J.K. and J.J. Kim (2007). Nonresponse weighting adjustment using estimated response probabilities, *The Canadian Journal of Statistics,* **35:4**, 501–514.

Kim, J.K. and Park, M. (2010). *International Statistical Review,* **78**, 21-39.

Kim, J.K. and Riddles, M.K. (2012). Some theory for propensity-score-adjustment estimators in survey sampling, *Survey Methodology*, **38:2**, 157–165.

Kott, P.S. (2013). Discussion, *Journal of Official Statistics*, **29:3**, 359–362

Kreuter, F. and K. Olson (2011). Multiple Auxiliary Variables in Nonresponse Adjustment. *Sociological Methods & Research,* **40:2**, 311–332.

Luce, R.D. (1959). *Individual Choice Behavior: A Theoretical Analysis*, Wiley, New York.

Mandell, L. (1974). When to weight: Determining nonresponse bias in survey data, *America Association for Public Opinion Research,* **38:2**, 247-252.

McFadden, D. (1974). The Measurement of Urband Travel Demand, *Journal of Public Economics*, 3, 303-328.

Oh, H.L. and F.J. Scheuren (1983). Weighting adjustment for unit nonresponse. In: Madow, W.G, Olkin, I. and D.B. Rubin (Eds.), Incomplete Data in Sample Surveys:Vol 2, Academic Press, New York, pp. 143-184.

Rizzo, L., Kalton, G., and M. Brick (1996). A comparison of some weighting adjustment methods for panel

nonresponse. *Survey Methodology*, **22**, 43-53.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice, *Survey Methodology,***33:2**, 99-119.

Särndal, C.-E. (2011). Three Factors to Signal Non-Response Bias With Applications to Categorical Auxiliary Variables, *International Statistical Review,* **79:2**, 233–254.

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

Särndal, C.-E. and S. Lundström (2008). Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator. *Journal of Official Statistics*, **24:2**, 167–191

Särndal, C.-E, Swensson, B. and J. Wretman (1992). *Model Assisted Survey Sampling.* Springer, New York.

Schouten, B. (2007). A selection strategy for weighting variables under a Not-Missing-at-Random assumption, *Journal of Official Statistics,* **23**, 51–68.

Singh, H.P. and Kumar, S. (2011). Combination of regression and ratio estimate in presence of nonresponse. *Brazilian Journal of Probability and Statistics,* **25:2**, 205–217 DOI: 10.1214/10-BJPS117

West, B.T. and R.J.A. Little (2012). Non-response adjustment of survey estimates based on auxiliary variables subject to error, *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **176**, 211–225.