RESPONSE MODEL SELECTION IN SMALL AREA ESTIMATION UNDER NOT MISSING AT RANDOM NONRESPONSE

Michael Sverchkov
Bureau of Labor Statistics, Washington DC, USA

Danny Pfeffermann
Hebrew University of Jerusalem, Israel, and University of Southampton, UK.

ABSTRACT

Sverchkov and Pfeffermann (2018) consider Small Area Estimation (SAE) under informative probability sampling of areas and within the sampled areas, and not missing at random (NMAR) nonresponse. To account for the nonresponse, the authors assume a given response model, which contains the outcome values as one of the covariates and estimate the corresponding response probabilities by application of the Missing Information Principle, which consists of defining the likelihood as if there was complete response and then integrating out the unobserved outcomes from the likelihood by employing the relationship between the distributions of the observed and unobserved data.

A key condition for the success of this approach is the "correct" specification of the response model. In this article we consider the likelihood ratio test and information criteria based on the appropriate likelihood and show how they can be used for the selection of the response model. We illustrate the approach by a small simulation study.

Key words: AIC, BIC information criteria, likelihood ratio test, population distribution, respondents' model, sample distribution,

1. INTRODUCTION

There exists almost no survey without nonresponse, but in practice most methods that deal with this problem assume either explicitly or implicitly that the missing data are 'missing at random' (MAR). However, in many practical situations, this assumption is not valid, since the probability to respond often depends on the outcome value, even after conditioning on available covariate information. In such cases, the use of methods that assume that the nonresponse is MAR can lead to large bias of parameter estimators and distort subsequent inference.

The case where the missing data are not MAR (NMAR) can be treated by postulating a parametric model for the distribution of the outcomes before non-response and a model for the response mechanism. These two models define a parametric model for the observed outcomes, so that the parameters of these models can be estimated from the observed data. See, for example, Pfeffermann and Sverchkov (2009) for details, with overview of related literature.

Modeling the distribution of the outcomes before non-response can be problematic since only the observed data are available. Sverchkov (2008) proposes an alternative approach, which allows to estimate the parameters of the response model without postulating a parametric model for the distribution of the outcomes before nonresponse. To account for the nonresponse, Sverchkov (2008) assumes a given response model and estimates the corresponding response probabilities by application of the missing information principle (MIP), which consists of defining the likelihood as if there was complete response, and then integrating out the unobserved outcomes from the likelihood, employing the relationship between the distributions of the observed and unobserved data. Sverchkov and Pfeffermann (2018) apply this approach for small area estimation (SAE) under informative probability sampling of areas and within the sampled areas, and NMAR nonresponse. We describe the main steps of this approach in Sections 2 and 3.

A key condition for the success of this approach is the "correct" specification of the response model. In section 4 we consider the likelihood ratio test and information criteria based on the appropriate likelihood and show how they can be used for the selection of the response model. Section 5 illustrates the application of the approach by a small simulation study.

2. NOTATION AND MODELS

Let $\{y_{ij}, \mathbf{x}_{ij}; i = 1,...,M,\ j = 1,...,N_i\}$ represent the data in a finite population of N units, comprised of M areas with N_i units in area i, $\sum_{i=1}^{M} N_i = N$, where y_{ij} is the value of the outcome variable for unit j in area i and $\mathbf{x}'_{ij} = (x_{ij,1},..., x_{ij,K})$ is a vector of corresponding K covariates. We assume that the covariates are known for every unit in the population. Suppose that the population outcome values follow the generic two-level model:

$$\begin{aligned} &y_{ij} \mid \mathbf{x}_{ij}, u_i^U \sim f(y_{ij} \mid \mathbf{x}_{ij}, u_i^U),\ i = 1,...,M,\ j = 1,...,N_i \\ &u_i^U \sim f(u_i^U);\ E(u_i^U) = 0,\ V(u_i^U) = \sigma^2_{u^U}, \end{aligned} \tag{2.1}$$

where u_i^U is the i^{th} area level random effect. The target is to estimate the area means $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}, i = 1,...,M$, based on a sample obtained by the following two-stage sampling scheme: **i)**- select a sample s of m out of the M population areas with inclusion probabilities $\pi_i = \Pr(i \in s)$; **ii)** select a sample s_i of $n_i > 0$ units from selected area i with probabilities $\pi_{j|i} = \Pr(j \in s_i \mid i \in s)$. Denote by I_i, I_{ij} the sample indicators; $I_i = 1$ if area i is selected in the first stage and 0 otherwise, $I_{ij} = 1$ if unit j of selected area i is sampled in the second stage and $I_{ij} = 0$ otherwise. Let $w_i = 1/\pi_i, w_{j|i} = 1/\pi_{j|i}$ denote the first- and second-stage sampling weights.

In practice, not every unit in the sample responds. Define the response indicator; $R_{ij} = 1$ if unit $j \in s_i$ responds and $R_{ij} = 0$ otherwise. The sample of respondents is

thus $R=\{(i,j): I_i=1, I_{ij}=1, R_{ij}=1\}$ and the sample of nonrespondents among the sampled units is $R^c=\{(i,k): I_i=1, I_{ik}=1, R_{ik}=0\}$. The response process is assumed to occur stochastically, independently between units. We assume $\sum_{j=1}^{n_i} R_{ij} > 0$ in all the sampled areas. The sample of respondents defines therefore a third, self-selected stage of the sampling process with unknown response probabilities. (Särndal and Swensson, 1987).

Define, $u_i = u_i^U - E(u_i^U \mid i \in s)$. Then, under the population model (2.1), the observed data follow the two-level 'respondents' model:

$$\begin{aligned} &f_R(y_{ij} \mid \mathbf{x}_{ij}, u_i) = f(y_{ij} \mid \mathbf{x}_{ij}, u_i, (i,j) \in R); \\ &u_i \sim f(u_i \mid i \in s),\ E(u_i \mid i \in s) = 0. \end{aligned} \tag{2.2}$$

The model (2.2) is again general and all that we state at this stage is that under informative sampling and/or NMAR nonresponse, the population and the respondents' models differ; $f_R(y_{ij} \mid \mathbf{x}_{ij}, u_i) \neq f(y_{ij} \mid \mathbf{x}_{ij}, u_i^U)$.

Remark 1. The respondents' model refers to the observed data and hence can be estimated and tested by standard SAE methods. See Pfeffermann (2013) and Rao and Molina (2015) for estimation and testing procedures in SAE, with references.

Let $p_r(y_{ij}, \mathbf{x}_{ij}) = \Pr[R_{ij}=1 \mid y_{ij}, \mathbf{x}_{ij}, i \in s, j \in s_i]$. If the probabilities $p_r(y_{ij}, \mathbf{x}_{ij})$ were known, the sample of respondents could be considered as a two-stage sample from the finite population with known sampling probabilities π_i and $\tilde{\pi}_{j|i} = \pi_{j|i} p_r(y_{ij}, \mathbf{x}_{ij})$. In this case, the area means $\bar{Y}_i$ can be estimated as in Pfeffermann and Sverchkov (2007). Also, if known, the response probabilities could be used for imputation of the missing data within the selected areas, by application of the relationship between the sample and sample-complement distributions, (Sverchkov and Pfeffermann, 2004);

$$f(y_{ij} \mid \mathbf{x}_{ij}, u_i, (i,j) \in R^c) = \frac{[p_r^{-1}(y_{ij}, \mathbf{x}_{ij}) - 1] f(y_{ij} \mid \mathbf{x}_{ij}, u_i, (i,j) \in R)}{E\{[p_r^{-1}(y_{ij}, \mathbf{x}_{ij}) - 1] \mid \mathbf{x}_{ij}, u_i, (i,j) \in R\}}. \tag{2.3}$$

See Sverchkov and Pfeffermann (2018), and Pfeffermann and Sverchkov (2019) for details.

3. ESTIMATION OF RESPONSE PROBABILITIES

Unlike the sampling probabilities, the response probabilities are generally unknown. We assume therefore a parametric model, which is allowed to depend on the outcome and the covariate values; $\Pr[R_{ij}=1|\, y_{ij},\mathbf{x}_{ij}, i\in s, j\in s_i;\boldsymbol{\gamma}] = p_r(y_{ij},\mathbf{x}_{ij};\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is a vector of unknown coefficients. We assume that $p_r(y_{ij},x_{ij};\boldsymbol{\gamma})$ is differentiable with respect to $\boldsymbol{\gamma}$ and satisfies the same mild regularity conditions as in Sverchkov and Pfeffermann (2018).

Under these assumptions, if the missing outcome values were observed, $\boldsymbol{\gamma}$ could be estimated by solving the likelihood equations:

$$\sum_{(i,j)\in R}\frac{\partial \log p_r(y_{ij},\mathbf{x}_{ij};\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}+\sum_{(i,\mathrm{k})\in R^c}\frac{\partial \log[1-p_r(y_{ik},\mathbf{x}_{ik};\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}}=0\,. \qquad (3.1)$$

In practice, the missing data are unobserved and hence the likelihood equations (3.1) are not operational. However, one may apply in this case the missing information principle:

Missing Information Principle (MIP, Cepillini et al. 1955, Orchard and Woodbury, 1972)**:** Let $O=\{y_{ij},n_i,(i,j)\in R;\ \mathbf{x}_{ht},h=1,...,M,\ \mathrm{t}=1,...,N_i\}$ represent the known observed data used below. Since no observations are available for $(i,k)\in R^c$, solve instead,

$$E_U\left\{\left[\sum_{(i,j)\in R}\frac{\partial \log p_r(y_{ij},\mathbf{x}_{ij};\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}+\sum_{(i,\mathrm{k})\in R^c}\frac{\partial \log[1-p_r(y_{ik},\mathbf{x}_{ik};\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}}\right]\middle| O\right\}$$

$$\overset{by(2.3)}{=}\sum_{(i,j)\in R}\frac{\partial \log p_r(y_{ij},\mathbf{x}_{ij};\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}$$

$$+\sum_{(i,k)\in R^c}E_s\left(\frac{E_{re}\left\{[p_r^{-1}(y_{ik},\mathbf{x}_{ik};\boldsymbol{\gamma})-1]\dfrac{\partial \log[1-p_r(y_{ik},\mathbf{x}_{ik};\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}}\middle|\,\mathbf{x}_{ik},u_i,(i,k)\in R\right\}}{E_{re}\{[p_r^{-1}(y_{ik},\mathbf{x}_{ik};\boldsymbol{\gamma})-1]\,|\,\mathbf{x}_{ik},u_i,(i,k)\in R\}}\middle| O\right)=0. \tag{3.2}$$

See Sverchkov (2008) and Sverchkov and Pfeffermann (2018) for derivation of (3.2). In these equations, $E_U, \mathrm{E}_s, \mathrm{E}_{re}$ define respectively expectations with respect to the population distribution, the sample distribution and the respondents' distribution. Notice that the internal expectations in the last expression are with respect to the model holding for the observed data for the respondents.

Remark 2. When the response probabilities $p_r(y_{ij},\mathbf{x}_{ij};\boldsymbol{\gamma})$ depend on only $\mathbf{x}_{ij}$, they are referred to as *propensity scores*, and the missing data are missing at random. This kind of response mechanism may hold in establishment surveys, for example, when the response probability is related to the known size of the establishment. The estimating equations in (3.2) reduce in this case to the common log-likelihood equations,

$$\sum_{(i,j)\in R}\frac{\partial \log p_r(\mathbf{x}_{ij};\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}+\sum_{(i,k)\in R^c}\frac{\partial \log[1-p_r(\mathbf{x}_{ik};\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}}=0, \tag{3.3}$$

where $p_r(\mathbf{x}_{ij};\boldsymbol{\gamma})=\Pr(\mathrm{R}_{ij}=1\,|\,\mathbf{x}_{ij};\boldsymbol{\gamma})$.

Sverchkov and Pfeffermann (2018) propose to solve the equations (3.2) by maximizing the log-likelihood leading to them, i.e., maximizing,

$$l(\boldsymbol{\gamma}) = \sum_{(i,j)\in R} \log p_r(y_{ij}, \mathbf{x}_{ij}; \boldsymbol{\gamma})$$

$$+ \sum_{(i,k)\in R^c} E_s\left(\frac{E_{re}\{[p_r^{-1}(y_{ik}, \mathbf{x}_{ik}; \boldsymbol{\gamma}^*) - 1]\log[1 - p_r(y_{ik}, \mathbf{x}_{ik}; \boldsymbol{\gamma})] \mid \mathbf{x}_{ik}, u_i, (i,k) \in R\}}{E_{re}\{[p_r^{-1}(y_{ik}, \mathbf{x}_{ik}; \boldsymbol{\gamma}^*) - 1] \mid \mathbf{x}_{ik}, u_i, (i,k) \in R\}} \middle| O \right). \quad (3.4)$$

We distinguish between $\boldsymbol{\gamma}^*$ and $\boldsymbol{\gamma}$ because by (3.2), the derivatives should only be taken with respect to $\boldsymbol{\gamma}$.

We maximize the likelihood (3.4) by replacing u_i by $\hat{u}_i$, obtained by fitting a model of the form (2.2), and dropping the external expectation- E_s. The maximization is carried out iteratively, by maximizing in the (q+1) iteration the expression,

$$\sum_{(i,j)\in R} \log p_r(y_{ij}, \mathbf{x}_{ij}; \boldsymbol{\gamma}^{(q+1)})$$

$$+ \sum_{(i,k)\in R^c} \frac{E_{re}\{[p_r^{-1}(y_{ik}, \mathbf{x}_{ik}; \boldsymbol{\gamma}^{(q)}) - 1]\log[1 - p_r(y_{ik}, \mathbf{x}_{ik}; \boldsymbol{\gamma}^{(q+1)})] \mid \mathbf{x}_{ik}, \hat{u}_i, (i,k) \in R\}}{E_{re}\{[p_r^{-1}(y_{ik}, \mathbf{x}_{ik}; \boldsymbol{\gamma}^{(q)}) - 1] \mid \mathbf{x}_{ik}, \hat{u}_i, (i,k) \in R\}}$$

(3.5)

with respect to $\boldsymbol{\gamma}^{(q+1)}$. The maximization can be carried out, for example, by SAS Proc NLIN. See Sverchkov (2022) and the examples following Remark 3 for details.

Remark 3. A fundamental question regarding the solution of the MIP equations is the existence of a unique solution or more generally, the identifiability of the response model. Riddles et al. (2016) propose a similar approach to deal with NMAR nonresponse in the general context of survey sampling inference and establish the following fundamental condition for the response model identifiability: the covariates $\mathbf{x}$ can be decomposed as $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ with $dim(\mathbf{x}_2) \geq 1$, such that $\Pr(R_{ij} = 1 \mid y_{ij}, \mathbf{x}_{ij}) = \Pr(R_{ij} = 1 \mid y_{ij}, \mathbf{x}_{1ij})$. In other words, the covariates in $\mathbf{x}_2$ that appear in the outcome model do not affect the response probabilities, given the outcome and the other covariates. Variable(s) of this property may or may not exist in a general set up, but interesting enough, SAE models actually contain such a variable, namely, the random effects. The random effects play a fundamental role in SAE models, so the outcome clearly depends on them, but it is reasonable to

assume that the response probabilities do not depend on the random effect, given the outcome value. In practice, the random effects are unobservable, but we estimate them and then solve the equations (3.5) by conditioning on the estimated effects. So, it is actually the estimated random effects that play the role of the covariates $\mathbf{x}_2$. (Other covariates that are predictive of the outcome but not of the response might exist as well).

Clearly, the larger is the absolute values of the random effects, the more they affect the values of the outcome values and hence also the values of the response probabilities. In the simulation study of Sverchkov and Pfeffermann (2018), the authors study the effect of the magnitude of the variance of the random effects on the prediction of the area means. The conclusions from that study is that although the estimators of response model parameters become biased as the variance of the random effects increases, the biases are relatively very small and so are the standard deviations of the estimators. Increasing the variance of the random effects has negligible effect on the estimation of the true response probabilities and the predictors of the true small are means remain virtually unbiased in each of the areas.

Riddles et al. (2016) prove asymptotic normality of the estimate $\hat{\boldsymbol{\gamma}}$ under general regularity conditions.

Example 1. (Sverchkov and Pfeffermann 2018): *Mixed logistic model for the outcome variable*.

Suppose that the model fitted to the observed data of the respondents is the mixed generalized logistic model,

$$p_y(x_{ij},u_i)=\Pr(y_{ij}=1\,|\,x_{ij},u_i,(i,j)\in R;\boldsymbol{\beta})=\frac{\exp(\beta_0+\beta_1 x_{ij}+u_i)}{1+\exp(\beta_0+\beta_1 x_{ij}+u_i)},\ u_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma_u^2).$$

Consider a generic response model, $p_r(y_{ij},x_{ij};\boldsymbol{\gamma})=\Pr[R_{ij}=1\,|\,y_{ij},x_{ij},i\in s,j\in s_i;\boldsymbol{\gamma}]$.

The components of (3.2) can be written in this case as,

$$E_{re}\left\{[p_r^{-1}(y_{ij},x_{ij};\boldsymbol{\gamma})-1]\frac{\partial\log[1-p_r(y_{ij},x_{ij};\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}}\,\middle|\,x_{ij},u_i,(i,j)\in R\right\}=$$

$$p_y(x_{ij},u_i)[p_r^{-1}(1,x_{ij};\boldsymbol{\gamma})-1]\frac{\partial \log[1-p_r(1,x_{ij};\boldsymbol{\gamma})]}{\partial \gamma}+$$

$$[1-p_y(x_{ij},u_i)][p_r^{-1}(0,x_{ij};\boldsymbol{\gamma})-1]\frac{\partial \log[1-p_r(0,x_{ij};\boldsymbol{\gamma})]}{\partial \boldsymbol{\gamma}}; \quad (3.6)$$

$$E_{re}\{[p_r^{-1}(y_{ij},x_{ij};\boldsymbol{\gamma})-1]\,|\,x_{ij},u_i,(i,j)\in R\}=p_y(x_{ij},u_i)[p_r^{-1}(1,x_{ij};\boldsymbol{\gamma})-1]+$$

$$[1-p_y(x_{ij},u_i)][p_r^{-1}(0,x_{ij};\boldsymbol{\gamma})-1]. \quad (3.7)$$

The random effects u_i and the logistic probabilities $p_y(x_{ij},\mathrm{u}_i)$ can be estimated by use of the SAS procedure PROC NLMIX.

Example 2. (Sverchkov 2022): *General continuous model*.

In Example 1, the outcomes follow a discrete distribution. In this section, we consider continuous outcomes. The proposed algorithm consists of two parts:

Part 1: Fit (estimate) the model (2.2). The output of this part (input for Part 2) contains the model parameter estimates, the estimated random effects, $\hat{u}_i$, and for each $(i,j)\in R$, estimates of $p_y^{(l)}(\mathbf{x}_{ij},\hat{u}_i)=P_R(a_l\le y_{ij}<a_{l+1}\,|\,\mathbf{x}_{ij},\hat{u}_i,(i,j)\in R)$, $l=0,...,L+2$; $a_0=-\infty$, $a_{L+2}=\infty$, $a_l=\min(y_{ij})+(l-1)\frac{\max(y_{ij})-\min(y_{ij})}{L}$, $l=1,...,L+1$. The max and min are over all the observed values y_{ij}.

Part 2: Approximate the expectations in (3.2) similarly to (3.6) and (3.7):

$$E_{re}\left\{[p_r^{-1}(y_{ij},\mathbf{x}_{ij};\boldsymbol{\gamma})-1]\frac{\partial \log[1-p_r(y_{ij},\mathbf{x}_{ij};\boldsymbol{\gamma})]}{\partial \boldsymbol{\gamma}}\,\middle|\,\mathbf{x}_{ij},u_i,(i,j)\in R\right\}\cong$$

$$\sum_{l=1}^{L+1}\hat{p}_y^{(l)}(\mathbf{x}_{ij},\hat{u}_i)[p_r^{-1}(a_l,\mathbf{x}_{ij};\boldsymbol{\gamma})-1]\frac{\partial \log[1-p_r(a_l,\mathbf{x}_{ij};\boldsymbol{\gamma})]}{\partial \boldsymbol{\gamma}}, \quad (3.8)$$

$$E_{re}\left\{[p_r^{-1}(y_{ij},\mathbf{x}_{ij};\boldsymbol{\gamma})-1]\,\middle|\,\mathbf{x}_{ij},u_i,(i,j)\in R\right\}\cong\sum_{l=1}^{L+1}\hat{p}_y^{(l)}(\mathbf{x}_{ij},\hat{u}_i)[p_r^{-1}(a_l,\mathbf{x}_{ij};\boldsymbol{\gamma})-1], \quad (3.9)$$

where $p_r(a_l,\mathbf{x}_{ij};\boldsymbol{\gamma})=\Pr[R_{ij}=1\,|\,y_{ij}=a_l,\mathbf{x}_{ij},i\in s,j\in s_i;\boldsymbol{\gamma}]$.

Substitute (3.8) and (3.9) into (3.2) and estimate $\boldsymbol{\gamma}$ by iteratively maximizing (3.5).

4. SELECTION OF A RESPONSE MODEL

There is no direct way to test the appropriateness of a chosen response model because the outcome values, which are part of the model, are unknown for the nonresponding units. If the model for the outcomes before nonresponse was known, one could derive the model holding for the observed outcomes based on this model and the model assumed for the responding units, and test the resulting model by use of standard tests that compare the cumulative hypothesized distribution of the observed data with the corresponding empirical distribution, and/or by testing moments of the assumed model. See, e.g., Pfeffermann and Landsman (2011) and Pfeffermann and Sikov (2011). However, in the approach described in Section 3, we start with a model fitted to the observed outcomes, which does not include the response model and therefore, we cannot use a similar strategy.

When following the approach proposed in Section 3, the likelihood (3.4) suggests at least two procedures for the selection of the response model in SAE under NMAR nonresponse. **1-** Compare different models based on information criteria such as the Akaike information criterion, $\text{AIC}= -2l(\boldsymbol{\gamma}) + 2\dim(\boldsymbol{\gamma})$, or Schwarz information criterion, $\text{BIC}= -2l(\boldsymbol{\gamma}) + \dim(\boldsymbol{\gamma})\log(n)$, $n = \sum_{i \in s} n_i$; **2-** test a saturated versus a nested model based on the likelihood ratio test. In Section 5 we illustrate via a simulation study how the likelihood (3.4) can be used for the application of these selection procedures.

5. SIMULATION STUDY

5.1 *Simulation set-up*

We start by defining the sample model before nonresponse because as stated in Section 4, our approach for estimating the response model is based on fitting a model to the observed outcomes, which does not include the response model. For convenience, we assume noninformative sampling of areas and within the areas, such that the sample model before nonresponse is the same as the population model. Note that although the sampling design defines the observed model (2.2),

once this model is estimated, the sampling design does not affect the estimation of the response probabilities in section 3.

The simulation study consists of the following steps:

Generate auxiliary values x_{ij}, $i=1,...,100$, $j=1,...,20$ from a Uniform(0,2) distribution. Next, generate sample values from the small area model,

$$y_{ij} \mid x_{ij}, u_i \sim N(x_{ij}+u_i, 1),\ i=1,...,100,\ j=1,...,20;\ u_i \sim N(0,1). \qquad (5.1)$$

Consider three unit response models (no selection of areas):

$p_r^{(1)}(y_{ij}, x_{ij}) = \text{logit}(-x_{ij}/2 + 2y_{ij})$,

$p_r^{(2)}(y_{ij}, x_{ij}) = \text{logit}(-x_{ij}/2 + 2y_{ij} - 0.3y_{ij}^2)$,

$p_r^{(3)}(y_{ij}, x_{ij}) = \text{logit}(1.5x_{ij})$.

Select 3 sets of respondents:

R1 uses Poisson sampling, independently between the units with response probabilities $p_r^{(1)}(y_{ij}, x_{ij})$,

R2 is the same as R1 but with response probabilities $p_r^{(2)}(y_{ij}, x_{ij})$,

R3 is the same as R1 but with response probabilities $p_r^{(3)}(y_{ij}, x_{ij})$.

The 3 response probabilities yield similar response rates of 65 - 75 per cent.

The working model for the observed data for the responding units is,

$$y_{ij} \mid x_{ij}, u_i \sim N(\theta_0 + \theta_1 x_{ij} + u_i, \theta_2),\ i=1...100,\ j=1...20;\ u_i \sim N(0, \sigma_u^2). \qquad (5.2)$$

Remark 4. The working model (5.2) is correct for the observed sample R3 that corresponds to MAR nonresponse, but not for R1 and R2, under which the nonresponse is NMAR.

Define three working response models:

M1: $p_r(y_{ij}, x_{ij}; \boldsymbol{\gamma}^1) = \text{logit}(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij})$,

M2: $p_r(y_{ij}, x_{ij}; \boldsymbol{\gamma}^2) = \text{logit}(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij} + \gamma_3 y_{ij}^2)$,

M3: $p_r(y_{ij}, x_{ij}; \boldsymbol{\gamma}^3) = \text{logit}(\gamma_0 + \gamma_1 x_{ij})$,

Note that $p_r(y_{ij}, x_{ij}; \boldsymbol{\gamma}^3)$ is nested in $p_r(y_{ij}, x_{ij}; \boldsymbol{\gamma}^1)$ and $p_r(y_{ij}, x_{ij}; \boldsymbol{\gamma}^2)$, and $p_r(y_{ij}, x_{ij}; \boldsymbol{\gamma}^1)$ is nested in $p_r(y_{ij}, x_{ij}; \boldsymbol{\gamma}^2)$. The response probability $p_r(y_{ij}, x_{ij}; \boldsymbol{\gamma}^3)$ defines MAR nonresponse and hence, can be estimated by solving (3.3).

Estimate the unknown parameters $\theta_0, \theta_1, \theta_2$ in (5.2) by SAS Proc NMIX, and then estimate $\boldsymbol{\gamma}$ by maximizing (3.4), as described in Section 3. The maximization was carried out by use of SAS Proc NLIN under the following 9 scenarios, as defined by the true response model and the assumed working response model:

S1: R1 set of respondents, M1 working response model.
S2: R1 set of respondents, M2 working response model.
S3: R1 set of respondents, M3 working response model.

S4: R2 set of respondents, M1 working response model.
S5: R2 set of respondents, M2 working response model.
S6: R2 set of respondents, M3 working response model.

S7: R3 set of respondents, M1 working response model.
S8: R3 set of respondents, M2 working response model.
S9: R3 set of respondents, M3 working response model.

Select the response model based on:

1- The Likelihood Ratio Test (LRT); test a saturated model [$l(\boldsymbol{\gamma}^{**})$] against a nested model [$l(\boldsymbol{\gamma}^{*})$], assuming the χ^2 distribution under the null hypothesis H_0 that the nested model with a smaller number of parameters is correct. The test statistic is $\lambda_{LRT} = -2[l(\hat{\boldsymbol{\gamma}}^{*}) - l(\hat{\boldsymbol{\gamma}}^{**})] \sim \chi^2_{[\dim(\boldsymbol{\gamma}^{**}) - \dim(\boldsymbol{\gamma}^{*})]}$. Reject H_0 at the $\alpha = .05$ level.

2 - AIC selection criterion: compare the values of the AIC as obtained for the corresponding two models;

3 - BIC selection criterion: compare the values of the BIC as obtained for the corresponding two models.

Repeat the whole process independently 500 times.

5.2 *Results*

S1 Vs S2 (R1 – set of respondents, M1 – correct model, M2 – saturated model). Note that although M2 is a saturated model, it is also correct but with an additional term. The LRT selects the model M1 in 368 out of the 500 simulations. AIC selects M1 in 305 out of 500 simulations, BIC selects M1 in 324 simulations.

S1 Vs S3 (R1 – set of respondents, M1 – correct model, M3 – incorrect nested model). The LRT selects the correct model M1 in 500 out of the 500 simulations. AIC and BIC likewise select M1 in all the 500 simulations.

S4 Vs S5 (R2 – set of respondents, M2 – correct model, M1 – incorrect nested model). The LRT selects the correct model M2 in 433 out of the 500 simulations. AIC and BIC select M2 in 483 simulations.

S4 Vs S6 (R2 - set of respondents, M2 – correct model, M3 – incorrect nested model). The LRT selects the correct model M2 in 490 out of the 500 simulations. AIC and BIC select the correct model in all the simulations.

S7 Vs S8 (R3 - set of respondents, M3 – correct model, M1 – also correct but a saturated model). The LRT selects the model M3 in 241 out of 500 simulations. AIC selects M3 in 225 out of 500 simulations; BIC selects M3 in 420 simulations.

S7 Vs S9 (R3 - set of respondents, M3 – correct model, M2 – also correct but a saturated model). The LRT selects the M3 model in 241 out of 500 simulations. AIC selects M3 in 361 out of 500 simulations, BIC selects M3 in 450 simulations.

Note that when R3 is the set of respondents and M3 is the correct model, M1 and M2 also produce correct estimates of the response probabilities, although with additional estimated parameters. Thus, the fact that the LRT test and the AIC select the M3 model in about half of the simulations is not surprising. The use of the BIC criterion performs better in these cases.

The results so far are summarized in table 1.

Table 1. Percentages out of 500 simulations in which each of the three selection procedures selected the correct model, for different combinations of correct (rows) and working (columns) response probability models.

	M1			M2			M3		
	LRT	AIC	BIC	LRT	AIC	BIC	LRT	AIC	BIC
R1, M1 correct	---	---	---	73.6	61.0	64.8	100	100	100
R2, M2 correct	86.6	96.6	96.6	---	---	---	98	100	100
R3, M3 correct	48.2	45	82	48.2	72.2	90	---	---	---

Finally, we consider the case where a working model is incorrect but might be a good approximation of the correct model: let R1 be the set of respondents such that M1 is the correct working model. Let M4 be the following working model: $p_r(y_{ij}, x_{ij}; \boldsymbol{\gamma}^4) = \text{logit}(\gamma_0 + \gamma_1 x_{ij} + \gamma_2 y_{ij}^2 + \gamma_3 y_{ij}^3)$. Compare M4 with M1 (correct model). In this case, AIC selects the correct M1 model in 430 out of the 500 simulations and BIC selects the correct model in 431 simulations.

Sverchkov (2013) suggested testing whether the response is NMAR or MAR by testing the significance of the corresponding estimated coefficients in the saturated response model. We applied this idea by testing the significance of the estimated coefficients $\hat{\gamma}_2$ and $\hat{\gamma}_3$ under the response models $p_r^{(1)}(y_{ij}, x_{ij})$ and $p_r^{(2)}(y_{ij}, x_{ij})$, (both assume NMAR nonresponse), when in fact the true response model is $p_r^{(3)}(y_{ij}, x_{ij})$ (MAR) or $p_r^{(1)}(y_{ij}, x_{ij})$ (NMAR), using the standard t-tests. (SAS Proc NLIN provides standard errors of the estimated coefficients.)

We considered two samples of respondents, R3 and R1. For R3, we found that when testing the working response model $p_r^{(1)}(y_{ij}, x_{ij})$, in 432 out of the 500 simulations, $\hat{\gamma}_2$ was not significant at the 0.05 level. When testing the working response model $p_r^{(2)}(y_{ij}, x_{ij})$, in 350 out of the 500 simulations, $\hat{\gamma}_2$ was not

significant at the 0.05 level, and in 398 simulations $\hat{\gamma}_3$ was not significant. Recall that for the respondents' sample R3, the working SAE model (5.2) for the observed outcomes is correct since the response is MAR.

For the respondents' sample R1, the response model $p_r^{(1)}(y_{ij}, x_{ij})$ is correct and in all the 500 simulations, the estimator $\hat{\gamma}_2$ was found significant. However, when testing the response model $p_r^{(2)}(y_{ij}, x_{ij})$, in 388 simulations the estimator $\hat{\gamma}_3$ was significant, even at the 0.01 level, and $\hat{\gamma}_2$ was significant in 498 simulations. This result might be explained by the fact that the working outcome model (5.2) is not correct when the response model is NMAR and thus, the likelihood (3.5), which conditions on the estimated random effects for the estimation of the γ coefficients is incorrect.

6. SUMMARY

In this paper we investigate the use of the likelihood function of the observed respondents' data for selecting an appropriate response model under possible NMAR nonresponse. For estimating the hypothesized model, we applied the missing information principle. Despite of what seems to be a rather complex estimation process, we find in our simulation study that the AIC and BIC information criteria and the LRT test, when applicable, perform well for model selection. Clearly, the use of other likelihood-based tests and selection criteria should be investigated as well.

REFERENCES

Cepillini, R., Siniscialco, M., and Smith, C.A.B. (1955). The estimation of gene frequencies in a random mating population. *Annals of Human Genetics*, **20**, 97-115.

Orchard, T., and Woodbury, M.A. (1972). A missing information principle: theory and application. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 697-715.

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science*, **28**, 40-68.

Pfeffermann, D., and V. Landsman (2011), "Are Private Schools Better than Public Schools? Appraisal for Ireland by Methods for Observational Studies," Annals of Applied Statistics, 5, 1726–1751.

Pfeffermann, D. and Sikov N. (2011). Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information. *Journal of Official Statistics*, **27**, 181–209.

Pfeffermann, D., and Sverchkov, M. (2007). Small-Area Estimation under Informative Probability Sampling of Areas and Within Selected Areas. *Journal of the American Statistical Association*, **102**, 1427-1439.

Pfeffermann, D., and Sverchkov, M. (2009). Inference under Informative Sampling. In: Handbook of Statistics 29B; *Sample Surveys: Inference and Analysis*. Eds. D. Pfeffermann and C.R. Rao. North Holland. pp. 455-487.

Pfeffermann, D. and Sverchkov, M. (2019). Multivariate small area estimation under nonignorable nonresponse, *Statistical Theory and Related Fields,* **3**, pp. 213-223.

Rao, J.N.K., and Molina, I. (2015), *Small Area Estimation*, 2nd Edition, Wiley.

Riddles, K.M., Kim, J.K. and Im, J. (2016). A propensity-score adjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology*, **4**, 215-245.

Särndal, C.E. and Swensson B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, **55**, 279-294.

Sverchkov, M. (2008). A new approach to estimation of response probabilities when missing data are not missing at random. *Joint Statistical Meetings*, *Proceedings of the Section on Survey Research Methods*, 867-874.

Sverchkov, M. (2013). Is it MAR or NMAR? *2013 JSM Meetings, Proceedings of the Section on Survey Methods Research,* pp. 2307-2311

Sverchkov, M. (2022). An Algorithm for Small Area Estimation under Not Missing At Random Non-response. *Joint Statistical Meetings*, *Proceedings of the Section on Survey Research Methods*, pp. 1735-1745.

Sverchkov, M., and Pfeffermann, D. (2004). Prediction of Finite Population Totals Based on the Sample Distribution. *Survey Methodology*, **30**, 79-92.

Sverchkov, M. and Pfeffermann, D. (2018). Small area estimation under informative sampling and not missing at random non-response. *Journal of Royal Statistical Society, ser. A,* 181, *Part* 4, *pp.* 981–1008.