# RESPONSE MODEL SELECTION IN SMALL AREA ESTIMATION UNDER NOT MISSING AT RANDOM NONRESPONSE December 2024

Michael Sverchkov
Bureau of Labor Statistics, Washington DC, USA

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

## 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 that 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.

## 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 also $\sum_{j=1}^{n_i} R_{ij} > 0$ 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:

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

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 $\pi_i$ 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);

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

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; $\Pr[R_{ij} = 1 \mid y_{ij}, \mathbf{x}_{ij}, i \in s, j \in s_i; \boldsymbol{\gamma}]$ $= p_r(y_{ij}, \mathbf{x}_{ij}; \boldsymbol{\gamma})$, where $\gamma$ 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, as it was shown in Sverchkov (2008) and Sverchkov and Pfeffermann (2018), the parameter $\gamma$ can be estimated by maximizing the log-likelihood (assuming $\boldsymbol{\gamma}^*$ in (3.1) be the "true" value of $\gamma$, see (3.2) for clarification),

$$l(\boldsymbol{\gamma}) = \sum_{(i,j)\in R} \log p_r(y_{ij}, \mathbf{x}_{ij}; \boldsymbol{\gamma})$$
$$+ \sum_{(i,k)\in R^c} E\left( \frac{E\{[p_r^{-1}(y_{ik}, \mathbf{x}_{ik}; \boldsymbol{\gamma}^*) - 1] \log[1 - p_r(y_{ik}, \mathbf{x}_{ik}; \boldsymbol{\gamma})] \mid \mathbf{x}_{ik}, u_i, (i,k) \in R\}}{E\{[p_r^{-1}(y_{ik}, \mathbf{x}_{ik}; \boldsymbol{\gamma}^*) - 1] \mid \mathbf{x}_{ik}, u_i, (i,k) \in R\}} \middle| O \right). \quad (3.1)$$

We maximize the likelihood (3.1) by replacing $u_i$ by $\hat{u}_i$, obtained by fitting a model of the form (2.2), and dropping the external expectation. The maximization is carried out iteratively by maximizing in the (q+1) iteration the expression,

$$\sum_{(i,j)\in R} \log p_r(y_{ij}, \mathbf{x}_{ij}; \boldsymbol{\gamma}^{(q+1)})$$
$$+ \sum_{(i,k)\in R^c} \frac{E\{[p_r^{-1}(y_{ik}, \mathbf{x}_{ik}; \boldsymbol{\gamma}^{(q)}) - 1] \log[1 - p_r(y_{ik}, \mathbf{x}_{ik}; \boldsymbol{\gamma}^{(q+1)})] \mid \mathbf{x}_{ik}, \hat{u}_i, (i,k) \in R\}}{E\{[p_r^{-1}(y_{ik}, \mathbf{x}_{ik}; \boldsymbol{\gamma}^{(q)}) - 1] \mid \mathbf{x}_{ik}, \hat{u}_i, (i,k) \in R\}} \qquad (3.2)$$

with respect to $\boldsymbol{\gamma}^{(q+1)}$. The maximization can be carried out, for example, by SAS Proc NLIN. See Sverchkov (2022) for details. Riddles et al. (2016) derive sufficient conditions under which the above maximization procedure leads to unique solution.

## 4. SELECTION OF A RESPONSE MODEL

There is no direct way to test the appropriateness of a chosen response model since 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 obtain the distribution of the observed outcomes based on this distribution and the response model, and test the resulting model fitted to the responding units by using standard tests that compare the cumulative hypothesized distribution with the corresponding empirical distribution, and/or by testing moments of the hypothesized model. See e.g., Pfeffermann and Landsman (2011) and Pfeffermann and Sikov (2011). However, in the approach described in Section 3, we start with the distribution for the observed outcomes, which does not include the response model and therefore, we cannot use the same strategy.

When following the approach proposed in Section 3, the likelihood (3.1) 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, $\mathrm{AIC}=-2l(\boldsymbol{\gamma})+2\dim(\boldsymbol{\gamma})$, or Schwarz information criterion, $\mathrm{BIC}=-2l(\boldsymbol{\gamma})+\dim(\boldsymbol{\gamma})\log(n),\ n=\sum_{i\in s}n_i$ ; **2-** test a saturated versus a nested model based on likelihood ratio test.

## REFERENCES

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.

Sarndal, 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. (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.