MULTIVARIATE SMALL AREA ESTIMATION UNDER INFORMATIVE SAMPLING AND NONRESPONSE

Michael Sverchkov[1] **and Danny Pfeffermann**[2]

[1] *Bureau of Labor Statistics*, Washington DC, USA

[2] *National Statistician of Israel*; *Professor*: *Hebrew University of Jerusalem, Israel* & *University of Southampton, UK*. (**)

Key words: Distribution of missing data, missing information principle, not missing at random

(**) *The opinions expressed in this paper are of the authors and do not necessarily represent the policies of the U.S. Bureau of Labor Statistics and the Israel Central Bureau of Statistics.*

INTRODUCTION

We consider multivariate small area estimation under informative sampling and not missing at random (NMAR) nonresponse.

We define a response model that accounts for the different patterns of the observed outcomes (which values are observed and which ones are missing), and estimate the response probabilities by application of the Missing Information Principle. By this principle, we first define the likelihood score equations as if the missing outcomes were actually observed, and then integrate out the unobserved outcomes from the score equations with respect to the distribution holding for the missing data. The latter distribution is obtained from the distribution fitted to the observed data.

Finally, the integrated score equations are solved with respect to the unknown parameters underlying the response model. See Sverchkov (2008), Sverchkov and Pfeffermann (2018) and Riddles et al. (2016) for application of this approach in the univariate case.

Once the response probabilities are estimated, we impute the missing outcomes and then apply the approach of Pfeffermann and Sverchkov (2007) to the complete data set (observed and imputed values), to obtain the small area predictors.

1. Notation and Models

$\{\mathbf{y}_{ij}, \mathbf{x}_{ij}; i = 1,..., M,\ j = 1,..., N_i\}$ - the data in a finite population of N units belonging to M areas with N_i units in area i, $\sum_{i=1}^{M} N_i = N$,

$\mathbf{y}_{ij} = (y_{ij,1},..., y_{ij,K})'$ - the value of the vector of outcome values for unit j in area i

$\mathbf{x}_{ij} = (x_{ij,1},..., x_{ij,L})'$ - a vector of corresponding L covariates. We assume that the covariates are known for every unit in the population.

The outcome values follow the generic two-level population model:

$$\begin{aligned} &\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i \overset{ind}{\sim} f(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i),\ i = 1...M,\ j = 1...N_i, \\ &\mathbf{u}_i \overset{ind}{\sim} f(\mathbf{u}_i);\ E(\mathbf{u}_i) = \mathbf{0} = (0,...,0)',\ V(\mathbf{u}_i) = \Sigma, \end{aligned} \tag{1}$$

where $\mathbf{u}_i = (u_{i,1},..., u_{i,K})'$ is an unobserved random effect.

The target is to estimate $f(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i)$ and $f(\mathbf{u}_i)$based on the observed data, as obtained under incomplete response.

By incomplete response we mean that some or all of the outcomes $y_{ij,k}$ are unobserved for some of the sampled units.

Once the two distributions are estimated, we estimate the expectations of target outcomes in the different areas and use them for predicting the area means, the common targets in small area estimation.

Alternatively, one can use the estimated distributions for imputation of the missing data.

Although we do not consider any selection from finite population in the present paper, all results can be generalized to the case where first a sample was selected from the finite population by some (informative) sampling scheme and then nonresponse occurs, see Pfeffermann and Sverchkov (2007) and Sverchkov and Pfeffermann (2018).

Define the response indicator $R_{ij,k} = 1(0)$ if $y_{ij,k}$ is observed (unobserved),

let $\mathbf{R}_{ij} = (R_{ij,1},...,R_{ij,K})'$,

$\mathbf{r}$ be any K-dimension vector with 0, 1 components,

and assume a parametric model for the response probabilities that depends on the outcome, the random effects and the covariates, indexed by the vector parameter γ; $p_{\mathbf{r}}(\mathbf{y}_{ij},\mathbf{x}_{ij},\mathbf{u}_i;\gamma) = \Pr[\mathbf{R}_{ij} = \mathbf{r} \mid \mathbf{y}_{ij},\mathbf{x}_{ij},\mathbf{u}_i;\gamma]$, with $p_{\mathbf{r}}(\mathbf{y}_{ij},\mathbf{x}_{ij},\mathbf{u}_i;\gamma)$ differentiable with respect to γ.

Assumption 1.

(a) The response occurs independently between the units,

(b) $\Pr[\mathbf{R}_{ij} = \mathbf{r} \mid (\mathbf{y}_{i^*j^*},\mathbf{x}_{i^*j^*},,\mathbf{u}_{i^*}), i^* = 1...M,\ j^* = 1...N_i] = \Pr[\mathbf{R}_{ij} = \mathbf{r} \mid \mathbf{y}_{ij},\mathbf{x}_{ij},\mathbf{u}_i]$,

and (c) there is no area non-response.

Note that under (1) and Assumption 1:

$$f(\mathbf{y}_{ij} \mid \mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij},\{(\mathbf{y}_{i^*j^*},\mathbf{x}_{i^*j^*},\mathbf{R}_{i^*j^*},\mathbf{u}_{i^*}), i^* = 1...M,\ j^* = 1...N_i; (i^*,j^*) \neq (i,j)\}) = f(\mathbf{y}_{ij} \mid \mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij}) \quad (2)$$

We assume a parametric form for the "completely observed" outcomes,

$$\begin{aligned}&\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{1} = (1,...,1)' \sim f_R(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i; \theta_1) = f(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{1}; \theta_1); \\ &\mathbf{u}_i \sim f(\mathbf{u}_i; \theta_2), E(\mathbf{u}_i; \theta_2) = 0,\end{aligned} \tag{3}$$

where we assume that both functions in (3) are differentiable with respect to the vector parameters θ_1 and θ_2 respectively.

We do not require that the subset $\{(i, j) : \mathbf{R}_{ij} = \mathbf{1}\}$ of the observed data is not empty. Even if it is, the model (3) is properly defined by (1) and Assumption 1.

On the other hand (3) and assumption 1 define (1), therefore by estimating the parameters γ and $\theta = (\theta_1, \theta_2)$ one can estimate (1).

2. Estimation of γ and θ

If the missing outcome values and random effects were actually observed, γ could be estimated by solving the likelihood equations:

$$\sum_{\mathbf{r}=(0,\ldots,0)'}^{(1,\ldots,1)'} \sum_{(i,j):\mathbf{R}_{ij}=\mathbf{r}} \frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij},\mathbf{x}_{ij},\mathbf{u}_i;\gamma)}{\partial \gamma} = 0, \qquad (4)$$

where the first summation is over all the K-dimension vectors with 0, 1 components.

In practice, the missing data and the random effects are unobserved and hence the likelihood equations (4) are not operational. However, one may apply in this case the missing information principle:

Missing Information Principle (Cepillini et al. 1955, Orchard and Woodbury, 1972**):**

Denote the observed data by $O = \{(y_{ij.k} : R_{ij,k} = 1), \mathbf{x}_{ij}, i = 1,\ldots,M,\ j = 1,\ldots,N_i\}$.

Since no observations are available for $(i,j,k): R_{ij,k} = 0$, solve instead the best predictor of (3) given the observed data,

$$\sum_{\mathbf{r}=(0,\ldots,0)'}^{(1,\ldots,1)'} \sum_{(i,j):\mathbf{R}_{ij}=\mathbf{r}} E\left(\frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij},\mathbf{x}_{ij},\mathbf{u}_i;\gamma)}{\partial \gamma} \middle| O, \mathbf{R}_{ij} = \mathbf{r} \right) = 0, \qquad (5)$$

$$E\left(\frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij},\mathbf{x}_{ij},\mathbf{u}_i;\gamma)}{\partial\gamma}\middle| O,\mathbf{R}_{ij}=\mathbf{r}\right)=E\left[E\left(\frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij},\mathbf{x}_{ij},\mathbf{u}_i;\gamma)}{\partial\gamma}\middle| O,\mathbf{u}_i,\mathbf{R}_{ij}=\mathbf{r}\right)\middle| O,\mathbf{R}_{ij}=\mathbf{r}\right]$$

$$=\int E\left(\frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij},\mathbf{x}_{ij},\mathbf{u}_i;\gamma)}{\partial\gamma}\middle| O,\mathbf{u}_i,\mathbf{R}_{ij}=\mathbf{r}\right) f(\mathbf{u}_i;\theta_2)d\mathbf{u}_i,$$

and $E\left(\frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij},\mathbf{x}_{ij},\mathbf{u}_i;\gamma)}{\partial\gamma}\middle| O,\mathbf{u}_i,\mathbf{R}_{ij}=\mathbf{r}\right)$ can be obtained as follows:

let α be the set of indexes for which components of $\mathbf{r}$ are equal to 1 and β is complement to α, i.e. $\mathbf{y}_{ij,\alpha}=(y_{ij,k}:r_k=1)$ and $\mathbf{y}_{ij,\beta}=(y_{ij,k}:r_k=0)$,

$$E\left(\frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij},\mathbf{x}_{ij},\mathbf{u}_i;\gamma)}{\partial\gamma}\middle| O,\mathbf{u}_i,\mathbf{R}_{ij}=\mathbf{r}\right)=$$

$$\int\frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij},\mathbf{x}_{ij},\mathbf{u}_i;\gamma)}{\partial\gamma} f(\mathbf{y}_{ij,\beta}\,|\,\mathbf{y}_{ij,\alpha},\mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij}=\mathbf{r})d\mathbf{y}_{ij,\beta}=$$

$$\int\frac{\partial \log p_{\mathbf{r}}(\mathbf{y}_{ij},\mathbf{x}_{ij},\mathbf{u}_i;\gamma)}{\partial\gamma}\frac{\{[\Pr(\mathbf{R}_{ij,\beta}=\mathbf{1}_\beta\,|\,\mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij,\alpha}=\mathbf{1}_\alpha,\mathbf{y}_{ij})]^{-1}-1\}f(\mathbf{y}_{ij,\beta}\,|\,\mathbf{y}_{ij,\alpha},\mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij}=\mathbf{1})d\mathbf{y}_{ij,\beta}}{\int[\Pr(\mathbf{R}_{ij,\beta}=\mathbf{1}_\beta\,|\,\mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij,\alpha}=\mathbf{1}_\alpha,\mathbf{y}_{ij})]^{-1}f(\mathbf{y}_{ij,\beta}\,|\,\mathbf{y}_{ij,\alpha},\mathbf{x}_{ij},\mathbf{u}_i,\mathbf{R}_{ij}=\mathbf{1})d\mathbf{y}_{ij,\beta}-1}$$

and $\Pr(\mathbf{R}_{ij,\beta} = \mathbf{1}_\beta \mid \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij,\alpha} = \mathbf{1}_\alpha, \mathbf{y}_{ij}) = \dfrac{p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}, \mathbf{u}_i; \gamma)}{\int p_{\mathbf{r}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}, \mathbf{u}_i; \gamma) f(\mathbf{y}_{ij,\beta} \mid \mathbf{y}_{ij,\alpha}, \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{1}) d\mathbf{y}_{ij,\beta}}$.

Finally one can solve (5) substituting $f(\mathbf{y}_{ij,\beta} \mid \mathbf{y}_{ij,\alpha}, \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{1}; \theta_1) = \dfrac{f_R(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i; \theta_1)}{\int f_R(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i; \theta_1) d\mathbf{y}_{ij,\beta}}$ in place of $f(\mathbf{y}_{ij,\beta} \mid \mathbf{y}_{ij,\alpha}, \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{R}_{ij} = \mathbf{1})$.

3. 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., 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.

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.

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., and Pfeffermann, D. (2018). Small area estimation under informative sampling and not missing at random nonresponse, *Journal of the Royal Statistical Society* JRSS-SA (Accepted)

Thank you.

Sverchkov.Michael@bls.gov