WHEN NONRESPONSE MAKES ESTIMATES FROM A CENSUS A SMALL AREA ESTIMATION PROBLEM: THE CASE OF THE SURVEY ON GRADUATES' EMPLOYMENT STATUS IN ITALY*

Maria Giovanna Ranalli [1], Fulvia Pennoni[2], Francesco Bartolucci[3], and Antonietta Mira[4]

[1] Department of Political Science, University of Perugia, IT (e-mail: maria.ranalli@unipg.it)

[2] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, IT (e-mail: fulvia.pennoni@unimib.it)

[3] Department of Economics, University of Perugia, IT (e-mail: francesco.bartolucci@unipg.it)

[4] Università della Svizzera italiana, CH, and Department of Economics, University of Insubria, IT (e-mail: antonietta.mira@uninsubria.it)

ABSTRACT: In this paper we frame the problem of obtaining estimates from the survey on the employment status of graduates in Italy as a Small Area Estimation problem because of unit nonresponse. We propose to use generalized linear mixed models and to include two variables that can be considered proxies of the response propensity among the set of covariates to make the MAR assumption more tenable. Estimates for degree programmes are obtained as (semi-parametric) empirical best predictions.

KEYWORDS: generalized linear mixed model, latent trait models, mixed-mode survey, nonparametric maximum likelihood, paradata.

1 Introduction

Since 1998 AlmaLaurea, a consortium of 80 Italian Universities, carries out an annual survey on the employment status of graduates. The survey is carried out one, three, and five years after graduation and provides a broad picture of graduates' job placement in the labour market. The 2022 edition has involved 660,000 first- and second-level graduates in 2020 (AlmaLaurea, 2022). The survey is a census and targets many variables of interest other than the

*We are grateful to AlmaLaurea for making the data available and to AlmaLaurea researchers for sharing their precious insights that motivated the research questions and helped with interpretation.

employment status, such as job characteristics, including type of contract and salary, and of the use of the skills gained at university.

As with all surveys, nonresponse occurs: the overall response rate for the graduates involved one year after graduation (the focus here) is 68.4%. This is the outcome of a two-fold process. First, a subset of graduates (approximately 92%) is identified as those who have given consent to be contacted according to the General Data Protection Regulation no. 2016/679. Then, these graduates are contacted using a dual survey technique: CAWI (Computer-Assisted Web Interviewing) and CATI (Computer-Assisted Telephone Interviewing). CATI is used to contact those who did not respond to the online questionnaire. This sequential mixed-mode CAWI-CATI methodology leads to a response rate of 74.2% among graduates contacted with their consent in accordance with the GDPR. Estimates for the overall population of graduates are adjusted for nonresponse by means of calibration on known population totals coming from administrative registers (AlmaLaurea, 2022; Kott, 2006).

The survey aims at providing estimates not only at the population level, but also for subpopulations (domains) of interest given by the degree programmes. In the last edition, there are almost 5,700 degree programmes for which unweighted count data are publicly released (AlmaLaurea, 2023). Some of these domains have a very small number of observations: this is due to a small number of observations in the population coupled with nonresponse. This setting resembles that of Small Area Estimation (SAE, Rao & Molina, 2015): a SAE problem arises when the sample size available in a domain (area) of interest is so small that direct estimates, albeit (approximately) unbiased, have unduly large variances. Here, re-weighting methods such as calibration are of little use. SAE methods, on the other hand, are indirect as they make use of observations coming from other areas and are model-based. In SAE, the small sample size is the outcome of a process (the sampling design) that is known to the researcher. Here, the SAE problem arises from a process (the response) that is unknown. Often, the (unverifiable) assumption that data is Missing At Random (MAR) given the covariates included in the model is made. In this paper we propose a modeling approach that tries to go beyond the classical MAR assumption by making use of all the available auxiliary information on the response behaviour of graduates from paradata and other survey data.

2 The proposed modeling approach

We adapt here the framework proposed in Marino *et al.* , 2019, and use their notation. Let U denote the finite population of AlmaLaurea graduates in 2020

of size N, which can be partitioned into m non-overlapping small areas (degree programmes), with U_i denoting the i-th small area with size N_i, $i = 1,\dots,m$. For a given degree programme i, population data consist of N_i measurements of a response variable Y_{ij} and a vector of covariates $\boldsymbol{x}_{ij} = (x_{ij1},\dots,x_{ijp})'$, with $j = 1,\dots,N_i$. For ease of notation, we consider here the case of one variable of interest. Covariates $\boldsymbol{x}$ come from administrative registers, as well as from previous surveys conducted by AlmaLaurea such as that on the Profile of Graduates. Also, let $\boldsymbol{\alpha}_1,\dots,\boldsymbol{\alpha}_m$ be iid, q-dimensional, vectors of area-specific random effects ($q \leq p$) with density $f_\alpha(\cdot)$, $E_\alpha(\boldsymbol{\alpha}_i) = 0$, and $E_\alpha(\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i') = \boldsymbol{\Sigma}$ for all $i = 1,\dots,m$. Last, let $\boldsymbol{w}_{ij}$ denote a q-dimensional subset of $\boldsymbol{x}_{ij}$ associated to $\boldsymbol{\alpha}_i$. Then a sample of size n of respondents is obtained from the above population and we denote by r_i the set containing the n_i population indexes of sample units belonging to degree programme i, with $n = \sum_{i=1}^{m} n_i$. Therefore, values of Y_{ij} are known only for the sample ($i = 1,\dots,m,\ j \in r_i$), while the values of $\boldsymbol{x}_{ij}$ and of $\boldsymbol{w}_{ij}$, are known for all units in the population ($i = 1,\dots,m, j = 1,\dots,N_i$).

Usually, it is assumed that the response process is non-informative for the small area distribution of $Y_{ij} \mid \boldsymbol{x}_{ij}$, allowing to use population level models with sample data. In order to make this assumption more tenable, we propose to include in each vector $\boldsymbol{x}_{ij}$ two covariates obtained as follows. The first one comes from paradata and has the following categories: "Response with CAWI", "Response with CATI", "Response with CATI recall", "Nonresponse", "No consent to GDPR". It can be considered as a proxy of the response propensity as these categories can be ordered along a decreasing response propensity. The second one exploits information on item nonresponse of graduates in the survey and in previous surveys to build a latent variable in the spirit of Matei & Ranalli, 2015. A set of binary indicators taking value 1 if the item is not missing and 0 if it is missing can be used to derive a latent trait using Item Response Theory models that can be interpreted as a proxy of the response propensity. Nonrespondents have all zeros and the smallest value of the latent trait.

We assume that, conditional on $\boldsymbol{\alpha}_i$, responses Y_{ij} from the same area i are independent with density $f_{y|\alpha}(y_{ij} \mid \boldsymbol{\alpha}_i;\boldsymbol{x}_{ij})$ in the Exponential Family with canonical parameter θ_{ij} modeled as $\theta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{w}_{ij}'\boldsymbol{\alpha}_i$. The marginal distribution of $\boldsymbol{y}_i$ is obtained as $f_y(\boldsymbol{y}_i;\boldsymbol{X}_i) = \int_{\mathbb{R}^q} f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i;\boldsymbol{X}_i) f_\alpha(\boldsymbol{\alpha}_i) d\boldsymbol{\alpha}_i$, where $f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i;\boldsymbol{X}_i) = \prod_{j\in r_i} f_{y|\alpha}(y_{ij} \mid \boldsymbol{\alpha}_i;\boldsymbol{x}_{ij})$ and $\boldsymbol{X}_i$ is the matrix of covariates for units in the i-th area. Typically, a parametric specification for $f_\alpha(\boldsymbol{\alpha}_i)$ is adopted, with a common choice being the $N_q(\boldsymbol{0},\boldsymbol{\Sigma})$ distribution. We also consider the more flexible alternative proposed in Marino *et al.* , 2019, in which the distribution of $\boldsymbol{\alpha}_i$ is left unspecified and nonparametric ML is used.

We use respondents data on Y_{ij} ($i = 1,\dots,m, j \in r_i$) and population data

on covariates $\boldsymbol{x}_{ij}$ ($i = 1,\ldots,m, j = 1,\ldots,N_i$) to predict a (possibly) non-linear function of fixed and random effects, say $\zeta(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\Sigma})$. According to Jiang, 2003, the Best Predictor (BP) of ζ in terms of minimum MSE is given by $\tilde{\zeta}^{BP} = E_{\alpha|y}[\zeta(\boldsymbol{\beta},\boldsymbol{\alpha}_i,\boldsymbol{\Sigma}) \mid \boldsymbol{y}] = \int_{\mathbb{R}^{\upsilon}} \zeta(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\Sigma}) f_{\alpha|y}(\boldsymbol{\alpha} \mid \boldsymbol{y}) d\boldsymbol{\alpha}$, where

$$f_{\alpha|y}(\boldsymbol{\alpha} \mid \boldsymbol{y}) = \frac{\prod_{i=1}^{m} f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i;\boldsymbol{X}_i) f_{\alpha}(\boldsymbol{\alpha}_i)}{\prod_{i=1}^{m} f_y(\boldsymbol{y}_i;\boldsymbol{X}_i)},$$

$\boldsymbol{y} = (\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m)$ and $\upsilon = m \times q$. Estimates of model parameters can be obtained by maximizing the observed data likelihood function: $L(\boldsymbol{\Phi}) = \prod_{i=1}^{m} f_y(\boldsymbol{y}_i;\boldsymbol{X}_i)$. To maximize $L(\boldsymbol{\Phi})$, numerical approximations (e.g., Gaussian quadrature techniques) or simulation based methods (e.g., Monte Carlo integration) may be required. Once parameters are estimated, we may compute the empirical BP of ζ, that is $\hat{\zeta}^{EBP} = \tilde{\zeta}^{BP}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\Sigma}})$. To evaluate the quality of such predictions, the second-order MSE estimator can be considered as in Jiang, 2003 and in Marino *et al.* , 2019.

References

ALMALAUREA. 2022. *24th Report Occupational Condition of Graduates, 2022 Summary Report*. https://www.almalaurea.it/sites/default/files/2022-09/sintesi_occupazione_rapporto_2022_en.pdf. Accessed: 2023-04-21.

ALMALAUREA. 2023. *Graduates' employment status, data*. https://www2.almalaurea.it/cgi-php/universita/statistiche/tendine.php?anno=2021&LANG=en&config=occupazione. Accessed: 2023-04-21.

JIANG, J. 2003. Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference*, **111**, 117–127.

KOTT, P. S. 2006. Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, **32**, 133–142.

MARINO, M. F., RANALLI, M. G., SALVATI, N., & ALFÒ, M. 2019. Semiparametric empirical best prediction for small area estimation of unemployment indicators. *The Annals of Applied Statistics*, **13**, 1166–1197.

MATEI, A., & RANALLI, M. G. 2015. Dealing with non-ignorable nonresponse in survey sampling: A latent modeling approach. *Survey Methodology*, **41**, 145–165.

RAO, J. N. K., & MOLINA, I. 2015. *Small Area Estimation*. John Wiley & Sons.