Modified profile likelihood in models for clustered data with missing values

Verosimiglianza profilo modificata in modelli per dati raggruppati con valori mancanti

Claudia Di Caterina and Nicola Sartori

Abstract Clustered data are frequently subject to missing values, especially those collected from longitudinal studies. The main focus of the analysts is usually not on the clustering variables, hence the group-specific parameters are treated as nuisance. If a fixed effects formulation is preferred and the total number of clusters is large relative to the single-group sizes, classical frequentist techniques are often misleading. We propose here to combine multiple imputation and the modified profile likelihood function to obtain accurate inferences on a parameter of interest under models with incidental parameters for incomplete grouped observations. Such solution is examined via simulation studies which shed light on the convenience for the imputation model to take into account the clustered structure of the data.

Abstract *Nei dati raggruppati si registrano abitualmente valori mancanti, soprattutto in quelli raccolti per studi longitudinali. L'attenzione degli analisti di solito non è rivolta alle variabili di raggruppamento, dunque i parametri specifici dei vari cluster sono considerati di disturbo. Se si preferisce adottare una formulazione ad effetti fissi ed il numero totale di gruppi è grande rispetto alle singole dimensioni di questi, le classiche tecniche frequentiste risultano spesso inadeguate. Qui proponiamo di combinare l'imputazione multipla e la verosimiglianza profilo modificata per ottenere un'inferenza accurata sul parametro d'interesse in modelli con parametri incidentali per osservazioni incomplete organizzate in cluster. Tale soluzione viene esaminata attraverso studi di simulazione che fanno luce sull'opportunità che il modello di imputazione tenga conto della struttura raggruppata dei dati.*

Key words: fixed effects, incidental parameters, missing at random, multiple imputation.

Claudia Di Caterina
Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121, Padova, Italy; e-mail: dicaterina@stat.unipd.it

Nicola Sartori
Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121, Padova, Italy; e-mail: sartori@stat.unipd.it

1 Introduction

Clustered, stratified or grouped data are either cross-sectional or longitudinal observations that can be arranged in groups. Missing values are ubiquitous in quantitative research analysis, particularly in clustered data resulting from clinical trials or panel surveys. Depending on the pattern and mechanism of missingness, a variety of techniques for handling inference in the presence of incomplete datasets can be used (see, e.g., Little and Rubin, 2002). When observations are organized in many groups of small to moderate size, statistical models which capture the unobserved heterogeneity across clusters via group-specific nuisance parameters are likely to suffer from the incidental parameters problem (Neyman and Scott, 1948). Such specifications are referred to as fixed effects models, in opposition to the random effects models which require to assume a distribution for the group features and their incorrelation with the covariates in the model.

Here we focus our attention on clustered observations characterized by both aspects, and propose a twofold strategy. On the one hand, tackling the incompleteness of the data by means of multiple imputation, and on the other, dealing with the incidental parameters assumed by the model through the modified profile likelihood function. More details on the two approaches can be found in Sections 2 and 3, respectively. Section 4 shows simulation results that help to investigate how the considered inferential tools should be combined in order to draw reliable conclusions on the parameter of interest.

2 Multiple imputation

The basic rationale behind multiple imputation (MI) is to exploit the distribution of the observed data in order to estimate a set of plausible values for the unobserved data. In particular, m multiply imputed datasets are created by substituting the missing observations in the original sample with draws from the posterior predictive distribution of the unobserved data conditional on the observed data. These completed datasets are then separately analyzed and the m results are pooled into overall estimates and standard errors using Rubin's rules (Rubin, 1987).

Various methods can be adopted to generating imputations (Little and Rubin, 2002, Section 10.2). Among those drawing from pragmatic conditional distributions when more variables are incomplete, multiple imputation by chained equations (MICE) (van Buuren and Oudshoorn, 1999) provides considerable flexibility in customizing imputation models for different data characteristics (Ji et al., 2018). For a thorough overview of the standard procedure and a helpful guidance for practice in case of data missing at random (MAR), we refer to White et al. (2011).

A well-known matter in MI inference is uncongeniality (Meng, 1994), which occurs when the imputer's model class and the ultimate analyst's model class are incompatible. Recently, Xie and Meng (2017) have pointed out many open problems connected with this topic. The general prescription is to include in the imputation

model all variables that are related to the missing data, so that to make the MAR assumption more plausible. This should reduce the need to make special adjustments for mechanisms that are not MAR (van Buuren and Oudshoorn, 1999). With specific reference to models for clustered observations with incidental parameters, a typical question concerns whether and how accounting for the groups when imputing the missing values. White et al. (2011) suggest to disregard the clustering in this phase, if this is not of direct interest. Results in Andridge (2011) also highlight the inadequate inferential performance due to the inclusion of the fixed effects in the imputation model, contrary to what happens with random effects. On the opposite, Reiter et al. (2006) conclude that completely ignoring the sampling design during MI can be a risky practice. Further evidence is surely needed in this area.

3 Modified profile likelihood

In fixed effects models for clustered data where the number of groups is much larger than the single group sizes, the incidental parameters problem descends from the magnitude of the bias of the profile score function (McCullagh and Tibshirani, 1990). Correcting for the presence of the nuisance components, Barndorff-Nielsen (1980, 1983) proposed to rely on the modified profile likelihood (MPL) for making adequate inference on the parameter of interest. In fact, its superiority with respect to the ordinary profile likelihood (PL) within the two-index asymptotic setting can be proved for independent clustered sample units (Sartori, 2003).

For observations y_{it} subdivided in N groups of sizes T_i, suppose the model

$$Y_{it} \sim f(y_{it}; x_{it}, \psi, \lambda_i), \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T_i, \tag{1}$$

where x_{it} is a p-dimensional vector of covariates. The global parameter is $\theta = (\psi, \lambda)$, where $\psi \in \Psi \subseteq \mathbb{R}^k$ denotes the component of interest and $\lambda = (\lambda_1, \ldots, \lambda_N) \in \Lambda$ indicates the vector with incidental parameters. Note that, here and henceforth, to avoid clutter the transpose symbol acting on vectors is omitted. Moreover, we assume $T_i = T$ and $\dim(\lambda_i) = 1$ $(i = 1, \ldots, N)$ for the sake of notational simplicity. With independent groups, the log-likelihood function about θ can be expressed by

$$l(\theta) = \sum_{i=1}^{N} l^i(\theta) = \sum_{i=1}^{N} l^i(\psi, \lambda_i),$$

with $l^i(\psi, \lambda_i) = \sum_{t=1}^{T} \log p(y_{it}; x_{it}, \psi, \lambda_i)$. Let us define the full maximum likelihood (ML) estimate for model (1) as $\hat{\theta} = (\hat{\psi}, \hat{\lambda}) = \arg\max_\theta l(\theta)$. Standard inference on the parameter of interest is typically based on the profile log-likelihood

$$l_P(\psi) = \sum_{i=1}^{N} l^i\big(\psi, \hat{\lambda}_{i\psi}\big) = \sum_{i=1}^{N} l_P^i(\psi),$$

where $\hat{\lambda}_{i\psi}$ is the constrained ML estimate of λ_i for fixed ψ obtained, under usual regularity conditions, by equating to zero the score $l_{\lambda_i}(\theta) = \partial l^i(\psi,\lambda_i)/\partial\lambda_i$ and solving for λ_i $(i = 1,\ldots,N)$. Given $\hat{\lambda}_\psi = (\hat{\lambda}_{1\psi},\ldots,\hat{\lambda}_{N\psi})$, the full constrained ML estimate for fixed ψ is denoted by $\hat{\theta}_\psi = (\psi,\hat{\lambda}_\psi)$.

The general expression taken by the logarithmic version of the MPL is $l_M(\psi) = l_P(\psi) + M(\psi)$, and one computationally convenient formulation of $M(\psi)$ is owed to Severini (1998). Specifically, using the additive form $M(\psi) = \sum_{i=1}^{N} M_i(\psi)$, the ith summand in Severini's modification term equals

$$M_i(\psi) = \frac{1}{2}\log j_{\lambda_i\lambda_i}(\hat{\theta}_\psi) - \log I_{\lambda_i\lambda_i}(\hat{\theta};\hat{\theta}_\psi), \qquad i = 1,\ldots,N.$$

In the above equation, we have $j_{\lambda_i\lambda_i}(\theta) = -\partial^2 l^i(\psi,\lambda_i)/(\partial\lambda_i\partial\lambda_i)$ and $I_{\lambda_i\lambda_i}(\hat{\theta};\hat{\theta}_\psi) = E_{\theta_0}\{l_{\lambda_i}(\theta_0)l_{\lambda_i}(\theta_1)\}\big|_{\theta_0=\hat{\theta},\theta_1=\hat{\theta}_\psi}$ indicating the scalar expected value calculated with regard to $\hat{\theta}$ of the product of partial scores evaluated at two different points in the parameter space.

4 Simulation studies

Monte Carlo experiments based on 1000 iterations can be run to study the effectiveness of the approach which incorporates MI and MPL inferences. The cluster size and number of groups considered are $T = 6$ and $N = 50, 100, 250$, respectively. For each couple (T,N), $p = 2$ covariates are randomly generated. The first, x_{1it}, is sampled from a Bernoulli(0.5) distribution. The second, x_{2it}, is drawn from the $N(0,1)$ random variable. We then simulate the binary clustered outcomes as independent realizations of $Y_{it} \sim Bern(\pi_{it})$ $(i = 1,\ldots,N, t = 1,\ldots,T)$. In particular $\pi_{it} = e^{\lambda_i+\beta x_{it}}/(1+e^{\lambda_i+\beta x_{it}})$, where $x_{it} = (x_{1it},x_{2it})$, $\beta = (\beta_1,\beta_2) = (1,1)$ and each λ_i is independently generated from the standard normal distribution. Here our interest is confined to datasets with completely observed response and MAR predictors, yet the same methodology could be applied in different contexts of incompleteness. Missing entries are thus created by deleting x_{1it} with probability ω_{1it} and x_{2it} with probability ω_{2it}. According to the dependence of the missingness on the stratification, two main scenarios may be distinguished. In case of missingness unrelated to the groups (setting I), we suppose $\omega_{1it} = e^{-2+y_{it}}/(1+e^{-2+y_{it}})$ and $\omega_{2it} = e^{0.5-y_{it}}/(1+e^{0.5-y_{it}})$. When instead the clustered structure plays a role in the probability of observing a covariate (setting II), we use $\omega_{1it} = e^{\lambda_i-2+0.2y_{it}}/(1+e^{\lambda_i-2+y_{it}})$ and $\omega_{2it} = e^{\lambda_i-y_{it}}/(1+e^{\lambda_i-y_{it}})$. Such values are chosen in order for the fraction of missing data in the datasets to be around 35%. The procedure starts by obtaining $m = 5$ complete samples through MICE, using a logistic regression for imputing x_{1it} and a Bayesian linear regression for x_{2it}, as implemented by the R package `mice` (van Buuren and Groothuis-Oudshoorn, 2010). In both imputation models, the dummy variables indicating the groups are either included or not. Inference on each completed dataset is then conducted using PL and MPL for the parame-

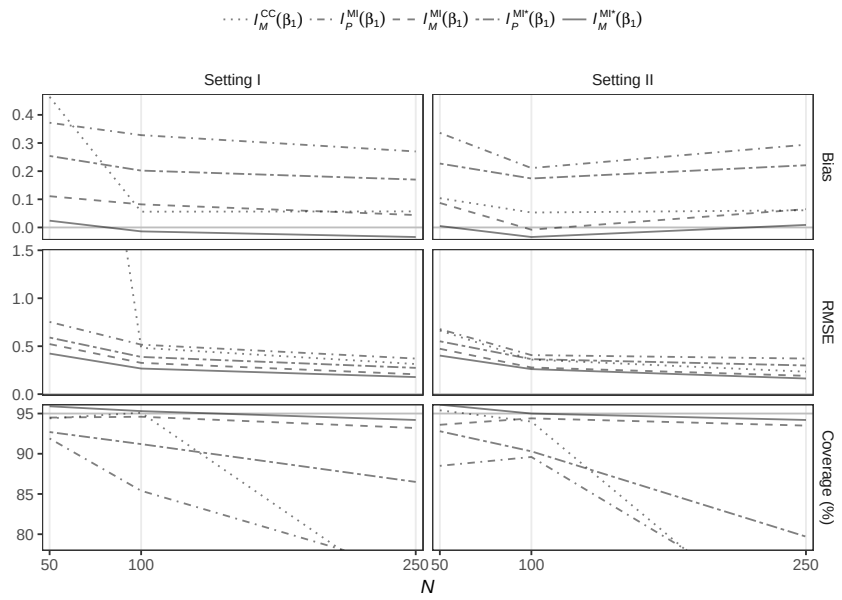

Fig. 1 Comparison between inferences on β_1 obtained via complete-case analysis $\left(l_M^{\mathrm{CC}}\right)$ and multiple imputation, either taking the groups into consideration when imputing the missing values $\left(l_P^{\mathrm{MI}}, l_M^{\mathrm{MI}}\right)$ or not $\left(l_P^{\mathrm{MI}^*}, l_M^{\mathrm{MI}^*}\right)$. Results based on 1000 clustered datasets simulated with $T = 6$ observations per group and $N = 50, 100, 250$ number of groups.

ter of interest β in the logistic regression with outcomes y_{it}, using the R package `panelMPL` (Bellio and Sartori, 2015). Rubin's rules are finally applied to pool together estimates and variances derived by the two likelihood functions. A complete-case analysis, which disregards units with missing values, is also carried out via both methods. Due to space constraints, just partial results of the experiments referred to β_1 are shown in Figure 1. Therein, empirical bias and root mean squared error (RMSE) of the various estimators can be compared, along with coverage of 95% Wald confidence intervals. Note that the performance of the complete-case PL is not reported, as it was found to be poorer than any other method in all scenarios. The output indicates that the solution combining MI and MPL outperforms the complete-case analysis, in terms of point and interval estimation. In addition, it seems that to neglect the clustering while imputing the unobserved covariates is recommendable, whether the incompleteness depends on the specific group features or not. One plausible reason is the incidental parameters problem observed under the imputation model. An improved fit of the latter might be achieved, for instance, by adopting bias reduction (Firth, 1993). This can be the subject of future research, as well as developments of the present work that consider other values for T and different patterns and mechanisms of missingness in the data.

References

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical journal 53*, 57–74.

Barndorff-Nielsen, O. E. (1980). Conditionality resolutions. *Biometrika 67*, 293–310.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika 70*, 343–365.

Bellio, R. and N. Sartori (2015). panelMPL: *Modified profile likelihood estimation for fixed-effects panel data models*. http://ruggerobellio.weebly.com/software.html.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*, 27–38.

Ji, L., S.-M. Chow, A. C. Schermerhorn, N. C. Jacobson, and E. M. Cummings (2018). Handling missing data in the modeling of intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–22.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (2nd ed.). Wiley, New York.

McCullagh, P. and R. Tibshirani (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological) 52*, 325–344.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science 9*, 538–558.

Neyman, J. and E. Scott (1948, January). Consistent estimates based on partially consistent observations. *Econometrica 16*, 1–32.

Reiter, J. P., T. E. Raghunathan, and S. K. Kinney (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology 32*, 143.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika 90*, 533–549.

Severini, T. A. (1998). An approximation to the modified profile likelihood function. *Biometrika 85*, 403–411.

van Buuren, S. and K. Groothuis-Oudshoorn (2010). MICE: Multivariate imputation by chained equations in R. *Journal of statistical software*, 1–68.

van Buuren, S. and K. Oudshoorn (1999). Flexible multivariate imputation by MICE. Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054. For associated software see http://www.multiple-imputation.com.

White, I. R., P. Royston, and A. M. Wood (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine 30*, 377–399.

Xie, X. and X.-L. Meng (2017). Dissecting multiple imputation from a multi-phase inference perspective: what happens when gods, imputers and analysts models are uncongenial. *Statistica Sinica 27*, 1485–1594.