Computational Statistics and Data Analysis 107 (2017) 32–47

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Poisson mixed models for studying the poverty in small areas

Miguel Boubeta [a], María José Lombardía [a,*], Domingo Morales [b]

[a] *Department of Mathematics, Universidade da Coruña, A Coruña, Spain*

[b] *Centro de Investigación Operativa, Universidad Miguel Hernández de Elche, Elche, Spain*

ARTICLE INFO

Article history:
Received 17 November 2015
Received in revised form 17 October 2016
Accepted 17 October 2016
Available online 24 October 2016

Keywords:
Bootstrap
Empirical best predictor
Mean squared error
Moment-based method
Poisson mixed models
Poverty proportion
Time dependency

ABSTRACT

Empirical best predictors are studied under area-level Poisson mixed models with time effects. Four cases are considered. The first two cases use independent time random effects. In the second two cases, the time effects follow an autoregressive process of order one. The four models are fitted by the moment-based method and the corresponding empirical best predictors are derived and compared with plug-in predictors. Several simulation experiments investigate the performance of both predictors. A parametric bootstrap procedure is considered for estimating the mean squared error. The developed methodology is applied to estimate the proportion of people under the poverty line by counties and sex in Galicia (a region in north-west of Spain).

1. Introduction

Developed countries are interested on the estimation of poverty indicators for the whole country and for small territories. In Galicia (north-west of Spain) the Spanish living conditions survey (SLCS) provides information about people with equivalent personal incomes below the poverty line. This survey has a planned sample size large enough for obtaining reliable direct estimates for Galicia, but not for their 53 counties. Small area estimation (SAE) deals with this kind of problems by introducing indirect estimators. See the monograph of Rao and Molina (2015) or the reviews of Jiang and Lahiri (2006), Rao (2008) and Pfeffermann (2013) for an introduction to SAE.

This paper takes data from the SLCS for estimating poverty proportions. The indicator of interest is defined as the count of people under the poverty line divided by the size of the small area. For estimating counts and proportions, the SAE model-based approach employs models at the unit or at the area level. In both cases linear mixed model (LMM) and generalized linear mixed models (GLMM) are used. Nevertheless, the literature shows more applications based on area-level than on unit-level models. This is because the former applications are in general easier to carry out because of the simplicity of the procedure and the availability of aggregated data from administrative registers.

Under the area-level approach, Esteban et al. (2011, 2012), Marhuenda et al. (2013, 2014) and Morales et al. (2015) derived poverty proportion estimators based on linear mixed models. They reported poverty indicators for Spanish provinces. Saei and Chambers (2003), Johnson et al. (2010), Chandra et al. (2011), Chambers et al. (2012) and López-Vizcaíno et al. (2013, 2015) applied area-level logit regression models to the estimation of domain counts or proportions. Similarly, Tzavidis et al. (2015) and Boubeta et al. (2016) applied Poisson regression models for estimating the same type of parameters. We follow the idea of using Poisson models for estimating counts and we introduce two temporal area-level mixed models.

* Corresponding author.
E-mail address: maria.jose.lombardia@udc.es (M.J. Lombardía).

http://dx.doi.org/10.1016/j.csda.2016.10.014

The first two Poisson models use independent time random effects. The second two models assume that time random effects follow an autoregressive process of order one. The four models are fitted by the moments-based (MB) method introduced by Jiang (1998) for GLMM.

We derive empirical best predictors (EBP), based on the introduced temporal area-level Poisson mixed models, for estimating counts and proportions. The statistical methodology is taken and adapted from Jiang and Lahiri (2001) and Jiang (2003). In addition to the EBP, a plug-in predictor is given and empirically studied in simulation experiments. For estimating the EBP mean squared error (MSE), we consider the parametric bootstrap MSE estimator introduced by González-Manteiga et al. (2007, 2008a) in the context of logistic and normal mixed models and later extended by González-Manteiga et al. (2008b) to a multivariate area-level model. We present an application of the developed methodology to data from the 2010–2013 SLCS of Galicia. The target of the application is the estimation of poverty proportions at county level by sex.

The paper is organized as follows. Section 2 introduces four area-level Poisson mixed models and the employed model-based fitting algorithm. Section 3 presents the EBP and the plug-in predictors of functions of fixed and small area specific random effects. Section 4 presents two simulation experiments. The first simulation studies the behavior of the MB fitting algorithm. The second simulation compares the performances of the EBP and the plug-in predictors. Section 5 applies the developed methodology to data from the 2010–2013 SLCS of Galicia. The target is the estimation of poverty proportions at county level by sex. Section 6 gives some conclusions. Appendix contains detailed mathematical derivations for implementing the MB algorithms.

2. The models

This section introduces four area-level Poisson mixed models with time effects and their fitting algorithms. The first two models (Models 1 and 1_2) have independent time random effects. The random effects of the second models (Models 2 and 2_2) follow an AR(1) autoregressive process within each domain. Along this paper, D and T denote the total numbers of domains and time instants respectively. The corresponding indexes are d and t, where $d = 1, \ldots, D$ and $t = 1, \ldots, T$.

2.1. Models with independent time effects

This section introduces two temporal models with independent time effects. Both models assume that the temporal correlation of the target variable is fully described by the auxiliary variables. Model 1 considers two independent sets of random effects such that $\{v_{1,d} : d = 1, \ldots, D\}$ and $\{v_{2,dt} : d = 1, \ldots, D, t = 1, \ldots, T\}$ are i.i.d. $N(0, 1)$. They denote the area and the interaction area-time effects that are not explained by the fixed part of the model. The distribution of the target variable y_{dt}, conditioned to the random effects $v_{1,d}$ and $v_{2,dt}$, is

$$y_{dt}|v_{1,d}, v_{2,dt} \sim \text{Poisson}(\mu_{dt}), \quad d = 1, \ldots, D, \ t = 1, \ldots, T. \tag{1}$$

Given the relationship between Poisson and binomial distributions, we take $\mu_{dt} = \nu_{dt} p_{dt}$, where ν_{dt} and p_{dt} are size and probability parameters respectively. In practice, ν_{dt} is known and equal to the sample size of domain d at time instant t. For the natural parameter, we assume that it can be expressed in terms of a set of auxiliary variables through a regression model, i.e.

$$\text{Model 1: } \log \mu_{dt} = \log \nu_{dt} + \boldsymbol{x}_{dt}\boldsymbol{\beta} + \phi_1 v_{1,d} + \phi_2 v_{2,dt}, \quad d = 1, \ldots, D, \ t = 1, \ldots, T, \tag{2}$$

where $\boldsymbol{\beta} = \text{col}_{1\le k\le p}(\beta_k)$ is the column vector of regression coefficients, $\boldsymbol{x}_{dt} = \text{col}'_{1\le k\le p}(x_{dtk})$ is the row vector of auxiliary variables and ϕ_1 and ϕ_2 are the variance component parameters. If we define $u_{1,d} = \phi_1 v_{1,d}$ and $u_{2,dt} = \phi_2 v_{2,dt}$, then ϕ_1 and ϕ_2 are the variances of $u_{1,d}$ and $u_{2,dt}$ respectively. These variances can be interpreted as the variability between domain and between time periods within each domain respectively.

Further, we assume that the y_{dt}'s are independent conditioned to $\boldsymbol{v}_1 = \text{col}_{1\le d\le D}(v_{1,d})$ and $\boldsymbol{v}_2 = \text{col}_{1\le d\le D}(\boldsymbol{v}_{2,d})$, where $\boldsymbol{v}_{2,d} = \text{col}_{1\le t\le T}(v_{2,dt})$. We have that

$$P\left(y_{dt}|\boldsymbol{v}_1, \boldsymbol{v}_2\right) = P(y_{dt}|v_{1,d}, v_{2,dt}) = \frac{1}{y_{dt}!}\exp\{-\nu_{dt}p_{dt}\}\nu_{dt}^{y_{dt}}p_{dt}^{y_{dt}},$$

where $p_{dt} = \exp\{\boldsymbol{x}_{dt}\boldsymbol{\beta} + \phi_1 v_{1,d} + \phi_2 v_{2,dt}\}$. For fitting the area-level Poisson mixed model with independent time effects, we use the MB algorithm based on the method of simulated moments suggested by Jiang (1998). A natural set of equations for applying this method is

$$\begin{aligned}
0 &= f_k(\boldsymbol{\theta}) = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}E_{\boldsymbol{\theta}}[y_{dt}]x_{dtk} - \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}y_{dt}x_{dtk}, \quad k = 1, \ldots, p,\\
0 &= f_{p+1}(\boldsymbol{\theta}) = \frac{1}{D}\sum_{d=1}^{D}E_{\boldsymbol{\theta}}[y_{d.}^2] - \frac{1}{D}\sum_{d=1}^{D}y_{d.}^2,\\
0 &= f_{p+2}(\boldsymbol{\theta}) = \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}E_{\boldsymbol{\theta}}[y_{dt}^2] - \frac{1}{DT}\sum_{d=1}^{D}\sum_{t=1}^{T}y_{dt}^2,
\end{aligned} \tag{3}$$