Journal of Multivariate Analysis 111 (2012) 213–221

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

Likelihood inference in small area estimation by combining time-series and cross-sectional data

Mahmoud Torabi [a,*], Farhad Shokoohi [b]

[a] *Department of Community Health Sciences, University of Manitoba, MB, R3E 0W3, Canada*

[b] *Department of Statistics, Shahid Beheshti University, Tehran, 1983963113, Iran*

ARTICLE INFO

Article history:
Received 7 September 2011
Available online 2 June 2012

AMS subject classifications:
62D99
62F12
62H10
62M10
62M20

Keywords:
Autocorrelated errors
Bayesian computation
Hierarchical model
Prediction interval and exponential family
Random effect

ABSTRACT

Using both time-series and cross-sectional data, a linear model incorporating autocorrelated random effects and sampling errors was previously proposed in small area estimation. However, in practice there are many situations that we have time-related counts or proportions in small area estimation; for example a monthly dataset on the number of incidences in small areas. The frequentist analysis of these complex models is computationally difficult. On the other hand, the advent of the Markov chain Monte Carlo algorithm has made the Bayesian analysis of complex models computationally convenient. Recent introduction of the method of data cloning has made frequentist analysis of mixed models also equally computationally convenient. We use data cloning to conduct frequentist analysis of small area estimation for Normal and non-Normal data situations with incorporating cross-sectional and time-series data. Another important feature of the proposed approach is to predict small area parameters by providing prediction intervals. The performance of the proposed approach is evaluated through several simulation studies and also by a real dataset.

1. Introduction

Small area estimation has received a lot of attention in recent years due to the growing demand for reliable small area statistics. Rao [17], Jiang and Lahiri [10] and Jiang [9] have given comprehensive accounts of model-based small area estimation. In particular, area level [4] and nested error linear regression models [1,15] are often used in small area estimation to obtain efficient model-based estimators of small area means.

Most of the research on small area estimation has focused on cross-sectional data at a given point in time, and the research based on time series in the context of small area estimation is limited. Scott and Smith [19], Jones [11] among others used time-series methods to develop efficient estimates of aggregated parameters from repeated surveys. Tiller [24] used the idea of a Kalman filter to combine a current-period state-wide estimate from the US Current Population Survey with past estimates for the same state. However, none of them studied small area estimation by combining cross-sectional and time-series data.

Pfeffermann and Burck [14] and Singh et al. [20] among others studied cross-sectional and time-series models for small area estimation using Kalman filter by assuming specific models for the sampling errors over time. In a pioneering paper, Rao and Yu [18] proposed a combined cross-sectional and time-series model involving autocorrelated random effects and sampling errors with an arbitrary covariance matrix over time. Datta et al. [3] applied the same model as the Rao–Yu model but replacing autoregressive (AR) random effects with the random walk model. Datta et al. [2] considered a similar model but added extra terms to reflect seasonal variation in their application. Torabi [25] extended the Datta et al. [2] model to account for spatial variation over regions.

* Corresponding author.
E-mail address: torabi@cc.umanitoba.ca (M. Torabi).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.05.016

The main purpose of this paper is to extend the Rao–Yu model for non-Normal data using the frequentist paradigm. There are many applications in small area estimation where responses are time-related counts or proportions. For example, we may be interested to analyze the monthly dataset of a number of incidences in small areas. Indeed, these types of models fall in the class of Generalized Additive Mixed Models (GAMMs). It is well known that the parameter estimation and prediction of small area statistics under GAMM are extremely difficult using the frequentist approach. The Bayesian approach, especially the non-informative Bayesian approach, has become quite popular because of its computational convenience and the ability to provide not just the point predictors but also the associated prediction intervals. However, the implementation of a non-informative Bayesian approach requires substantial care. The inferences may also depend on the choice of prior.

Recently, Lele et al. [12] introduced an alternative approach, called data cloning (DC), to compute the maximum likelihood (ML) estimates and their standard errors for general hierarchical models. Similar to the Bayesian approach, data cloning avoids high dimensional numerical integration and requires neither maximization nor differentiation of a function. Extending this work to the Generalized Linear Mixed Model (GLMM) situation, Lele et al. [13] described an approach to compute prediction and prediction intervals for the random effects. We use the idea of data cloning to extend the Rao–Yu model with incorporating cross-sectional and time-series to non-Normal data using the frequentist paradigm. Because these estimators are ML estimators, unlike the Bayesian estimators, they are independent of the choice of priors and non-estimable parameters are also flagged automatically.

In this paper, we use data cloning to propose a combined cross-sectional and time-series model with AR(1) for Normal and non-Normal data. In the next section, we describe the combined cross-sectional and time-series models. We then describe how data cloning can be used to obtain model parameters estimate and also to get prediction and prediction intervals for small area parameters. The performance of the proposed approach is reported through several simulation studies and also by an application to a real dataset. Finally, some concluding remarks are given.

2. Cross-sectional and time-series models

The basic model for the combined cross-sectional and time-series data can be described as follows. Let y_{it} be the variable of interest for the ith area in given time $t (t = 1, \ldots, T; i = 1, \ldots, m)$. The y_{it} are assumed to be conditionally independent with exponential family p.d.f.

$$f(y_{it}|\theta_{it}, \phi_{it}) = \exp[\{y_{it}\theta_{it} - a(\theta_{it})\}/\phi_{it} + b(y_{it}, \phi_{it})], \tag{2.1}$$

$(t = 1, \ldots, T; i = 1, \ldots, m)$. The density (2.1) is parameterized with respect to the canonical parameters θ_{it}, known scale parameters ϕ_{it} and functions $a(\cdot)$ and $b(\cdot)$. The exponential family (2.1) covers well-known distributions including Normal, binomial and Poisson distributions. The natural parameters θ_{it} are then modeled as

$$h(\theta_{it}) = \boldsymbol{x}'_{it}\boldsymbol{\beta} + v_i + u_{it} \ (t = 1, \ldots, T; i = 1, \ldots, m),$$

where h is a strictly increasing function, the $\boldsymbol{x}_{it} (p \times 1)$ are known design vectors, $\boldsymbol{\beta}(p \times 1)$ is an unknown vector regression coefficient, $v_i \overset{i.i.d.}{\sim} N(0, \sigma_v^2)$, and u_{it}'s are assumed to follow a common $AR(1)$ process for each i, that is,

$$u_{it} = \rho u_{i,t-1} + \epsilon_{it}, \quad |\rho| < 1,$$

with $\epsilon_{it} \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2)$.

As special case, under Normal distribution, $h(\theta_{it}) = \theta_{it}$, the Rao–Yu model is given by

$$\hat{\theta}_{it} = \theta_{it} + e_{it} (t = 1, \ldots, T; i = 1, \ldots, m),$$

where e_{it}'s are sampling errors normally distributed, given the θ_{it}'s, with zeros means and a known block diagonal covariance matrix Ψ with blocks Ψ_i. The errors $(v_i, \epsilon_{it}, e_{it})$ are also assumed to be independent of each other. For the case of an AR(1) model with ρ known, Rao and Yu [18] estimated σ_ϵ^2 and σ_v^2 by extending the simple transformation method of Fuller and Battese [5]. Replacing σ_ϵ^2 and σ_v^2 by their estimators $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_v^2$, they got the empirical best linear unbiased predictor (EBLUP) of θ_{it}, $\hat{\theta}_{it}(\rho)$, under the AR(1) model with ρ known. Rao and Yu [18] also obtained a second-order approximation to the estimator of mean squared prediction error (MSPE) of $\hat{\theta}_{it}(\rho)$ using the Taylor expansion. For the case of ρ unknown, Rao and Yu [18] obtained a consistent estimator $\hat{\rho}$ and pointed out that this estimator often takes values outside the admissible range (−1,1), particularly for small T or small σ_ϵ^2 relative to the sampling variation. To avoid this difficulty, they proposed a naive estimator of ρ, $\hat{\rho}_N$, which is inconsistent and underestimates ρ in the presence of sampling errors. Although, the resulting EBLUP estimator $\hat{\theta}_{it}(\hat{\rho}_N)$ was unbiased, but the corresponding estimator of MSPE was not correct to terms of order $o(m^{-1})$.

3. Frequentist inference using data cloning

Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)'$ be the observed data vector and, conditionally on the random effects, $\boldsymbol{w} = (v_1, \ldots, v_m, u_{11}, \ldots, u_{mT})'$, assume that the elements of $\boldsymbol{y}$ are independent and drawn from a distribution in the exponential family with parameters $\boldsymbol{\beta}$ where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})$, $(i = 1, \ldots, m)$. It is also assumed that the distribution for $\boldsymbol{w}$ depends on parameters $(\rho, \sigma_v^2, \sigma_\epsilon^2)$. The goal of the analysis is to estimate the model parameters $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \rho, \sigma_v^2, \sigma_\epsilon^2)'$ and predict the small area parameters $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{mT})'$.