Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 2720 – 2733
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model

W. González-Manteiga[a], M.J. Lombardía[a,*], I. Molina[b], D. Morales[c], L. Santamaría[c]

[a]*Department of Statistics and Operations Research, Universidad de Santiago de Compostela, 15782 Santiago de Compostela, Spain*
[b]*Department of Statistics, Universidad Carlos III de Madrid, C/Madrid 126, Getafe 28903, Madrid, Spain*
[c]*Operations Research Center, Universidad Miguel Hernández de Elche, Av/Universidad, s/n, Elche 03202, Spain*

Received 31 May 2005; received in revised form 19 January 2006; accepted 19 January 2006
Available online 10 February 2006

Abstract

An accuracy measure (mean squared error, MSE) is necessary when small area estimators of linear parameters are provided. Even in the case when such estimators arise from the assumption of relatively simple models for the variable of interest, as linear mixed models, the analytic form of the MSE is not suitable to be calculated explicitly. Some good and widely used approximations are available for those models. For generalized linear mixed models, a rough approximation can be obtained by a linearization of the model and application of Prasad–Rao approximation for linear mixed models. Resampling methods, although computationally demanding, represent a conceptually simple alternative. Under a logistic mixed linear model for the characteristic of interest, the Prasad–Rao-type formula is compared with a bootstrap estimator obtained by a wild bootstrap designed for estimating under finite populations. A simulation study is developed in order to study the performance of both methods for estimating a small area proportion.

Keywords: Resampling methods; Bootstrap; Superpopulation model; Logit mixed regression model; Small area estimation; Mean squared error

1. Introduction

Survey sample sizes are usually planned to produce reliable estimates for large geographic areas. However, the demand of estimates for smaller areas (subdomains) using the same survey data; that is, without increasing sample sizes, is growing in both public and private sectors. But the resulting sizes for those small areas are rarely large enough to provide direct estimates with acceptable precision, and consequently some new methods have been developed. The discipline called small area estimation gathers all that techniques for obtaining precise estimators for small areas. Generally, two main problems are treated, namely the estimation of some area characteristic and the assessment of the accuracy of proposed estimators. This manuscript deals with the second topic; more concretely, the estimation of the mean squared error (MSE) of predictors of linear parameters in small areas, when the characteristic of interest follows a logistic mixed model.

Models have become a powerful tool in small-area estimation, since the resulting estimators borrow strength from related areas and from auxiliary information. There are two main approaches related to the use of models. Model-based

* Corresponding author. Departamento de Estadística e Investigación Operativa, Facultad de Matemáticas, 15782 Santiago de Compostela, Spain. Tel.: +34 981 563100; fax: +34 981 597054.
E-mail address: mjoselc@usc.es (M.J. Lombardía).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2006.01.012

methods assume that the observations follow a particular model, and the inference is made with respect to the assumed model (see, e.g., Ghosh and Rao, 1994; Rao, 1999 for excellent reviews about model-based literature). Model-assisted approaches base the inference on the sampling distribution, using models to assist the estimation of hyperparameters (see, e.g., Estevao and Särndal, 1999). Closer to the scope of this paper, Lehtonen and Veijanen (1998) and Lehtonen et al. (2003) introduced a logistic generalized regression estimator by assuming a multinomial logistic model with fixed effects. For further references on the application of mixed models to small area estimation, see, e.g., Rao (2003). The framework of this paper is exclusively model based.

In current literature about model-based small area estimation, the underlying models often belong to the class of linear mixed models. But even in the simplest special cases of these models, the estimation of the MSE of estimated small area linear parameters requires approximation. Some approximations of the MSE appearing in the literature for these models have been given by Kackar and Harville (1981), Prasad and Rao (1990), Datta and Lahiri (2000) and Das et al. (2004). With more detail, Prasad and Rao (1990) provided an approximation of the MSE of the EBLUP under three types of linear mixed models, including the linearized model (7), when the variance components are estimated by Henderson's method 3. When the variance components are estimated by ML or RML, Das et al. (2004) showed that the same approximation given by Prasad and Rao is still valid.

A more complex situation for MSE estimation appears when a generalized linear mixed model is assumed for the target variable. There is some interesting research done in this area using hierarchical Bayes methodology (see, e.g., Malec et al., 1997; Ghosh et al., 1998; Farrell et al., 1997). In the frequentist framework, Jiang and Lahiri (2001) proposed an empirical best predictor (in terms of the MSE) and obtained an approximation of the MSE correct up to order $o(D^{-1})$, where D is the number of small areas. Jiang (2003) extended the Jiang–Lahiri results to generalized linear mixed models. In this paper a simpler (although not "best") predictor with a closed formula is used. The shape of the predictor allows linearization and subsequent approximation of its MSE.

Concerning the fit of the model, in generalized linear mixed models the marginal loglikelihood is represented by an integral that cannot be explicitly calculated. Several methods have been proposed to overcome this problem, most of them relying on Taylor linearizations. Goldstein (1991) considers a Taylor linearization of the inverse link function and then applies standard estimation procedures for linear multilevel models. Longford (1994) and Breslow and Clayton (1993) apply Laplace's method for integral approximations. Wolfinger and O'Connell (1993) approximate the conditional distribution of the difference between the response variable and its prediction, given the fixed and random effects, by a Gaussian distribution with the same first two moments. The procedure is implemented via iterated fitting of a weighted Gaussian linear mixed model to the modified dependent variable, which is obtained by a Taylor series approximation of the linked response. McCulloch (1994, 1997) and Booth and Hobert (1999) use EM-type algorithms assisted by Monte Carlo methods. Schall (1991) and McGilchrist (1994) use the PQL algorithm introduced by Breslow and Clayton (1993) in combination with a Gaussian approximation of the marginal density that provides approximate maximum likelihood estimators of variance components, obtaining a double iteration scheme. Although it is known that in some cases the method proposed by Schall (1991) may lead to inconsistent and biased estimators, this method works empirically well in the simulated example described in Section 5 (see Table 1). Further, the Schall method is conceptually very simple and allows for the adaptation of the prediction and MSE estimation theory under linear mixed models to the case of generalized linear models. For this reason, in this paper the Schall (1991) proposal is used.

Nowadays, bootstrap (Efron, 1979) is a computer intensive method widely used in the i.i.d. case. It has also been adapted to handle complex issues in survey sampling. In the framework of model-based small area estimation, parametric bootstraps for linear mixed models to estimate the MSE of the EBLUP (empirical best linear unbiased predictor) or the EBP (empirical Bayes predictor) have been proposed by Laird and Louis (1987), Pfefferman and Tiller (2002) and Butar and Lahiri (2003). For an application of Butar–Lahiri parametric bootstrap for nonnormal mixed models see Lahiri and Maiti (2002). Pfefferman and Tiller (2002) also propose a nonparametric bootstrap. However, there is lack of research about bootstrap estimation methods under generalized linear mixed models.

In this work, a logistic mixed model is assumed for the characteristic of interest, and two alternative approximations of the MSE of an empirical predictor are discussed and compared empirically. The first one is a Prasad–Rao-type MSE estimator derived from a Taylor series approximation, while the second is a bootstrap-based estimator. The specific version of the bootstrap technique is called "wild bootstrap" and it is well known for example in the context of linear modelling of heteroskedastic data where the data generating process is partly unknown (Wu, 1986). Simulations show that the (at first sight rough) Prasad–Rao-type estimator succeeds to approximate the true value except for areas for