Computational Statistics and Data Analysis 70 (2014) 268–280

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Information criteria for Fay–Herriot model selection

Yolanda Marhuenda [a,*], Domingo Morales [a], María del Carmen Pardo [b]

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

[b] *Departamento de Estadística e Investigación Operativa I, Universidad Complutense de Madrid, Spain*

ARTICLE INFO

Article history:
Received 7 December 2012
Received in revised form 18 September 2013
Accepted 18 September 2013
Available online 25 September 2013

Keywords:
Small area estimation
Fay–Herriot model
Akaike information criterion
Kullback symmetric divergence criterion
Model selection
Bootstrap

ABSTRACT

The selection of an appropriate model is a fundamental step of the data analysis in small area estimation. Bias corrections to the Akaike information criterion, *AIC*, and to the Kullback symmetric divergence criterion, *KIC*, are derived for the Fay–Herriot model. Furthermore, three bootstrap-corrected variants of *AIC* and of *KIC* are proposed. The performance of the eight considered criteria is investigated with a simulation study and an application to real data. The obtained results suggest that there are better alternatives than the classical *AIC*.

1. Introduction

We consider the problem of model selection for the well known area-level linear mixed model due to Fay and Herriot (1979). Estimates of small area parameters based on area-level models are generally more efficient than direct estimates. Random intercept models, like the Fay–Herriot model, are used when the variability between areas is not sufficiently explained by the auxiliary variables. Small area estimation is an increasingly important part of survey sample inference with applications to social and economic statistics. Small-area models improve the accuracy of direct area estimators by including, via modeling, the information of all sample observations and not just the one within the corresponding area, and by making use of auxiliary information. Statistical estimation and inference for these models have been largely described by Rao (2003) and Jiang and Lahiri (2006).

In the literature of both the general small area estimation and the Fay–Herriot model, there exist many works for variable selection. The very first paper of Fay and Herriot (1979) discussed variable selection. Early works often used ad hoc approaches such as validation samples and the classic Akaike information criterion (*AIC*). The *AIC* was introduced by Akaike (1973). Recent works were focused on general model selection approaches including *AIC* and other related criteria. Longford (2005) suggested a model averaging approach that requires deriving the covariances among estimates from different candidate models. Jiang et al. (2008) developed the adaptative fence method, Datta et al. (2011) focused on the random effect while fixing the choice of covariates. Burnham and Anderson (2002), Yang (2005), Kubokawa (2011) and Han (2013), among others, proposed *AIC* related criteria.

The *AIC* has been applied in different settings since its derivation is quite general. The *AIC* was designed to be an approximately unbiased estimator of the expected Kullback–Leibler information of a fitted model. In general, it is a good estimator

* Correspondence to: Centro de Investigación Operativa, Universidad Miguel Hernández de Elche, Avda. de la Universidad s/n, 03202 Elche, Alicante, Spain. Tel.: +34 966658536; fax: +34 966658721.

E-mail address: y.marhuenda@umh.es (Y. Marhuenda).

0167-9473/$ – see front matter
http://dx.doi.org/10.1016/j.csda.2013.09.016

when the sample size is large and the number of unknown parameters is small. In other settings the *AIC* may present a large negative bias when estimating the Kullback–Leibler information, which limits its usefulness as a model selection criterion. Hurvich and Tsai (1989) derived the corrected Akaike information criterion (*AICc*) for regression and autoregressive time series models. The *AICc* has been extended to autoregressive moving-average modeling, vector autoregressive modeling and multivariate regression modeling by Hurvich et al. (1990), Hurvich and Tsai (1993) and Bedrick and Tsai (1994), respectively. The *AICc* often outperforms the *AIC* but it is less used since its justification depends on the particular underlying model. Other interesting corrected variants of the *AIC* were given by Shang and Cavanaugh (2008). They proposed two bootstrap-corrected versions for the joint selection of the fixed and random components of a linear mixed model.

The Kullback–Leibler divergence, used to develop the *AIC*, assesses the dissimilarity between two statistical models. In fact, this is what every divergence measure does. Therefore, we can think of substituting the Kullback–Leibler divergence measure by any other divergence measure in order to define a new model selection criterion. Cavanaugh (1999, 2004) proposed to consider the Kullback symmetric divergence criterion for linear model selection. As the Kullback symmetric divergence is the sum of the Kullback–Leibler divergence and the divergence obtained by reversing the roles of the two models in the definition of the Kullback–Leibler measure, it captures different characteristics of the model. The proposed criteria (*KIC* and *KICc*) are the analogous of the *AIC* and the *AICc* for linear models.

In this paper, we derive the analogous to the *AICc* and the *KICc* for the Fay–Herriot model. We also propose bootstrap *AIC* and *KIC* variants. We compare through a simulation study the performance of the new criteria in relation to *AIC* and *KIC*. The rest of the paper is organized as follows. Sections 2 and 3 introduce *AIC* and *KIC* variants respectively. Section 4 presents a simulation study comparing the considered model selection criteria. Section 5 gives an application to real data where the use of the different information criteria is illustrated. The aim is the small area estimation of poverty proportions by using empirical best predictors based on Fay–Herriot models. Finally, Section 6 gives some concluding remarks.

2. The AIC variants

Let us postulate that target data $\boldsymbol{y}$ is generated by the Fay–Herriot model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} + \boldsymbol{e}, \tag{2.1}$$

where $\boldsymbol{y} = \underset{1\le d\le D}{\text{col}}(y_d)$, $\boldsymbol{u} = \underset{1\le d\le D}{\text{col}}(u_d)$, $\boldsymbol{e} = \underset{1\le d\le D}{\text{col}}(e_d)$, $\boldsymbol{X} = \underset{1\le d\le D}{\text{col}}(\boldsymbol{x}_d)$, $\boldsymbol{x}_d = \underset{1\le i\le p}{\text{col}'}(x_{di})$, $\boldsymbol{\beta} = \underset{1\le i\le p}{\text{col}}(\beta_i)$, $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{V}_u)$ with $\boldsymbol{V}_u = \lambda \boldsymbol{I}_D$, $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{V}_e)$ with $\boldsymbol{V}_e = \underset{1\le d\le D}{\text{diag}}(\sigma_d^2)$ and known variances σ_d^2, and $\boldsymbol{u}$ and $\boldsymbol{e}$ are independent. For $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda)$, the marginal probability density function (p.d.f.) $f(\boldsymbol{y}|\boldsymbol{\theta})$ is multivariate normal with mean $\boldsymbol{X}\boldsymbol{\beta}$ and variance matrix $\boldsymbol{V}_\lambda = \boldsymbol{V}_u + \boldsymbol{V}_e$. In the context of small area estimation, Rao (2003) describes the model (2.1) as the basic area-level (type A) model and he points out the differences with the basic unit-level (type B) model. The type B model is a particular case of the general linear mixed model appearing in Kubokawa (2011), but the type A model is not. This is why the study of AIC variants in the Fay–Herriot model deserves specific attention, see e.g. Han (2013).

Let us assume that the true model has also the form (2.1), but having the parameters $\boldsymbol{\beta}_0$ and λ_0 with $\beta_i = 0$ for $i = p_0 + 1, \dots, p$ and some $0 < p_0 \le p$. Let us denote $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \lambda_0)$ and $\boldsymbol{x}_{0d} = (x_{1d}, \dots, x_{p_0 d})$. We refer to $f(\boldsymbol{y}|\boldsymbol{\theta}_0)$ as the p.d.f. of the true generating model and $f(\boldsymbol{y}|\boldsymbol{\theta})$ as the approximating p.d.f. associated to the candidate model (2.1).

The Kullback–Leibler divergence between $f(\boldsymbol{y}|\boldsymbol{\theta}_0)$ and $f(\boldsymbol{y}|\boldsymbol{\theta})$ with respect to $f(\boldsymbol{y}|\boldsymbol{\theta}_0)$,

$$I(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}_0}\left[\log \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_0)}{f(\boldsymbol{y}|\boldsymbol{\theta})}\right],$$

reflects the separation between the true p.d.f. $f(\boldsymbol{y}|\boldsymbol{\theta}_0)$ and the approximating p.d.f. $f(\boldsymbol{y}|\boldsymbol{\theta})$.

For two arbitrary parametric densities $f(\boldsymbol{y}|\boldsymbol{\theta}_1)$ and $f(\boldsymbol{y}|\boldsymbol{\theta}_2)$, Cavanaugh (1997) defined

$$d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = E_{\boldsymbol{\theta}_1}\left[-2\log f(\boldsymbol{y}|\boldsymbol{\theta}_2)\right], \tag{2.2}$$

so that

$$2I(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = d(\boldsymbol{\theta}_0, \boldsymbol{\theta}) - d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0).$$

To discriminate among various candidate models $I(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ can be substituted by $d(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ since $d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$ does not depend on $\boldsymbol{\theta}$.

Akaike (1973) noted that $-2\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is an estimator of $\boldsymbol{\theta}$, is a biased estimator of $E_{\boldsymbol{\theta}_0}[d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})]$. Further, assuming that $\hat{\boldsymbol{\theta}}$ is the maximum likelihood (ML) estimator, the bias adjustment

$$B_1(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}_0}\big[d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})\big] - E_{\boldsymbol{\theta}_0}\big[-2\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})\big]$$

can be asymptotically estimated by twice the dimension of $\hat{\boldsymbol{\theta}}$. Therefore, the expected value of

$$AIC = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + 2(p+1)$$