Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Transforming response values in small area prediction

Shonosuke Sugasawa [a,*], Tatsuya Kubokawa [b]

[a] The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa-shi, Tokyo 190-8562, Japan
[b] Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

ARTICLE INFO

Article history:
Received 15 September 2015
Received in revised form 28 March 2017
Accepted 29 March 2017
Available online xxxx

Keywords:
Dual power transformation
Empirical Bayes estimation
Fay–Herriot model
Mean squared error
Positive-valued data
Small area estimation

ABSTRACT

In real applications of small area estimation, one often encounters data with positive response values. The use of a parametric transformation for positive response values in the Fay–Herriot model is proposed for such a case. An asymptotically unbiased small area predictor is derived and a second-order unbiased estimator of the mean squared error is established using the parametric bootstrap. Through simulation studies, a finite sample performance of the proposed predictor and the MSE estimator is investigated. The methodology is also successfully applied to Japanese survey data.

1. Introduction

Sample surveys are indispensable to estimate various characteristics of a population of interest. However, reliability of estimates from sample surveys depends on sample sizes, and direct estimates from small sample surveys have large variability, which is known as a small area estimation problem. In small area estimation methodology, a model-based approach has become very popular to produce indirect and improved estimates by 'borrowing strength' from related areas. Importance and usefulness of the model-based small area estimation approach has been emphasized in the literature. For a recent comprehensive review of small area estimation, see Pfeffermann (2014) and Rao and Molina (2015).

To describe the detailed setting, we define y_i as the direct survey estimator of the area mean θ_i, noting y_i is often unstable because of small area sample sizes. For producing a reliable estimate of θ_i, the most famous and basic small area model is the Fay–Herriot (FH) (Fay and Herriot, 1979) described as

$$y_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + v_i + \varepsilon_i, \quad i = 1, \ldots, m, \tag{1.1}$$

where $v_i \sim N(0, A)$ and $\varepsilon_i \sim N(0, D_i)$ for known D_i's, and the quantity of interest is $\theta_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + v_i$. It is well known that the best predictor $\widetilde{\theta}_i$ that minimizes the mean squared error is expressed as

$$\widetilde{\theta}_i = \gamma_i y_i + (1 - \gamma_i) \boldsymbol{x}_i^t \boldsymbol{\beta},$$

which is a weighted combination of the direct estimator y_i and the synthetic part $\boldsymbol{x}_i^t \boldsymbol{\beta}$ with weight $\gamma_i = A/(A + D_i)$. The weight is a decreasing function of D_i so that the weight on synthetic part $\boldsymbol{x}_i^t \boldsymbol{\beta}$ is large when y_i is not reliable, that is, the

* Corresponding author.
E-mail addresses: sugasawa@ism.ac.jp (S. Sugasawa), tatsuya@e.u-tokyo.ac.jp (T. Kubokawa).

http://dx.doi.org/10.1016/j.csda.2017.03.017
0167-9473/

Please cite this article in press as: Sugasawa, S., Kubokawa, T., Transforming response values in small area prediction. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.03.017

sampling variance D_i is large. Since it depends on unknown parameters $\boldsymbol{\beta}$ and A, the practical form of $\widetilde{\theta}_i$ is obtained by plugging estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{A}$ into $\widetilde{\theta}_i$, namely

$$\widehat{\theta}_i = \widehat{\gamma}_i y_i + (1-\widehat{\gamma}_i)\boldsymbol{x}_i^t\widehat{\boldsymbol{\beta}} = \frac{\widehat{A}y_i + D_i\boldsymbol{x}_i^t\widehat{\boldsymbol{\beta}}}{\widehat{A}+D_i},$$

which is called the empirical best linear predictor (EBLUP).

Until now, the EBLUP and the related topics have been extensively studied in the framework of the Fay–Herriot model. Chatterjee et al. (2008) and Diao et al. (2014) proposed the empirical Bayes confidence intervals of θ_i with second-order refinement. Li and Lahiri (2010) and Yoshimori and Lahiri (2014) were concerned with the problem of estimating the variance parameter A avoiding 0 estimate. Moreover, Ghosh et al. (2008) and Sinha and Rao (2009) suggested some robust estimating methods for the Fay–Herriot model. The Fay–Herriot model and EBLUP are simple and useful methods, but the setting of the Fay–Herriot model is sometimes inadequate for analysis of real data. Therefore, several extensions of the Fay–Herriot model have been proposed. Opsomer et al. (2008) suggested a nonparametric small area model using penalized spline regression. In relation to the assumption of known D_i's, Gonzàlez-Manteiga et al. (2010) proposed a nonparametric procedure for estimating D_i, and You and Chapman (2006) and Maiti et al. (2014) considered shrinkage estimation of D_i by assuming a hierarchical structure for D_i. Ybarra and Lohr (2008) and Arima et al. (2015) were concerned with the problem of measurement error in covariate $\boldsymbol{x}_i$. Datta et al. (2011) and Molina et al. (2015) suggested procedures of preliminary testing for existence of the random effect v_i. Datta and Mandal (2015) proposed a mixture of two models; one includes a random effect and the other does not include a random effect. Although all these papers treat important problems, the response values of the data are assumed to be normally distributed.

However, we often encounter positive-valued data (e.g. income, expense), which have skewed distributions and non-linear relationships with covariates. For such a data set, the traditional Fay–Herriot model with a linear structure between response values and covariates and a normally distributed error term is not appropriate. A typical alternative approach is using the log-transformed response values as discussed in Slud and Maiti (2006), but the log-transformation is not always appropriate and it may produce inefficient and biased prediction when the log-transformation is misspecified. Thus, a natural way to solve this problem is using a parametric family of transformations which enables us to select a reasonable transformation based on data. A famous family is the Box–Cox transformation (Box and Cox, 1964) defined as

$$h_\lambda^{BC}(x) = \begin{cases} \lambda^{-1}(x^\lambda - 1) & \lambda \neq 0 \\ \log x & \lambda = 0. \end{cases}$$

However, it suffers from a truncation problem that the range of the Box–Cox transformation is not the whole real line if $\lambda \neq 0$, which leads to inconsistency of the maximum likelihood estimator of λ. Moreover, the inverse transformation cannot be defined on whole real line, so that we cannot define a back-transformed predictor in the original scale. Alternatively, Yang (2006) suggested a novel family of transformations called the dual power transformation (DPT):

$$h_\lambda(x) = \begin{cases} (2\lambda)^{-1}(x^\lambda - x^{-\lambda}) & \lambda > 0 \\ \log x & \lambda = 0, \end{cases}$$

which can be seen as the average of two Box–Cox transformations, namely $h_\lambda(x) = \{h_\lambda^{BC}(x) + h_{-\lambda}^{BC}(x)\}/2$. The main advantage of the DPT is that its range is the whole real line for all $\lambda \geq 0$ so that DPT does not suffer from the truncation problem. Sugasawa and Kubokawa (2015) proposed the Fay–Herriot model in which response variables are transformed by general parametric transformations. In this paper, we focus on the FH model with DPT transformation described as

$$h_\lambda(y_i) = \boldsymbol{x}_i^t\boldsymbol{\beta} + v_i + \varepsilon_i, \quad i = 1, \ldots, m, \tag{1.2}$$

where $v_i \sim N(0, A)$ and $\varepsilon_i \sim N(0, D_i)$ for known D_i's.

Although Sugasawa and Kubokawa (2015) derived EBLUP of $\theta_i = \boldsymbol{x}_i^t\boldsymbol{\beta} + v_i$ and the MSE estimator, the parameter of most interest in the model (1.2) is $\mu_i = h_\lambda^{-1}(\theta_i)$ rather than θ_i, where $h_\lambda^{-1}(\cdot)$ is the inverse transformation of DPT:

$$h_\lambda^{-1}(x) = \left(\lambda x + \sqrt{\lambda^2 x^2 + 1}\right)^{1/\lambda}.$$

Thus, the method developed in Sugasawa and Kubokawa (2015) is not enough for practical applications. In this paper, we focus on the prediction of μ_i with its risk evaluation. Specifically, we derive the best predictor of μ_i as the conditional expectation and the empirical best predictor by plugging the parameter estimates in the best predictor. For risk evaluation, we construct a second order unbiased MSE estimator based on the parametric bootstrap.

The paper is organized as follows. In Section 2, we derive the best predictors of μ_i as well as the maximum likelihood estimation of model parameters. A second-order unbiased estimator of the mean squared error of the small area predictor is derived based on the parametric bootstrap. In Sections 3 and 4, we show some simulation studies and empirical applications, respectively. In Section 5, we give some concluding remarks. The technical details are given in the Appendix.

Please cite this article in press as: Sugasawa, S., Kubokawa, T., Transforming response values in small area prediction. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.03.017