Journal of Multivariate Analysis 127 (2014) 36–55

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

On small area estimation under a sub-area level model

Mahmoud Torabi [a,*], J.N.K. Rao [b]

[a] Department of Community Health Sciences, University of Manitoba, MB, R3E 0W3, Canada
[b] School of Mathematics and Statistics, Carleton University, Ottawa, ON, K1S 5B6, Canada

ARTICLE INFO

Article history:
Received 29 April 2013
Available online 15 February 2014

AMS subject classifications:
62D99
62F12
62H10
62J05
62M20

Keywords:
Best linear unbiased prediction
Fay–Herriot model
Linear mixed models
Mean squared error
Variance components

ABSTRACT

We propose an extension of the well-known Fay and Herriot (1979) area level model to sub-area level. Not only this model may be used to estimate small area means by borrowing strength from related areas, but also by borrowing strength from sub-areas to obtain more efficient sub-area estimators. Model-based empirical best linear unbiased prediction (EBLUP) estimators are obtained from the BLUP estimators by replacing the model parameters by suitable estimators, using an iterative method based on weighted residual sum of squares. Second order approximations to the mean squared error (MSE) of the EBLUP estimators are obtained and then used to drive MSE estimators unbiased to second order. Results of simulation studies on the performance of the proposed estimators are also provided.

1. Introduction

Sample surveys are generally designed to provide estimates of totals and means of items of interest for large subpopulations (or domains). Such estimates are "direct" in the sense of using only the domain-specific sample data, and the domain sample sizes are large enough to support reliable direct estimates that are "design based". The associated inferences (standard errors, confidence intervals, etc.) are based on the probability distribution induced by the sampling design with the population item values held fixed. Standard text books on sampling (e.g. Cochran [2], Thompson [19], Lohr [15]) provide extensive accounts of design-based direct estimation.

In recent years, demand for reliable estimates for small domains (small areas) has greatly increased worldwide due to their growing use in formulating policies and programmes, allocation of government funds, regional planning, marketing decisions at local level and other uses. Examples of small domain estimation include poverty counts of school-age children at the county level, income for small places, monthly unemployment rates for Census Metropolitan Areas, health-related estimates for local areas and so on (Rao [17, Chapter 5]). However, due to cost and operational considerations, it is seldom possible to procure a large enough overall sample size to support direct estimates for all domains of interest. We use the term "small area" to denote any domain for which direct estimates of adequate precision cannot be produced due to small domain-specific sample size. It is often necessary to employ "indirect" estimates for small areas that can increase the "effective" domain sample size by "borrowing strength" from related areas through linking models, using census and administrative data and other auxiliary data associated with the small areas. Such small area models may be classified into two broad types:

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

http://dx.doi.org/10.1016/j.jmva.2014.02.001
0047-259X/

(i) Area-level models that relate small area direct estimates to area-specific covariates; such models are used if unit-level data are not available. (ii) Unit-level models that relate the unit values of a study variable to associated unit-level covariates with known area means and area-specific covariates. A comprehensive account of model-based small area estimation under area-level and unit-level models is given by Rao [17]; see also Jiang and Lahiri [12], Datta [4], and Jiang [11] for recent overviews.

In this paper, we study model-based estimators for sub-areas nested within areas. We introduce a sub-area level model that relates a sub-area direct estimator to sub-area specific covariates, sub-area random effect and associated area random effect. Such a model is useful if unit level auxiliary variables are not available. The proposed model is a natural extension of the well-known Fay and Herriot [6] area-level model to sub-area level. The sub-area model is used to estimate small area means by borrowing strength from related areas. In addition, it can borrow strength from sub-areas to obtain more efficient sub-area estimators. Empirical best linear unbiased prediction (EBLUP) estimators of sub-area level and area level means are obtained from the BLUP estimators [10] by estimating the model parameters using an iterative method based on weighted residual sum of squares. We obtain second order approximations to the mean squared error (MSE) of the EBLUP estimators and then use them to derive MSE estimators unbiased to second order. Our approximations to MSE and its estimator assume that the number of sampled areas is large but the number of sampled sub-areas within a sampled area can be small. Our paper extends the results of Datta et al. [5] for the area level model to the sub-area level model.

The paper is organized as follows: In Section 2, we introduce the sub-area model and derive EBLUP estimators of area and sub-area means when the variance components σ_v^2 and σ_u^2, corresponding to areas and sub-areas, are estimated iteratively based on a weighted residual sum of squares method. In Section 3, we derive second order approximations to MSE of the EBLUP estimators. In Section 4, estimation of MSEs, unbiased to second order, is studied. Simulation studies, reported in Section 5, provide results on the performance of the proposed estimators.

2. Empirical best linear unbiased prediction

In the context of linear mixed models, we propose the following linking model for the sub-area means μ_{ij}:

$$\mu_{ij} = x_{ij}'\beta + v_i + u_{ij}, \quad i = 1, \ldots, m; j = 1, \ldots, N_i, \tag{2.1}$$

where j denotes a sub-area within area i, x_{ij} is a $p \times 1$ vector of sub-area level auxiliary variables ($m > p$), β is a $p \times 1$ vector of regression parameters, $v_i \overset{i.i.d.}{\sim} N(0, \sigma_v^2)$ are area random effects, and $u_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_u^2)$ are sub-area random effects. We assume that n_i sub-areas are sampled from the N_i sub-areas in the ith area.

On the other hand, the sampling model is given by

$$y_{ij} = \mu_{ij} + e_{ij}, \tag{2.2}$$

where y_{ij} is a direct estimator of μ_{ij} with sampling error e_{ij}, and $e_{ij}|\mu_{ij} \overset{ind}{\sim} N(0, \sigma_{eij}^2)$ with known sampling variances σ_{eij}^2. Assuming no sample selection bias, the sampling model (2.2) combined with the linking model (2.1) leads to the sub-area

$$y_{ij} = x_{ij}'\beta + v_i + u_{ij} + e_{ij}, \quad i = 1, \ldots, m; j = 1, \ldots, n_i. \tag{2.3}$$

Model (2.3) accounts for the sub-area level effect u_{ij} as well as the area level effect v_i. It enables us to estimate both small area means, μ_i, and sub-area means, μ_{ij}, by borrowing strength from related areas as well as sub-areas, where μ_{ij} is given by (2.1) and $\mu_i = \sum_{j=1}^{N_i} N_{ij}\mu_{ij}/N_{i+} = \bar{X}_i'\beta + v_i + \bar{U}_i$ is the mean of area i. Here, $N_{i+} = \sum_{j=1}^{N_i} N_{ij}$, $\bar{X}_i = \sum_{j=1}^{N_i} N_{ij}x_{ij}/N_{i+}$, $\bar{U}_i = \sum_{j=1}^{N_i} N_{ij}u_{ij}/N_{i+}$ and N_{ij} is the number of units in sub-area j of area i.

Fuller and Goyeneche [7] proposed a sub-area model, similar to our model (2.3), in the context of Small Area Income and Poverty Estimation (SAIPE) in the United States. In this application, county is the sub-area nested within a state (area) and direct county estimates obtained from the Current Population Survey (CPS) data. County-level auxiliary variables are ascertained from census and administrative records.

In matrix notation, the model (2.3) can be written as

$$\mathrm{y}_i = X_i\beta + v_i 1_{n_i} + u_i + e_i, \quad i = 1, \ldots, m,$$

where $\mathrm{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$ is a $n_i \times 1$ vector, X_i is a $n_i \times p$ matrix with rows x_{ij}', $(j = 1, \ldots, n_i)$, $u_i = (u_{i1}, \ldots, u_{in_i})'$ and $e_i = (e_{i1}, \ldots, e_{in_i})'$. Equivalently, we have

$$\mathrm{y}_i = X_i\beta + Z_i b_i + e_i, \quad i = 1, \ldots, m, \tag{2.4}$$

where $Z_i = (1_{n_i}|I_{n_i})$ with 1_{n_i} as the vector of ones and I_{n_i} as the identity matrix with dimension n_i, and $b_i = (v_i, u_i')'$. Model (2.4) is a linear mixed model with a block diagonal covariance structure with blocks $\mathrm{cov}(\mathrm{y}_i) = V_i$ with

$$V_i = \sigma_v^2 J_{n_i} + \mathrm{diag}(\sigma_u^2 + \sigma_{ei1}^2, \ldots, \sigma_u^2 + \sigma_{ein_i}^2), \tag{2.5}$$

where $J_{n_i} = 1_{n_i}1_{n_i}'$.