2011

Contextual effects in modeling for small domain estimation

Mohammad-Reza Namazi-Rad
University of Wollongong, mrad@uow.edu.au

David G. Steel
University of Wollongong, dsteel@uow.edu.au

Contextual effects in modeling for small domain estimation

Abstract

Many different Small Area Estimation (SAE) methods have been proposed to overcome the challenge of findingreliable estimates for small domains. Often, the required data for various research purposes are available at differentlevels of aggregation. Based on the available data, individual-level or aggregated-level models are used in SAE.However, parameter estimates obtained from individual and aggregated level analysis may be different, in practice.This may happen due to some substantial contextual or area-level effects in the covariates which may be misspecifiedin individual-level analysis. If small area models are going to be interpretable in practice, possible contextualeffects should be included. Ignoring these effects leads to misleading results. In this paper, synthetic estimators andEmpirical Best Linear Unbiased Predictors (EBLUPs) are evaluated in SAE based on different levels of linear mixedmodels. Using a numerical simulation study, the key role of contextual effects is examined for model selection inSAE.Key words: Contextual Effect; EBLUP; Small Area Estimation; Synthetic Estimator.

Keywords

Contextual, effects, modeling, for, small, domain, estimation

Disciplines

Engineering | Physical Sciences and Mathematics

Publication Details

Namazi-Rad, M. & Steel, D. (2011). Contextual effects in modeling for small domain estimation. In E. Beh, L. Park & K. Russell (Eds.), Proceedings of the 4th Applied Statistics Education and Research Collaboration (ASEARC) Conference (pp. 12-14). Wollongong: University of Wollongong.

Contextual Effects in Modeling for Small Domain Estimation

Mohammad-Reza Namazi-Rad

Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522, Australia
mohammad_ namazi@uow.edu.au

David Steel

Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522, Australia
david_ steel@uow.edu.au

Abstract

Many different Small Area Estimation (SAE) methods have been proposed to overcome the challenge of finding reliable estimates for small domains. Often, the required data for various research purposes are available at different levels of aggregation. Based on the available data, individual-level or aggregated-level models are used in SAE. However, parameter estimates obtained from individual and aggregated level analysis may be different, in practice. This may happen due to some substantial contextual or area-level effects in the covariates which may be misspecified in individual-level analysis. If small area models are going to be interpretable in practice, possible contextual effects should be included. Ignoring these effects leads to misleading results. In this paper, synthetic estimators and Empirical Best Linear Unbiased Predictors (EBLUPs) are evaluated in SAE based on different levels of linear mixed models. Using a numerical simulation study, the key role of contextual effects is examined for model selection in SAE.

Key words: Contextual Effect; EBLUP; Small Area Estimation; Synthetic Estimator.

1. Introduction

Sample surveys allow efficient inference about a national population when resources do not permit collecting relevant information from every member of the population. As a consequence, many sample surveys are conducted each year around the world to obtain statistical information required for policy making. In recent years, there is increasing need for statistical information at sub-national levels. Such statistics are often referred to as small area statistics, and methods for obtaining them from national surveys have become an important research topic, stimulated by demands from government agencies and businesses for data at different geographic and socio-demographic levels. In this context, Small Area Estimation (SAE) refers to statistical techniques for producing reliable estimates for geographic sub-populations (such as city, province or state) and socio-demographic sub-domains (such as age group, gender group, race group etc.) where the survey data available are insufficient for calculation of reliable direct estimates.

A fundamental property of SAE methods is that they combine related auxiliary variables with statistical models to define estimators for small area characteristics. Most statistical models used in SAE can be formulated either at the unit level or at the area level. When sample data are available at the individual or unit level, a common approach in SAE is to compute parameter estimates based on a unit-level mixed linear model. However, it is also often possible to fit this type of model at the area level, and to then compute the small area estimates based on this area-level mixed model.

In this paper we explore the relative performance of small area estimates based on area-level models with the same estimates based on unit-level models given both individual and aggregate (i.e. area level) data are available. We assume that the targets of inference are the area-level population means of a variable. A unit-level analysis is thus at a different level from which the final estimates will be calculated.

Our aim is to identify situations where aggregated-level analysis can provide more reliable estimates than unit-level analysis. This may happen due to the presence of contextual or area-level effects in the small area distribution of the target variable. Ignoring these effects in unit-level models can lead to biased estimates. However, such area-level effects are automatically included in area-level models in certain cases.

In this paper, matrices are displayed in bold type. Sample statistics are denoted by lowercase letters, with uppercase used for corresponding population statistics.

2. Area-level Approach

Fay and Herriot (1979) applied a linear regression with area random effects with unequal variances for predicting the mean value per capita income (PCI) in small geographical areas.

Throughout this paper we shall assume the target population divided into K sub-domains. In such a case, Fay-Herriot model is:

$$\hat{\bar{Y}}_k^{D} = \bar{Y}_k + \varepsilon_k \ ; \ k = 1, \ldots, K \tag{1}$$

where $\bar{Y}_k$ is the true population value for kth area mean for the target variable, $\hat{\bar{Y}}_k^{D}$ denotes its direct estimate and $\varepsilon_k | \bar{Y}_k \sim N(0, \sigma^2_{\varepsilon_k})$. $\bar{Y}_k$ is assumed in (1) to be related with P auxiliary variables as follows:

$$\bar{Y}_k = \bar{\mathbf{X}}'_k \beta + u_k \ ; \ \textit{where} \quad u_k \sim N(0, \sigma^2_u) \tag{2}$$

where $\bar{\mathbf{X}}_k$ is the vector of kth area population means for P auxiliary variables.

$$\bar{\mathbf{X}}'_k = [1 \ \ \bar{X}_{k1} \ \ \bar{X}_{k2} \ \ \ldots \ \ \bar{X}_{kP}]$$

Variance of the error term $(\sigma^2_{\varepsilon_k})$ is typically assumed to be associated with the complex sampling error for kth area and it is assumed to be known in (1). This strong assumption seems unrealistic in practice (Gonz*á*lez-Manteiga, *et. al.* (2010)). The implications of having to estimate variance components and the effectiveness of the aggregated-level approach in SAE is considered in following sections.

3. Unit-level Approach

A standard Linear Mixed Model (LMM) for individual-level population data is:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{3}$$

Supposing N to be the population size, $\mathbf{Y}$ is a column vector of N random variables, $\mathbf{X}$ is an $N \times (P+1)$ matrix of known quantities whose rows correspond to the statistical units, and β is a vector of $(P+1)$ parameters. $\mathbf{Z}$ is a $N \times K$ matrix of random-effect regressors, and finally, $\mathbf{u}$ and $\mathbf{e}$ are respectively $K \times 1$ and $N \times 1$ vectors of different random effects. Note that, $\mathbf{u}$ and $\mathbf{e}$ are assumed to be distributed independently with mean zero and covariance matrices $\mathbf{G}$ and $\mathbf{R}$, respectively.

$$E(\mathbf{u}) = \mathbf{0} \quad \& \quad E(\mathbf{e}) = \mathbf{0} \ ; \ Var\begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}$$

The variance-covariance matrix for $\mathbf{Y}$ is:

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}.$$

Under the general definition of LMM, Datta and Lahiri (2000) defined the target of inference as:

$$\mu_{\bar{Y}_k} = \bar{\mathbf{X}}'_k \beta + u_k$$

The BLUP for $\mu_{\bar{Y}_k}$ is:

$$\tilde{\mu}_{\bar{Y}_k} = \bar{\mathbf{X}}'_k \tilde{\beta} + \mathbf{l}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\tilde{\beta}) \tag{4}$$

where

$$\begin{aligned} \tilde{\beta} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \\ \mathbf{l}' &= (\underbrace{0, 0, \ldots, 0, 1}_{k}, 0, ..., 0) \end{aligned} \tag{5}$$

To calculate BLUP value for $\mu_{\bar{Y}_k}$ in (4), variance components are assumed to be known. Replacing the estimated values for the variance components in (4) and (5), a new estimator will be obtained. This estimator is presented by Harville (1991) as an "empirical BLUP" or EBLUP.

The mean value for the target variable within the kth area can be estimated based on the fitted working model through the synthetic estimation technique as:

$$\hat{\bar{Y}}_k^{Syn} = \bar{\mathbf{X}}'_k \tilde{\beta}$$

A similar approach can be used to calculate parameter estimates and consequent synthetic estimators and EBLUPs under Fay-Herriot model [Longford (2005)].

If individual-level data are available, small area estimation is usually based on models formulated at the unit level but they are ultimately used to produce estimates at the area level. Using aggregated-level analysis may cause loss of efficiency when the data is available at the individual level. When the data comes from a complex sample, it may not be very straightforward to calculate small area estimates. Therefore, a common approach is to use area-level estimates that account for the complex sampling and regression model of a form introduced in (1).

4. Contextual Models

Linear mixed models such as (3) are commonly used in SAE. However, area-level covariates can also be included in the unit-level models in order to improve the efficiency in certain cases. Supposing $\mathcal{T}_k$ to denote the vector of kth area-level covariates being included in unit-level population model, we have:

$$\begin{aligned} Y_{ik} &= (\mathbf{X}'_{ik}; \mathcal{T}'_k)\beta + u_k + e_{ik} \\ &\qquad\qquad i = 1, \ldots, N_k \ \ \& \ \ k = 1, \ldots, K \\ u_k &\sim N(0, \sigma^2_u) \ ; \ e_{ik} \sim N(0, \sigma^2_e) \end{aligned} \tag{6}$$

where $\mathbf{X}'_{ik} = [1 \ \ X_{ik1} \ \ X_{ik2} \ \ \ldots \ \ X_{ikP}]$. N_k denotes the population size for kth area.

In the statistical literature, area-level covariates such as those in (6) are sometimes referred to as 'contextual effects' and model (6) is then described as a 'contextual model'. A special case of $\mathcal{T}_k$ is where the contextual effects are small area population means. Then we have:

$$\begin{aligned} Bias_{\xi}\left(\hat{\bar{Y}}_k^{Syn}\right) &= \bar{\mathbf{X}}'_k Bias_{\xi}(\tilde{\beta}) \\ Bias_{\xi}\left(\hat{\bar{Y}}_k^{EBLUP}\right) &\approx (1 - \gamma_k)\bar{\mathbf{X}}'_k Bias_{\xi}(\tilde{\beta}) \end{aligned} \tag{7}$$

where:

$$\gamma_k = \frac{\sigma_u^2}{\sigma_u^2 + \psi_k} \quad ; \quad \psi_k = Var(\bar{e}_k|\bar{Y}_k)$$

Note that, the subscript ξ denotes the bias under the assumed population model (6).

It is often the case in practice that unit-specific and area-specific coefficient estimates would have different expectations. This may happen as a result of area-level miss-specifications in individual-level analysis and can cause an error in the interpretation of statistical data.

5. Monte-Carlo Simulation

A model-assisted design-based simulation study is presented in this section to assess the empirical Mean Square Error (MSE) of resulting synthetic estimators and EBLUPs based on individual-level and aggregated-level analysis. To develop the numerical study, the gross weekly income is considered as the target variable. Available data on the length of education and training experience for different individuals aged 15 and over is then used as auxiliary information. The target variable is assumed to be related with the auxiliary variable through a linear mixed model.

Available information for 57 statistical sub-divisions in Australia about the mentioned characteristics and area population sizes is used in this study. Area means are also included in the individual-level population model for generating population data in this monte-carlo simulation.

Considering the actual area means to be the target of inference, synthetic estimates and EBLUPs are then calculated based on two working models fitted on the sample data as follows:

$$\begin{aligned} y_{ik}^{(W_1)} &= (1; x_{ik})\beta + u_k + e_{ik} \\ u_k &\sim N(0, \sigma_u^2) \;\; ; \;\; e_{ik} \sim N(0, \sigma_e^2) \\ & \qquad i = 1, \ldots, n_k \;\; \& \;\; k = 1, \ldots, K \\ \bar{y}_k^{(W_2)} &= (1; \bar{x}_k)\beta + u_k + \bar{e}_k \\ \bar{\mathbf{e}} &\sim N\Big(\mathbf{0}\ ,\ diag(\tfrac{\sigma_e^2}{n_1},\ \ldots,\ \tfrac{\sigma_e^2}{n_K})\Big) \end{aligned} \tag{8}$$

where n_k is the sample size allocated to kth area. The first working model *(W1)* can be fitted on individual-level sample data while the second working model *(W2)* uses aggregated-level sample data for estimation purposes.

This allows a comparison to be made among the performance of the models in (8) in case of having actual area means as possible contextual effects in population model. In this study, the model parameters β, σ_e^2 and σ_u^2 are empirically estimated in both unit-level and area-level models by Fisher scoring algorithm as a general method for finding maximum likelihood parameter estimates (Longford (2005)).

Figure (1) summarizes the results by giving the ratio of the MSEs for the SAEs based on unit-level and area-level models for $K = 57$ areas in the simulation.

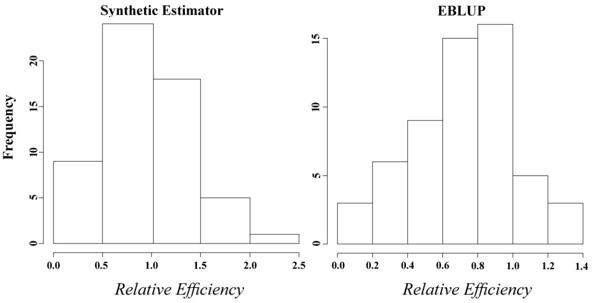

Figure 1: The Relative Efficiency of Unit-level to Area-level Model

A value less than 1 for the relative efficiency in figure (1) indicates that the unit-level approach based on (*W1*) is more precise comparing with the area-level approach based on (*W2*) in terms of MSE in each case. Using synthetic approach, it is difficult to say which model helps to obtain more precise estimates. The ratio varies below and above 1 for the synthetic estimation, while this value is generally below 1 for the EBLUP. This can be due to the effect of the shrinkage factor used in EBLUP technique (see (7)).

6. Conclusion

Individual-level analysis usually results in more stable small area estimates. However, if the unit-level working model is misspecified by exclusion of important auxiliary variables, parameter estimates obtained from the individual and aggregated level analysis will have different expectations.

In particular, if an existing contextual variable is ignored, the parameter estimates calculated from an individual-level analysis will be biased, whereas an aggregated-level analysis can lead to small area estimates with less bias. Even if contextual variables are included in an unit-level modeling, there may be an increase in the variance of parameter estimates due to increased number of variables in the working model.

References

[1] Datta, G. S., and Lahiri, P. (2000). A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems. *Statistica Sinica*. **10**, 613-627.

[2] Fay, R. E., and Herriot, R. A. (1979). Estimates of Income for Small Places: an Application of James-Stein Procedures to Census Data. *Journal of The American Statistical Association.* **74**, 269-277.

[3] Gonzάlez-Manteiga, W., Lombardίa, M. J., Molina, I., Morales, D., and Santamarίa, L. (2010). Small Area Estimation under FayHerriot Models with Non-parametric Estimation of Heteroscedasticity. *Statistical Modelling*. **10**, 215-239.

[4] Harville D. A. (1991). That BLUP is a Good Thing: The Estimation of Random Effects, (Comment). *Statistical Science*. **6**, 35-39.

[5] Longford N. T. (2005). *Missing Data and Small Area Estimation*. Spring-Verlag.