2011

Contextual Effects in Modeling for Small Domains

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

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

Publication Details

Namazi-Rad, Mohammad-Reza; and Steel, David, Contextual Effects in Modeling for Small DomainsProceedings of the Fourth Annual ASEARC Conference, 17-18 February 2011, University of Western Sydney, Paramatta, Australia.

Contextual Effects in Modeling for Small Domains

Abstract
During last two decades, different Small Area Estimation (SAE) methods have been proposed to overcome the challenge of finding reliable small area estimates. This happens a lot that the required data for various research purposes are available at different levels. Based on availability of data, individual-level or aggregated-level models are implied in SAE. However, the estimated values for model parameters obtained from individual-level analysis can be different from the one obtained based on analysis of aggregate data. Generally, this is referred to as the ecological fallacy. This happens due to some substantial contextual or area-level effects in the covariates. To have a good interpretation of available data, possible contextual effects must be carefully included, measured, and accounted for in statistical models for calculating reliable estimates. Ignoring these effects leads to misleading results. The main advantage of contextual models is to help statisticians in studying aggregated-level data without concerning about the issue of ecological fallacy. In this paper, synthetic estimators and Empirical Best Linear Unbiased Predictors (EBLUPs) are studied based on different levels of linear mixed models. Using a numerical simulation study, the key role of contextual area-level effects is examined for model selection in SAE.

Keywords
Contextual Effect; EBLUP; Ecological Fallacy; Small Area Estimation; Synthetic Estimator

Publication Details
Namazi-Rad, Mohammad-Reza; and Steel, David, Contextual Effects in Modeling for Small DomainsProceedings of the Fourth Annual ASEARC Conference, 17-18 February 2011, University of Western Sydney, Paramatta, Australia.

Contextual Effects in Modeling for Small Domains

Mohammad-Reza Namazi-Rad and David Steel

Centre for Statistical and Survey Methodology
University of Wollongong, NSW 2522, Australia

February 2011

Abstract

During last two decades, different Small Area Estimation (SAE) methods have been proposed to overcome the challenge of finding reliable small area estimates. This happens a lot that the required data for various research purposes are available at different levels. Based on availability of data, individual-level or aggregated-level models are implied in SAE. However, the estimated values for model parameters obtained from individual-level analysis can be different from the one obtained based on analysis of aggregate data. Generally, this is referred to as the ecological fallacy. This happens due to some substantial contextual or area-level effects in the covariates. To have a good interpretation of available data, possible contextual effects must be carefully included, measured, and accounted for in statistical models for calculating reliable estimates. Ignoring these effects leads to misleading results. The main advantage of contextual models is to help statisticians in studying aggregated-level data without concerning about the issue of ecological fallacy. In this paper, synthetic estimators and Empirical Best Linear Unbiased Predictors (EBLUPs) are studied based on different levels of linear mixed models. Using a numerical simulation study, the key role of contextual area-level effects is examined for model selection in SAE.

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

1. Introduction

Sample surveys allow efficient estimation and other forms of inference about a large population when the resources available do not permit collecting relevant information from every member of the population. Each year, sample surveys are conducted in the world to obtain statistical information required for various decisions and policy making. The demand has grown markedly in recent years for comprehensive statistical information not only at the national levels but also for sub-national domains.

Working on different types of small area statistics have become an important research topic in survey methods in the last few decades, stimulated by increasing demands in government agencies and various advertising, marketing and business sectors for data at different geographic and socio-demographic levels. Small Area Estimation (SAE) involves statistical techniques producing a number of estimates for geographic sub-population (such as city, province, state or country etc.) and socio-demographic sub-domains (such as age group, gender group, race group etc.) in which available survey data is not enough to calculate reliable estimates. Usually, related auxiliary variables are used in statistical models to find required estimates in different small area estimation techniques [5].

Statistical models in SAE can be formulated at the unit level or area level. Unit-level models use available data for different individuals while area-level models work with available information at the area level and use aggregate data for estimation purposes. Area-level models are useful when available data is accessible just at the area levels. The area-level model can be also derived using aggregating (averaging) techniques on the individual data. In this paper, assuming the target of inference to be at the area level, the performance of area-level models is explored comparing with unit-level models when both individual and aggregate data are available.

The main purpose is to find situations in which directly aggregated-level analysis can provide more reliable estimates. This can happen due to substantial contextual or area-level effects in the covariates. Ignoring these effects in unit-level working models can cause biased estimates which is referred to as the ecological fallacy. However, these area-level effects can be automatically covered in area-level models in especial cases.

2. FayHerriot model

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 is not very straightforward to find likelihood for unit level sample data from complex designs. Therefore, a common approach is to use area-level estimates that account for the complex sampling and regression models of a form introduced by Fay and Herriot (1979).

Fay and Herriot (1979) applied a linear regression with area random effects in the context of unequal variances for predicting the mean value per capita income (PCI) in small geographical areas [4].

Considering the population divided into K subdomains, Fay-Herriot model is presented as:

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

where $\varepsilon_k|\bar{Y}_k \sim N(0, \sigma^2_{\varepsilon_k})$. In Fay-Herriot model, it is also assumed that the true mean is correlated with P auxiliary variable through a linear model.

$$\bar{Y}_k = (1; \underline{\bar{X}}_k')\underline{\beta} + u_k \;\; ; \;\; where \;\; u_k \sim N(0, \sigma_u^2) \tag{2}$$

where $\underline{\bar{X}}_k$ is the vector of mean values of P auxiliary variables within kth area. Variance of the fixed error term (ε_k) is typically assumed to account for the complex sampling error for kth area and σ_{ε_k} is considered be known in the Fay-Herriot model. This strong assumption seems unrealistic in practice [3]. Usually, it is useful to use underlying unit-level models to obtain more realistic parameter estimates. In this way, the model parameters will be estimated using the individual-level data, firstly. Then, the unit-level estimates will be used to estimate the variable of interest at the required area-level by the aggregating the data. The implications of having to estimate the sampling variance and the effectiveness of a unit-level approach is considered in following sections.

3. EBLUP Techniques

A straightforward definition of general Linear Mixed Models (LMM) with P auxiliary variables is given as:

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

where $\underline{Y}$ is an $N \times 1$ column vector of random variables, $\mathbf{X}$ is an $N \times P$ matrix of known quantities whose rows correspond to the statistical units, and $\underline{\beta}$ is a $P \times 1$ vector of parameters. $\mathbf{Z}$ is an $N \times q$ matrix of random-effect regressors, and finally, $\underline{u}$ and $\underline{e}$ are respectively

$q \times 1$ and $n \times 1$ random and fixed effects vectors. Note that, $\underline{u}$ and $\underline{e}$ are assumed to be distributed independently with mean zero and covariance matrices $\mathbf{G}$ and $\mathbf{R}$, respectively.

$$Var\begin{pmatrix} \underline{u} \\ \underline{e} \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix} \quad , \quad E(\underline{e}) = \underline{0} \ \ \& \ \ E(\underline{u}) = \underline{0} \quad (4)$$

The mean vector and covariance matrix for $\underline{Y}$ are respectively, $\underline{\mu}_Y = \mathbf{X}\underline{\beta}$ and $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$.

Under the general definition of linear mixed model, a linear combination of the fixed and random effects' prediction is discussed by Datta and Lahiri (2000) as:

$$\theta = \underline{b}'\underline{\beta} + \underline{l}'\underline{u} \quad (5)$$

where the elements $\underline{b}$ and $\underline{l}$ are defined as below:

$$\underline{b}' = (1; \underline{\bar{X}}_k') \qquad \& \qquad \underline{l}' = (\underbrace{0, 0, \ldots, 0, 1}_{k}, 0, ..., 0)$$

then in this especial case, the mentioned linear combination is presented as:

$$T(\theta, \underline{\bar{Y}}) = \bar{\mathbf{X}}_k'\underline{\beta} + u_k \quad (6)$$

and the BLUP (or BLUE) for this combination is: [Henderson (1975)]

$$\hat{T}(\theta, \underline{\bar{Y}}) = \bar{\mathbf{X}}_k'\underline{\tilde{\beta}} + \underline{l}'\mathbf{GZ}'\mathbf{V}^{-1}(\underline{Y} - \mathbf{X}\underline{\tilde{\beta}}) \quad (7)$$

To calculate BLUP value for $T(\theta, \underline{Y})$ in above equation, variance components have been assumed to be know. Replacing the estimated values for the variance components in the mentioned equation, a two-stage estimator will be obtained. This estimator in presented in statistical literature as an "empirical BLUP" or EBLUP.

4. Contextual Models

It is common to derive the mixed models at the individual levels, but sometimes some covariates may be available in the model which can improve the efficiency in the final conclusions. Suppose $\underline{\mathcal{T}}_k$ denotes the area-level covariate which is added to the general linear mixed model. Then, the linear population model can be presented as below:

$$Y_{ik} = (1; \underline{X}_{ik}'; \underline{\mathcal{T}}_k')\underline{\beta}^* + u_k^* + e_{ik}^* \qquad i = 1, \ldots, N_k \ \ \& \ \ k = 1, \ldots, K \quad (8)$$
$$u_k^* \sim N(0, \sigma_{u^*}^2) \ \ ; \ \ e_{ik}^* \sim N(0, \sigma_{e^*}^2)$$

In statistical literatures, the mentioned area-level covariate is discussed as a 'contextual effect' and the model above is mentioned as a 'contextual model'. As it can be seen in the model above, both individual and aggregate data are involved in a contextual model, simultaneously. This is the main advantage of using contextual models which helps statisticians to use aggregate data in modeling without concerning about the issue of ecological fallacy. Ecological fallacy, which is often called 'ecological inference fallacy' occurs when researchers want to draw a conclusion about an individual-level inference based on aggregated-level data analysis. This causes an error in the interpretation of statistical data as the results based on purely aggregated-level analysis may not be true for describing the inference about an individual-based characteristic. This is referred to as an ecological fallacy [6].

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 synthetic and EBLUP based on individual-level and aggregated-level analysis. To develop the numerical study, a linear relationship is considered for the weekly income in Australia as the required variable. The length of education and training experience for different individuals aged 15 and over is also considered as the auxiliary variable. Note that, there are 6 states and 3 mainland territories in Australia and each is divided into some statistical sub-divisions. Totally, there are 57 statistical sub-divisions which are being used in different survey designs in Australian Bureau of Statistics (ABS).

In this monte-carlo simulation, available information in ABS web-site is used in order to simulate the population based a contextual model as below:

$$\begin{aligned} Y_{ik} &= (1; X_{ik}; X_k)\underline{\beta}^* + u_k^* + e_{ik}^* \\ u_k^* &\sim N(0, \sigma_{u^*}^2) \;\; ; \;\; e_{ik}^* \sim N(0, \sigma_{e^*}^2) \\ & i = 1, \ldots, N_k \;\; \& \;\; k = 1, \ldots, K \end{aligned} \tag{9}$$

Synthetic estimates and EBLUPs are then calculated based on two working models fitted on the sample data presented as:

$$\begin{aligned} y_{ik}^{(W_1)} &= (1; x_{ik})\underline{\beta} + u_k + e_{ik} \\ u_k &\sim N(0, \sigma_u^2) \;\; ; \;\; e_{ik} \sim N(0, \sigma_e^2) \\ & i = 1, \ldots, n_k \;\; \& \;\; k = 1, \ldots, K \\ \bar{y}_k^{(W_2)} &= (1; \bar{x}_k)\underline{\beta} + u_k + \bar{e}_k \\ \underline{\bar{e}} &\sim N\Big(\underline{0}\,,\; diag(\tfrac{\sigma_e^2}{n_1},\; \ldots,\; \tfrac{\sigma_e^2}{n_K})\Big) \end{aligned} \tag{10}$$

This allows a comparison to be made among unit-level and area-level working models which can be fitted on the sample data in order to predict values for the required variable in the population for each case.

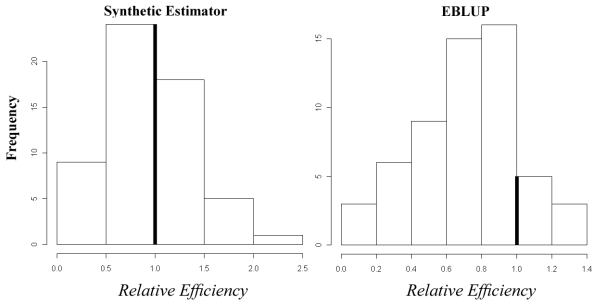

Figure 1: The Relative Efficiency of Unit-level to Area-level Model

Figure (1) summarizes the results by giving the ratio of the MSEs for the SAEs based on unit-level and area-level model for $K = 57$ regressors in the simulation. In the simulation, the parameter estimates for both working models are estimated using Fisher scoring method. 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.

6. Conclusion

Usually, choosing unit-level analysis helps to produce better 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 important contextual variable is omitted, the parameter estimates obtained from an individual-level analysis will be biased, whereas an aggregated-level analysis can produce unbiased estimates. Even if contextual variables are included in an individual-level model analysis, 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] González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., and Santamaría, L. (2010). Small Area Estimation under FayHerriot Models with Non-parametric Estimation of Heteroscedasticity. *Statistical Modelling*. **10**, 215239.

[4] Pfeffermann, D. (2002). Small Area Estimation-New Developments and Directions. *International Statistical Review*. **70**, 125-143.

[5] Rao, J. N. K. (2003). *Small Area Estimation*. Wiley; New York.

[6] Seiler, F. A., and Alvarez, J. L. (2000). Is the Ecological Fallacy a Fallacy? *Human and Ecological Risk Assessment*. **6**, 921-941.