STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 35–39, DOI 10.21307/stattrans-2020-025
Received – 31.01.2020; accepted – 30.06.2020

Discussion of "Small area estimation: its evolution in five decades", by Malay Ghosh

Yan Li[1]

Prof. Ghosh leads us step gradually into the realm of small area estimation (SAE) through the evolution of SAE for the past five decades, introducing various SAE methods of synthetic estimators, composite estimators, and model-based estimators for small area parameters, mean squared error approximations, adjustment methods of benchmarking and transformation, etc. The paper broadens and deepens our understanding of different perspectives of the SAE and provides a few illustrative real-life applications. It is a great review paper for general audience, especially for our graduate students in survey statistics and related areas, who wish to have a snapshot of the SAE research.

Prof. Ghosh focuses his review on the inferential aspects of the two celebrated small area models ----- the Fay-Herriot (FH) area model and the unit level nested error regression (NER) model. In the implementation of these models, variable selection plays a vital role and my discussion centers around this topic, which complements Professor Ghosh's paper.

There is a vast literature on variable selection, a subtopic of model selection. We refer to the Institute of Mathematical Statistics Monograph edited by Lahiri (2001) for different approaches and issues in model selection and the book by Jiang and Nguyen (2015) for model selection methodology especially designed for mixed models. Variable selection methods for general linear mixed model can be, of course, applied to select variables for the FH and NER models as they are special cases of the general linear mixed model. Many data analysts not familiar with mixed models, however, use software meant for linear regression models to select variables. This approach may result in loss of efficiency in variable selection. Lahiri and Suntornchost (2015) and Li and Lahiri (2019) proposed simple adjustment methods so that the data users can select reasonable models by calculating their favorite variable selection criteria, such as AIC, BIC, Mallow's C_p, and adjusted R^2, which are developed for standard linear regression model assuming independent identically distributed (*iid*) errors. The goal of the two

[1] Joint Program in Survey Methodology and Department of Epidemiology and Biostatistics, University of Maryland, College Park, USA. E-mail: yli6@umd.edu. ORCID: https://orcid.org/0000-0001-8241-7464.

papers is to propose adjustment methods, instead of advocating a specific variable selection method. Cai et al. (2020), with the same goal, creatively combined the two variable selection methods (Lahiri and Suntornchost, 2015 and Li and Lahiri, 2019) and proposed a variable selection method for another popular two-fold subarea model.

The above-mentioned three methods consider commonly used variable selection criteria under a standard regression model with *iid* errors, including

1) Adjusted R^2: adjRsq $= 1 - \frac{MSE_k}{MST}$,
2) Mallows C_P: $C_P = \frac{SSE_k}{MSE_K} + 2k - n$,
3) AIC: $AIC = 2k + n \cdot \log\left(\frac{SSE_k}{n}\right)$, and
4) BIC: $BIC = k \cdot \log(n) + n \cdot \log\left(\frac{SSE_k}{n}\right)$,

where

$$MSE_k = \frac{SSE_k}{n-k} \text{ with}$$
$$SSE_k = y^T[I - X_k(X_k^T X_k)^{-1}X_k^T]y, \text{ and}$$
$$MST = \frac{SST}{n-1} \text{ with}$$
$$SST = y^T[I - n^{-1}11^T]y.$$

Note that $y = (y_1, \dots, y_n)$ is a vector of observations on the dependent variable; X_k is a $n \times (1+k)$ design matrix with columns of one's and k auxiliary variables, corresponding to the intercept and k unknown parameters; $SSE_k(MSE_k)$ is the SSE (MSE) based on the standard regression model for $k = 1, \dots, K$. Here K is the total number of auxiliary variables considered in model selection and n is the sample size. When $k = K$, $MSE_K = \frac{SSE_K}{n-K}$ is the MSE based on the full model with all K auxiliary variables. As noted, these variable selection criteria can be expressed as a smooth function of MSE_k and MST.

Next, adjustments proposed for the three small area models are briefly discussed before above variable selection criteria designed for standard regression model can be used.

1. Consider the Fay-Herriot area model given by:

$$y_i = \theta_i + e_i \text{ and } \theta_i = x_i^T\beta + v_i, \quad (1)$$

where θ_i is the unobserved true mean for small area i; y_i is the survey-weighted estimate of θ_i; $v_i$is the random effect for small area i; v_i's and e_i's are independent with $v_i \sim N(0, A)$ and $e_i \sim N(0, D_i)$ $i = 1, \dots, m$. Let $\epsilon_i = v_i + e_i$, and its variance is $A + D_i$. The vector $\beta = (\beta_0, \beta_1, \dots, \beta_k)^T$ is a vector of length k+1 of unknown parameters.

Lahiri and Suntornchost (2015) proposed a simple adjustment to the standard variable selection methods by replacing MSE_k and MST in above variable selection criteria by

$$\widehat{MSE}_k = MSE_k - \bar{D}_w$$

and

$$\widehat{MST} = MST - \bar{D},$$

where $\bar{D}_w = \frac{\sum_{i=1}^{m}(1-h_{ii})D_i}{m-k}$, $h_{ii} = x_i^T(X^TX)^{-1}x_i$, and $\bar{D} = m^{-1}\sum_{i=1}^{m} D_i$. The new variable selection criteria track the corresponding true variable selection criteria much better than naïve methods. Lahiri and Suntornchost (2015) also proposed a transformation method and a truncation method to prevent negative values of $\widehat{MSE}_k$ and $\widehat{MST}$. As noted, the Lahiri-Suntornchost method can be implemented using two simple steps: 1) adjusting MSE_k and *MST*, and 2) computing the variable selection criteria of users' choice under the standard regression model with adjusted $\widehat{MSE}_k$ and $\widehat{MST}$.

2. Consider a unit level nested error regression model given by:

$$y_{ij} = x_{ij}^T\beta + v_i + e_{ij} \tag{2}$$

for unit $j = 1, \ldots, n_i$ in area $i = 1, \ldots, m$, where n_i is the sample size for small area *i* and the total sample size $n = \sum_{i=1}^{m} n_i$. In Model (2), we assume the area effect v_i ~ iid $N(0, \sigma_v^2)$ is independent of e_{ij} ~ iid $N(0, \sigma_e^2)$. Define $\sigma^2 = \sigma_e^2 + \sigma_v^2$. The outcome in unit *j* of area *i* is denoted by y_{ij}, and $x_{ij} = \left(1, x_{ij1}, x_{ij2}, \ldots, x_{ijk}\right)$ is a vector of length *k*+1 with the values of the covariates $x_1, x_2, \ldots x_k$ for unit *j* in area *i*. In order to make the observations independent and at the same time to avoid the estimation of the intra-cluster correlation, Li and Lahiri (2019) specified P_i to be an $(n_i - 1) \times n_i$ matrix such that $\begin{pmatrix} n_i^{-\frac{1}{2}} 1_{n_i}^T \\ P_i \end{pmatrix}$ is orthogonal for $i = 1,2, \ldots, m$, and transformed the data by

$$\begin{gathered} y_i^{LL} = P_i y_i, \\ x_i^{LL} = P_i x_i, \text{and} \\ u_i^{LL} = P_i u_i. \end{gathered}$$

The transformed model can then be written as:

$$y_i^{LL} = x_i^{LL}\beta + u_i^{LL} \text{ for } i = 1, 2, \ldots, m, \tag{3}$$

where the vector of the error term in area *i* follows $u_i^{LL} \sim N(0, \sigma^2(1-\rho)I_{n_i-1})$ with I_{n_i-1} a $(n_i - 1) \times (n_i - 1)$ identity matrix. The MSE_k and MST estimated from Model (3) can then be plugged into the various variable selection criteria, from which users can pick their favorite to select model variables. Same as the Lahiri-Suntornchost

method, the Li-Lahiri (LL) method is implemented with two steps, but with a different first step: estimating MSE_k and MST by fitting the LL-transformed data to Model (3): a standard regression model with *iid* error.

3. Consider two-fold subarea model given by:

$$y_{ij} = \theta_{ij} + e_{ij} \text{ and } \theta_{ij} = x_{ij}^T\beta + v_i + \gamma_{ij}. \quad (4)$$

Compared to the unit-level nested error regression model (2), an additional error term γ_{ij}~iid $N(0, \sigma_\gamma^2)$ is assumed and independent of v_i or e_{ij}. Cai et al. (2020) first employed the LL data transformation to construct a new linking model for θ_{ij}, given by

$$\theta_i^{LL} = x_i^{LL}\beta + u_i^{LL}, \quad (5)$$

which is similar to Model (3) but with unobserved response θ_i^{LL}. The Lahiri-Suntornchost method are then employed to adjust the MSE_k and MST in estimating the information criteria under Model (5) with MSE_k and MST estimated by replacing the unobserved response θ_i^{LL} by y_i^{LL}, the LL-transformed observed response.

All the three papers aim at making simple adjustments to the regression packages available to data users, and their objective is not to decide on the best possible regression model selection criterion, but to suggest ways to adjust the MSE_k and MST before employing a data user's favorite model selection criterion. Given the conceptual and computational simplicity of the methods and wide availability of software packages for the standard regression model, these adjustments are likely to be adopted by users. To carry out variable selection under an assumed model (Fay-Herriot area model, nested error regression model, or two-fold subarea model), users can choose one of the above information criteria and estimate its values for a set of submodels under consideration with adjusted MSE and MST. The submodel with the smallest estimated information criterion value is selected as the final model.

Prof. Ghosh discussed various inferential aspects, including MSE approximations, under the FH and NER models, assuming the underlying model is true. In practice, variable selection is often conducted to select the optimal model so that inferential accuracy can be improved conditional on the selected model. An important follow-up question is how we can incorporate this additional uncertainty introduced by model selection into the MSE approximation at the inferential stage.

REFERENCES

CAI, S., RAO, J. N. K., DUMITRESCU, L., CHATRCHI, G., (2020). Effective transformation-based variable selection under two-fold subarea models in small area estimation. Statistics in Transition (*to appear*).

HAN, B., (2013). Conditional Akaike information criterion in the Fay-Herriot model. Statistical Methodology, 11, pp. 53–67.

JIANG, J., THUAN, N., (2015). The Fence Methods. World Scientific Publishing Co. Pte. Ltd., Singapore.

LAHIRI, P. ed., (2001). Model Selection. Beachwood, OH: Lecture Notes–Monograph Series, Institute of Mathematical Statistics.

LAHIRI, P., SUNTORNCHOST, J., (2015). Variable selection for linear mixed models with applications in small area estimation. Sankhya B, 77(2), pp. 312–320.

LI, Y., LAHIRI, P., (2019). A simple adaptation of variable selection software for regression models to select variables in nested error regression models. Sankhya B, 81(2), 302–371.

MEZA, J. L., LAHIRI, P., (2005). A note on the PC statistic under the nested error regression model. Survey Methodology 31, pp. 105–109.

RAO, J. N. K., MOLINA, I., (2015). Small Area Estimation, 2nd Edition. Hoboken: Wiley.