Small Area Estimation for Grouped Data

Yuki Kawakubo and Genya Kobayashi
Graduate School of Social Sciences, Chiba University

We are concerned with small area estimation for grouped data or frequency distribution. There are two fundamental models for model based small area estimation: one is the Fay–Herriot model for area level data, the other is the nested error regression model for unit level data. Because it is difficult to access unit level data in many cases, Fay–Herriot model is more widely used in practice. However, it is often the case that we can observe the frequency distribution of some quantity of interest in each area, which contains more information than the area level aggregated data. For such data, we propose a new model-based approach for small area estimation. We assume that the unit level unobserved quantity of interest follows the linear mixed model. We develop an Monte Carlo EM algorithm to estimate the unknown parameters in the model and calculate the empirical best predictors of the area means by a simple Gibbs sampling algorithm. The numerical performance of our proposed method is examined by simulations. In addition, we apply the proposed method to the income dataset in Japan.

The detailed model setup is explained as follows. Let z_{ij} be some quantity of interest of the jth unit in the ith area ($i = 1, \ldots, m; j = 1, \ldots, N_i$). We assume that some transformed value of z_{ij} follows the linear mixed model:

$$h(z_{ij}) = x_i^\top \beta + b_i + e_{ij}, \quad b_i \sim N(0, \tau^2), \quad e_{ij} \sim N(0, \sigma^2),$$

where $h(\cdot)$ is some known function, x_i is the area specific p-dimensional auxiliary variable, β is the unknown parameter vector of regression coefficients, b_i is the random area effect with the unknown variance parameter τ^2, e_{ij} is the error term with the unknown variance parameter σ^2 and b_i's and e_{ij}'s are mutually independent. We consider the situation where we cannot directly observe z_{ij}'s but can their frequency distribution in each area. Let $y_{ig} = \sum_{j=1}^{N_i} I(c_{g-1} \leq z_{ij} < c_g)$ for $g = 1, \ldots, G$, where $c_0, \ldots, c_G$ are thresholds for the groups, G is the number of groups and $I(\cdot)$ is the indicator function.

Based on the observed data $y_i = (y_{i1}, \ldots, y_{iG})^\top$'s and corresponding auxiliary variables x_i's, we estimate (or predict) the area means $\bar{z}_i$'s, where $\bar{z}_i = N_i^{-1} \sum_{j=1}^{N_i} z_{ij}$. We calculate the empirical best predictor of $\bar{z}_i$ by a simple Gibbs sampling algorithm. The unknown parameters in the model are estimated by the maximum likelihood method based on the marginal likelihood integrating out with respect to the random effects. Because the analytical evaluation of the marginal likelihood is difficult, we develop an Monte Carlo EM algorithm, where the E-step includes an Monte Carlo integration.