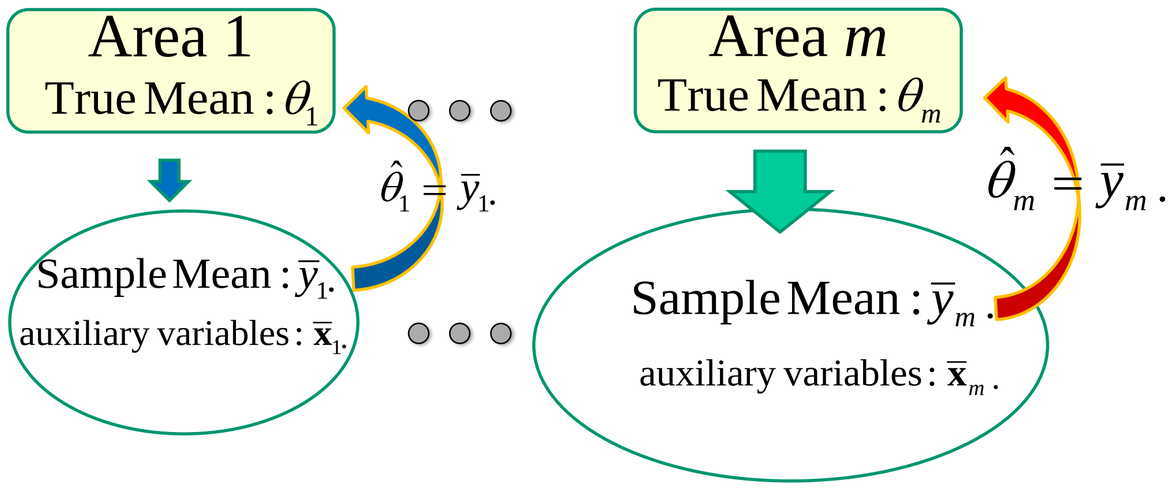
小地域推定におけるモデル分散パラメーターに対する0推定値の発生とその回避法

廣瀬　雅代　データ科学研究系　助教

1. Introduction: Estimation for each area (or domain) mean



Sample mean is not reliable

when there are many small areas

Utilizing specific Linear Mixed model (Fay and Herriot, 1979 etc.)

Fay Herriot (FH) model

$$\bar{y}_{i\cdot} = \theta_i + \bar{e}_{i\cdot},\ \theta_i = \bar{\mathbf{x}}_i'\boldsymbol{\beta} + u_i,$$
$$where\ \mathbf{u} \sim N(\mathbf{0}, \psi\mathbf{I}),\ \mathbf{e} \sim N(\mathbf{0}, diag(d_1, \cdots, d_m))$$
$$\text{Assumption}: d_1, \cdots, d_m \text{ are known}$$

Prediction: each true Mean $\theta_i = \bar{\mathbf{x}}_i'\boldsymbol{\beta} + \mathbf{u}_i$

Empirical Best Linear Unbiased predictor **(EBLUP)**

$$\hat{\theta}_i^{EB}(\mathbf{y}, \hat{\psi}) = (1 - B_i(\hat{\psi}))\,\bar{y}_i + B_i(\hat{\psi})\bar{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}(\hat{\psi})$$

$\hat{\psi}$: PR estimator, PML estimator, **REML estimator**

2. Practical Problem

PML and REML could yield zero estimates

[1] Avoiding zero estimates of ψ through specified adjustment factors (term)

[Li and Lahiri, 2010; Lahiri and Li, 2009]

$$\hat{\psi}_{\mathrm{AR}} = \arg\max_{0<\psi<\infty} h(\psi) L_{RE}(\psi, \mathbf{y})$$

$$h_{LL}(\psi) = A^{c_1/2}\,|\mathbf{V}|^{c_2/2m}\,|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|^{c_3/2}, c_1 > 0,\ e.g.\ \{A, A\,|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|^{1/2}\} \subset \{h_{LL}(\psi)\}$$

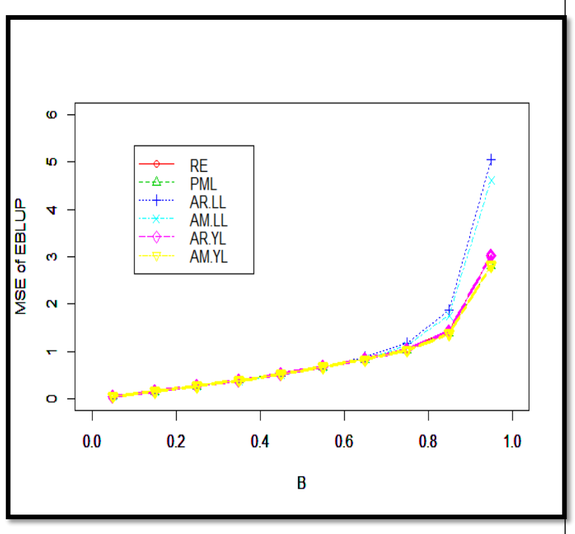
Mean Squared Prediction Error (MSPE) inflation could be caused by

their strictly positive estimators of ψ when B gets closer to one

[2] Preserving from such problems [Yoshimori and Lahiri, 2014]

$$h_{YL}(\psi) = \left[\arctan[tr\{I_m - diag(B_1, \cdots, B_m)\}]\right]^{1/m}$$

- New AM.YL and AR.YL provide strictly positive estimates, not like RE and PML.
- MSPEs of new EBLUPs perform well, compared with these of previous estimators (AR.LL and AM.LL).



Disadvantage:

Considered only in the context of the FH model

Study goal :

Extension to Nested Error Regression model which has two unknown variance parameters for avoiding zero estimates but keeping precision of EBLUP with REML estimator.

3. Extension to Nested Error regression model (NERM)

NERM (Battesse et al., 1988) : unknown two variance parameters

$$\bar{y}_{i\cdot} = \theta_i + \bar{e}_{i\cdot},\ \theta_i = \bar{\mathbf{x}}_i'\boldsymbol{\beta} + u_i,$$
$$where\ u_i \sim iid.N(0, \sigma_u^2),\ \bar{e}_{i\cdot} \sim indep.N(0, \sigma_e^2/n_i\}),\ Cov(\mathbf{u}, \bar{\mathbf{e}}) = 0$$
$$\boldsymbol{\psi} = (\sigma_u^2, \sigma_e^2)'$$

$$\hat{\theta}_i^{EB}(\mathbf{y}, \hat{\psi}) = \gamma_i(\hat{\psi})\,\bar{y}_i + (1 - \gamma_i(\hat{\psi}))\bar{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}(\hat{\psi})$$

$$\gamma_i(\hat{\psi}) = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_i}$$

Suggestion: $(\hat{\sigma}_u^2, \hat{\sigma}_e^2)' = \hat{\boldsymbol{\psi}}'_{\mathrm{AR}} = \arg\max_{\boldsymbol{\psi} \in \mathbf{R}_+^2} h(\boldsymbol{\psi}) L_{RE}(\boldsymbol{\psi}, \mathbf{y})$

where $h(\boldsymbol{\psi}) = \left(h_1(\sigma_u^2),\ 0\right)'$ $\quad e.g.\ h_1(\boldsymbol{\psi}) = [\arctan(\sigma_u^2)]^{1/\sqrt{m}}$

Theorem. 1. Under some regularity conditions, we have, for large *m*:

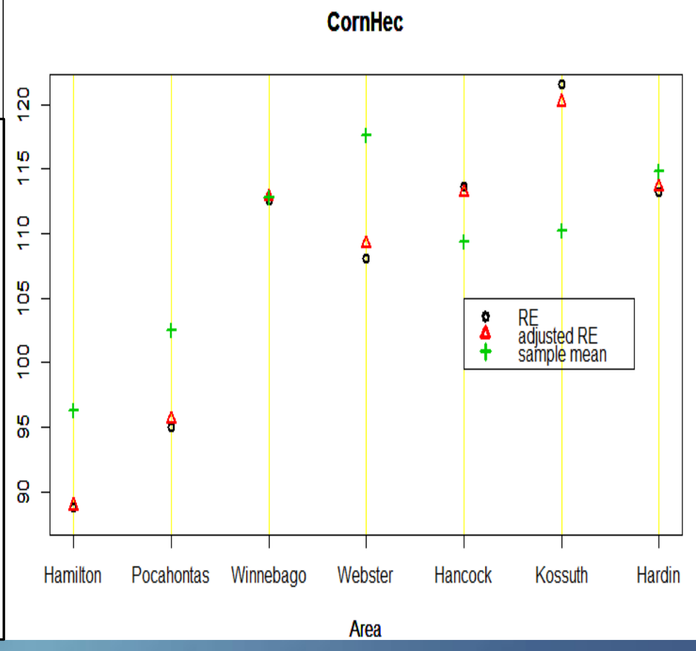
$$E[\hat{\boldsymbol{\psi}}_{\mathrm{AR}} - \hat{\boldsymbol{\psi}}_{\mathrm{RE}}] = [o(m^{-1})]_{2\times 1};\ E[(\hat{\boldsymbol{\psi}}_{\mathrm{AR}} - \boldsymbol{\psi})(\hat{\boldsymbol{\psi}}_{\mathrm{AR}} - \boldsymbol{\psi})'] = E[(\hat{\boldsymbol{\psi}}_{\mathrm{RE}} - \boldsymbol{\psi})(\hat{\boldsymbol{\psi}}_{\mathrm{RE}} - \boldsymbol{\psi})'] + [o(m^{-1})]_{2\times 2}$$
$$E\left[\left(\hat{\theta}_i(\hat{\boldsymbol{\psi}}_{\mathrm{AR}}) - \theta_i\right)^2\right] = E\left[\left(\hat{\theta}_i(\hat{\boldsymbol{\psi}}_{\mathrm{RE}}) - \theta_i\right)^2\right] + o(m^{-1});$$

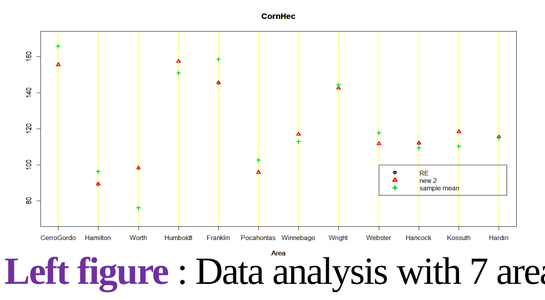
Advantages: Avoiding zero estimates, but keeping precision of EBLUP with REML estimator asymptotically. [with some assumptions given in Das et al. (2004) and Kubokawa (2011), included $\sup_{i\geq 1} n_i < \infty$].
Hirose (2015) suggested a class of adequate adjustment factor for avoiding zero estimates.

4. Data Analysis

Data: Corn soy bean data (Battese et al., 1988)

Survey and satellite data for corn and soy beans in 12 Iowa counties, obtained from the 1978 June Enumerative Survey of the U.S. Department of Agriculture and from land observatory satellites (LANDSAT) during the 1978 growing season. 37 observations. (Molina et al., 2015 [R package "sae"])





Left figure : Data analysis with 7 areas [In this case, REML estimate gets zero] we can see differences of EBLUP between the usage of RE and Adjusted RE.

Right figure : Data analysis with All 12 areas [In this case, REML estimate gets non- zero]
Adjusted RE provides very similar results to EBLUP with REML.

Reference

[1] G.E., Battese, R.M., Harter, and W. A. Fuller. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.*, **83**: 28-36
[2] K. Das, J. Jiang and J.N.K. Rao (2004). Mean squared error of empirical predictor, *Ann. Statist.*, **32:** 818-840
[3] R.E. Fay and R.A. Herriot (1979). Estimates of income for small places: an application of James-Stein procedures to census data, *J. Amer. Statist. Assoc.*, **74:** 269-277
[4] M.Y. Hirose (2015). A Class Extension of Adequate Adjustment factor for Small Area Inference, EAR-BC 2015
[5] T. Kubokawa (2011). ON MEASURING UNCERTAINTY OF SMALL AREA ESTIMATORS WITH HIGHER ORDER ACCURACY, J. Japan Statist. Soc. 41: 93-119
[6] P. Lahiri and H. Li (2009). Generalized Maximum Likelihood Method in Linear Mixed Models with an Application in Small Area Estimation. *In Proceedings of the Federal Committee on Statistical Methodology Research Conference*
[7] H. Li and P. Lahiri (2010). An adjusted maximum likelihood method for solving small area estimation problems, *J. Multivariate Anal.*, **101:**882-892
[8] I. Molina and Y. Marhuenda (2015). sae: An R Package for Small Area Estimation. *The R Journal*, **7**: 81–98
[9] M. Yoshimori and P. Lahiri (2014). A new adjusted maximum likelihood method for the Fay-Herriot small area model, *J. Multivariate Anal.*, **124:**281-294.