小地域推定問題に対する ”モデルに基づくアプローチ” の新たな課題 ---海外の事例を通して---

統計数理研究所　統計思考院･データ科学研究系

廣瀬雅代

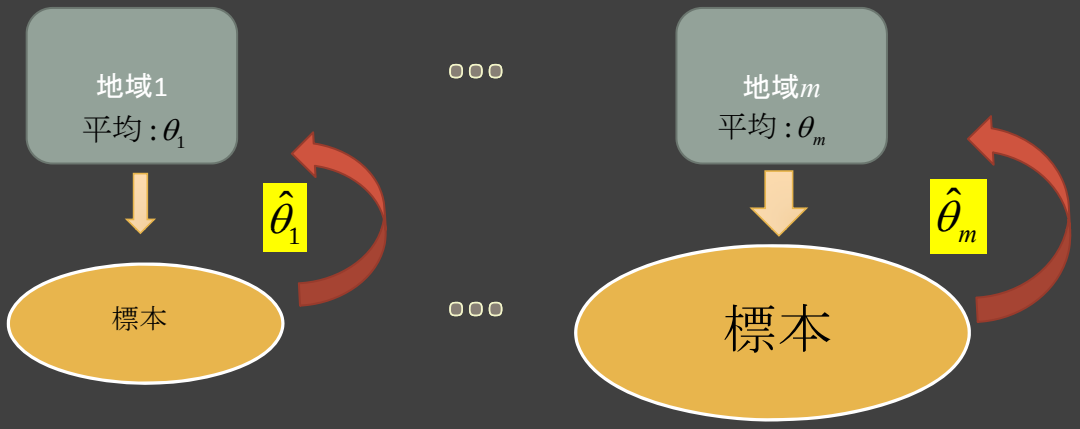
目次

1. 小地域推定問題
2. モデルに基づくアプローチと活用事例
3. モデルに基づくアプローチに対する課題と取り組み

1. 小地域推定問題

Small area estimation

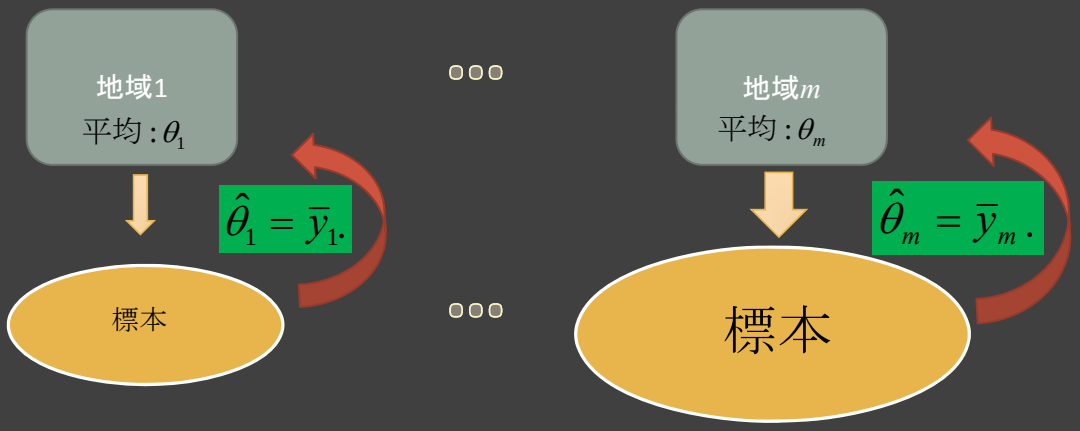
- 目的: 各地域 (州, 学区, domain) に対する特性値(平均等)の推定



Design Based approach

[Direct estimation]

- 目的: 各地域 (州, 学区, domain)に対する平均を推定



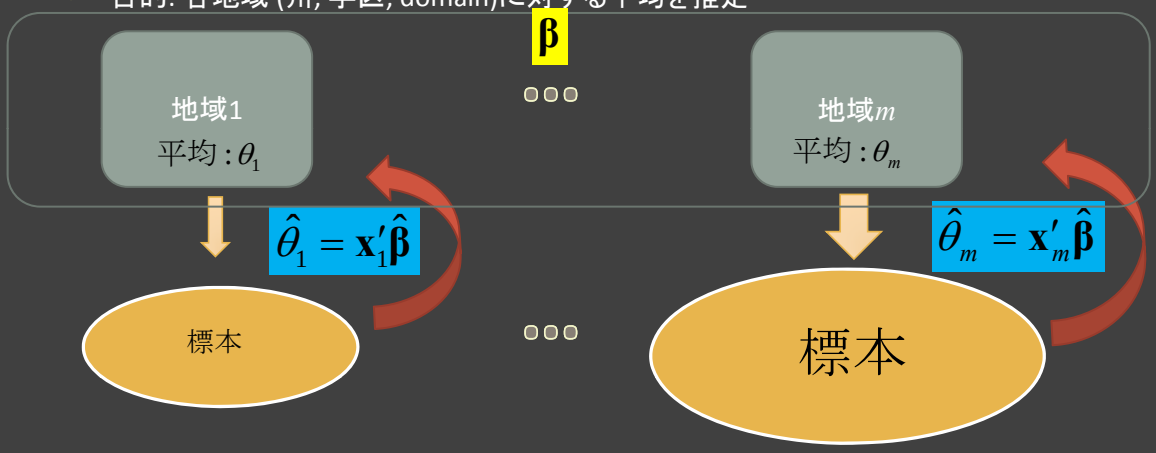
推定が不安定

モデルに基づくアプローチと活用事例

Implicit model

[Synthetic estimation/Composite estimation]

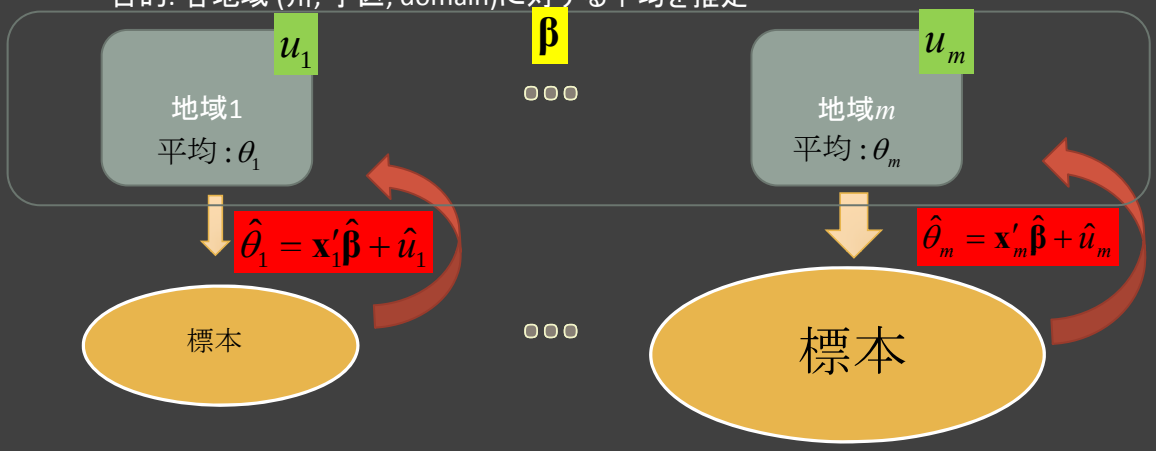
- 目的: 各地域 (州, 学区, domain)に対する平均を推定



Explicit model

[Empirical Best Linear Unbiased Predictor (EBLUP)/ Empirical Bayes (EB) /Hierarchical Bayes (HB)]

- 目的: 各地域 (州, 学区, domain)に対する平均を推定



Explicit model の有用性

Rao and Molina (2015)

1. 仮定したモデル下での最適な推定が可能
2. 各地域ごとのMSEの評価が可能
3. データからモデルを検証できる
4. 反応変数の自然な性質と複雑なデータ発生構造をモデルに組み入れることが可能
   - Spatial, time series structures

→ **EBLUP, EB, HB**

Model Based Approach

- Implicit model
- Explicit model
  - Aggregate Level model
    Unit levelのデータを使うことができないとき
  - Unit Level model
    Unit levelのデータを使うことができるとき

Aggregate level model

Direct estimates (直接推定値) と地域特有な補助情報との関係を表したモデル

-Fay Herriot model (1979)

$$g(\mathbf{y}) = \boldsymbol{\theta} + \mathbf{e},\ \boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

- 各地域の標本平均: $\mathbf{y}_{m\times 1} = (\bar{y}_{1.}, \cdots, \bar{y}_{m.})'$
- 各地域の平均: $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)'$
- 補助変数: $\mathbf{X}_{m\times p} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)'$
- 回帰係数: $\boldsymbol{\beta} \in R^p$

  $rank(\mathbf{X}) = p < \infty,\ \sup_{i,k\geq 1} |\{\mathbf{X}\}_{ik}| < \infty$
- 地域の差異 $\mathbf{u} = (u_1, \ldots, u_m)',\ \mathbf{e} = (e_1, \ldots, e_m)'$

  それぞれ独立に $u_i \sim N(0, a),\ e_i \sim N(0, d_i),\ (d_1, \cdots, d_m$: 既知$)$

EBLUP

Unit level model

Unitごとの変数の値とunitごとの共変量との関連を表したモデル

-Nested Error Regression model (Battese et al., 1988)

$$\mathbf{y} = \boldsymbol{\theta}^* + \mathbf{e},\ \boldsymbol{\theta}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u}$$

- 各Unitの値: $\mathbf{y} = (\mathbf{y}_1', \cdots, \mathbf{y}_m')'\ where\ \mathbf{y}_i = (y_{i1}, \cdots, y_{in_i})'$
- 各domainの平均: $\boldsymbol{\theta} = (\bar{\theta}_1^*, \cdots, \bar{\theta}_m^*)'$
- 補助変数: $\mathbf{X}^*_{(N\times p)} = (\mathbf{X}^*_1, \ldots, \mathbf{X}^*_m)',\ N = \sum_{i=1}^{m} n_i$
- 回帰係数: $\boldsymbol{\beta} \in R^p$

  $rank(\mathbf{X}^*) = p < \infty,\ \sup_{i,k\geq 1} |\{\mathbf{X}^*\}_{ik}| < \infty,\ \sup_{i\geq 1} n_i < \infty$
- 地域の差異 $\mathbf{u} = (u_1, \ldots, u_m)',\ \mathbf{e} = (e_1, \ldots, e_N)',\ \mathbf{Z}^* = diag\{\mathbf{1}_{n_1}, \cdots, \mathbf{1}_{n_m}\}$

  それぞれ独立に $u_i \sim N(0, \sigma_u^2),\ e_{ij} \sim N(\mathbf{0}, \sigma_e^2)$

EBLUP

θ_iの予測: EBLUP

Empirical Best Linear Unbiased Predictor (EBLUP)

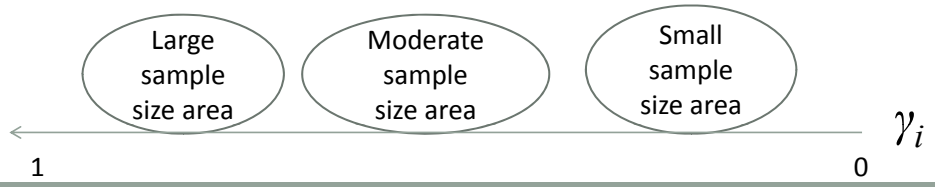
- Nested error regression model (Battesse et al., 1988; Prasad and Rao, 1990)

$$\hat{\theta}_i^{\text{EBLUP}} = \hat{\gamma}_i \bar{y}_i . + (1-\hat{\gamma}_i)\bar{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y},\ \hat{\mathbf{V}} = diag\{\hat{\mathbf{V}}_i\},\ \hat{\mathbf{V}}_i = \hat{\sigma}_u^2\mathbf{1}_{\mathbf{n_i}}\mathbf{1}'_{\mathbf{n_i}} + \hat{\sigma}_e^2\mathbf{I}_{\mathbf{n_i}}$$

- 縮小因子γ_i の推定量

$$\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / n_i}$$



事例: Per Capita Income [アメリカ]

Per Capita Income (PCI)の推定 (Fay and Herriot, 1979)

1. Design based approach [直接推定量]の活用
   - 約15000の地域で500人未満の地域
     - 約500人の地域　CV:約13%
     - 約100人の地域　CV:約30%
2. Small area placesに対して、Model based approachを用いて推定
3. Rao and Molina (2015) “1974年国勢調査局にて採用“

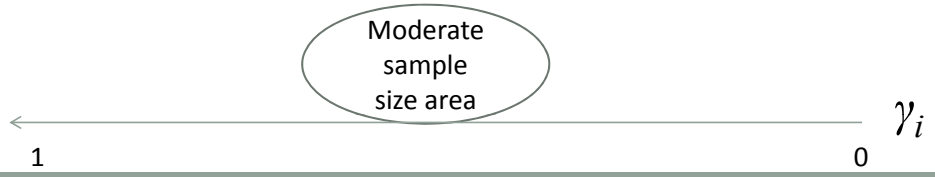
θ_iの予測: EBLUP

EBLUP

$$\hat{\theta}_i^{\text{EBLUP}} = \hat{\gamma}_i \bar{y}_i . + (1-\hat{\gamma}_i)\bar{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}$$

ある正則条件下で$\mathrm{E}[(\hat{\theta}_i^{EBLUP} - \theta_i)^2] \approx \frac{1}{\gamma_i}\mathrm{E}[(\bar{y}_i - \theta_i)^2] \text{ as } m \to \infty$

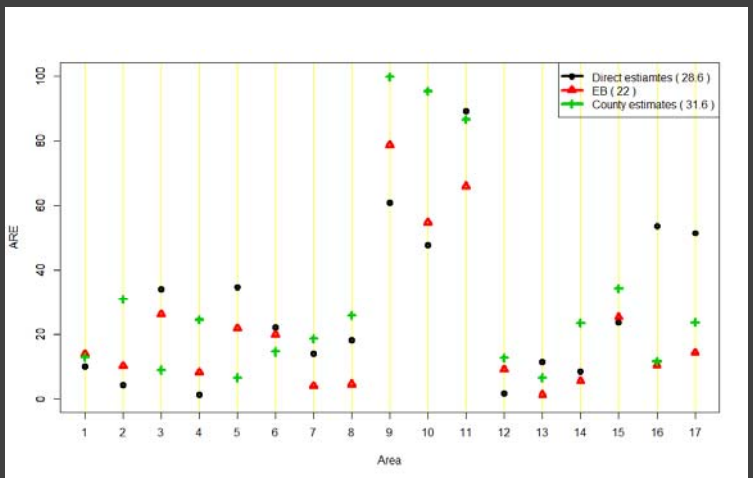
- たとえば $\gamma_i = 1/2$のとき、$\mathrm{E}[(\hat{\theta}_i^{EBLUP} - \theta_i)^2] \approx 2\mathrm{E}[(\bar{y}_i - \theta_i)^2] \text{ as } m \to \infty$



事例: Per Capita Income [アメリカ]

1973年Special complete census実施地域における1972年の値と1972年の推定値との比較

(Data source: Rao and Molina, 2015; Fay and Herriot, 1979)



$$ARE_i = \frac{|\hat{\theta}_i - T_i|}{T_i} \times 100$$

事例: Poverty counts [アメリカ]

SAIPE program in U.S.Census Bureau
(http://www.census.gov/did/www/saipe/about/index.html)

- the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program provides annual estimates of income and poverty statistics for all school districts, counties, and states.
- The main objective of this program is to provide estimates of income and poverty for the administration of federal programs and the allocation of federal funds to local jurisdictions.
- In addition to these federal programs, state and local programs use the income and poverty estimates for distributing funds and managing programs.

例: Title I fund: Over $7 billion dollars of funds [Rao (2003), National research council (2000)]

- 恵まれないこどもたちのための補償教育プログラム

➡ Model based approach

モデルに基づくアプローチに対する課題と取り組み

RAO AND MOLINA (2015)

Model based approachの課題
直接推定量の活用

SAIPEにおける直接推定値

(Rao and Molina, 2015; http://www.census.gov/did/www/saipe/methods/statecounty/20102014county.html)

1. 2005年まで:人口動態調査(CPS)を基に導出

2. 2005年以後: American Community Survey (ACS)結果を基に導出
   - CPSより大規模なサンプリングが行われている

有用な直接推定値の活用

   - County estimates (約3140) : Model based approach

Model based approachの課題

Model misspecification [Model, Linking model の仮定が崩れたとき]
- Jiang et al. (2011), You and Rao (2002) etc.

元のスケールへの変換 [非線形変換を行うとバイアスが生じる]
- Slud and Maiti (2006) etc.

Benchmarkingに対する問題

$$\sum_{i=1}^{m} \omega_i \bar{y}_i \neq \sum_{i=1}^{m} \omega_i \hat{\theta}_i^{EB},$$

- You and Rao (2002) , Wang, Fuller and Qu (2008), etc.

モデルに基づくアプローチの新たな課題

Model based approachの新たな課題1

0推定値の発生 $\hat{a} = 0$

例: 1989-1992, US 5-17 years old state poverty rate (Bell, 1999)

$$\hat{\theta}_i^{\text{EBLUP}} = (1-\hat{B}_i)\bar{y}_{i\cdot} + \hat{B}_i \mathbf{x}_i'\hat{\boldsymbol{\beta}},\ \hat{B}_i = d_i/(\hat{a}+d_i)$$

$$\hat{a} = 0$$

$$\hat{B}_i = 1\ (0 < B_i < 1) \Rightarrow \hat{\theta}_i^{\text{EBLUP}} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$$

①非現実的 ($0<a<\infty$)

②地域の差異が反映されない

すべての地域に対してsynthetic estimatesを使用

③Parametric bootstrap 法による信頼区間利用が不可能

- Hall and Maiti (2006), Chatterjee et al.(2008)

狭義正の推定値を得る方法

Wang and Fuller (2003)

Morris and Tang (2011)

Li and Lahiri (2010)

Yoshimori and Lahiri (2014a)

Mix estimator

Mix estimator (Rubin-Bleuer and You, 2013; Molina et al., 2015)

$$\hat{a}_{Mix.LL} = \begin{cases} \hat{a}_{RE} & \text{if } \hat{a}_{RE} > 0 \\ \hat{a}_{LL} & \text{otherwise} \end{cases}$$

ある正則条件下で以下が成り立つ

$$E[\hat{a}_{Mix.LL} - a] = E[\hat{a}_{RE} - a] + o(m^{-1}), as\ m \to \infty$$

Yoshimori and Lahiri(2014a)と同様の漸近的結果

Yoshimori and Lahiri (2014a)

新たな調整項 [Under the Fay-Herriot model]

$$h_{\mathrm{YL}}(a) = \arctan[tr(\mathbf{I} - \mathbf{B})]^{1/m}$$

$$\mathbf{B} = \mathrm{diag}(B_1, \cdots, B_m),\ B_i = \frac{d_i}{a + d_i}$$

New adjusted PML estimator (AM.YL)

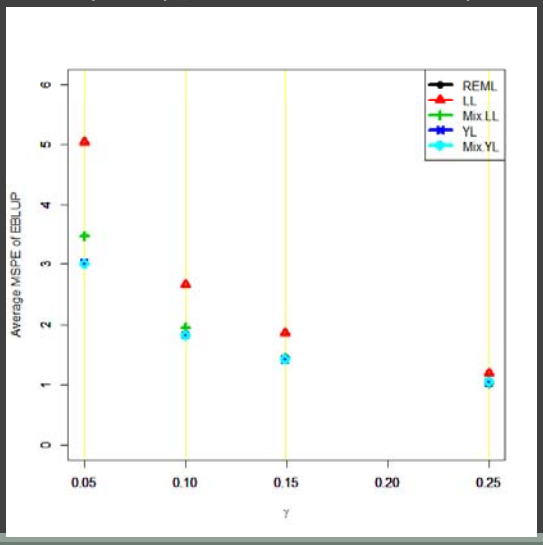
$$\hat{a}_{\mathrm{AM.YL}} = \arg\max_{0<a<\infty} h_{\mathrm{YL}}(a) L_p(a, \mathbf{y})$$

New adjusted REML estimator (AR.YL)

$$\hat{a}_{\mathrm{AR.YL}} = \arg\max_{0<a<\infty} h_{\mathrm{YL}}(a) h_{\mathrm{RE}}(a) L_p(a, \mathbf{y})$$

MIX VS YL estimator

Data Source: Rao and Molina, (2015) [Yoshimori and Lahiri (2014b)]



Model based approachの新たな課題2

計算機の利用

統計手法のソフトウェア化

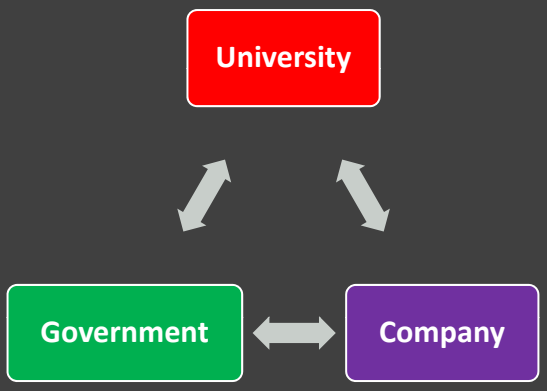
- SAS: Mukhopadhyay et al. (2011)
- R package: sae (Molina and Marhuenda, 2015) etc

計算量負荷から計算量軽減へ

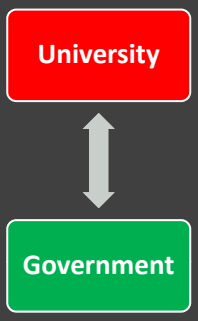
- Bootstrap methodの活用に消極的
- Yoshimori and Lahiri (2014c) Fay-Herriotモデル下での信頼区間構築

課題への取り組みと連携

アメリカ



カナダ



ヨーロッパ

The European working group on small area estimation (EWORSAE)

http://sae.wzr.pl/index.php

SAE(Small Area Estimation) conference

2001: Maryland, US

2005: Jyvaskyla, Finland

2007: Pisa, Italy

2009: Elche, Spain

2011: Trier, Germany

2013: Bangkok, Thailand

2014: Poznan, Poland

2015 : Santiago, Chile

2016: Maastricht, Netherland

Small area estimation project in Japan?

Reference (1/3)

1. Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83:** 28-36.
2. Bell, W. R. (1999). Accounting for uncertainty about variances in small area estimation. *Bulletin of the International Statistical Institute*, *52*.
3. Chatterjee, S., Lahiri, P., & Li, H. (2008). On small area prediction interval problems. *The Annals of Statistics*, **36:** 1221-1245.
4. Fay III, R. E., & 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.
5. Hall, P., & Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B*, **68:** 221-238.
6. Jiang, J., Nguyen, T., & Rao, J. S. (2011). Best predictive small area estimation. *Journal of the American Statistical Association*, **106***:* 732-745.
7. Li, H., & Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of multivariate analysis*, **101***:* 882-892.
8. Molina, I., & Marhuenda, Y. (2015). Package 'sae'.
9. Molina, I., & Morales, D. (2009). Small Area Estimation of Poverty Indicators. *Boletín de Estadística e Investigación Operativa*, **25***:* 218-225.
10. Molina, I., Rao, J. N. K., & Datta, G. S. (2015). Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random area effects. *Survey Methodology*, **41:** 1-19.

Reference (2/3)

11. Morris, C., & Tang, R. (2011). Estimating random effects via adjustment for density maximization. *Statistical Science*, **26**: 271-287.

12. Mukhopadhyay, P. K., & McDowell, A. (2011). Small area estimation for survey data analysis using SAS software. In *Proceedings of the SAS Global Forum 2011 Conference. http://support. sas. com/resources/papers/proceedings11/336-2011. pdf*.

13. National Research Council. (2000). *Small-Area Estimates of School-Age Children in Poverty:: Evaluation of Current Methodology*. In Citro, C. F. & Kalton, G. (Eds.). National Academies Press.

14. Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association*,, **85:** 163-171.

15. Rao, J.N.K. (2003). *Small Area Estimation*, Wiley.

16. Rao, J. N. K., & Molina, I. (2015). *Small Area Estimation* 2nd edition, Wiley.

17. Slud, E. V., & Maiti, T. (2006). Mean-squared error estimation in transformed Fay–Herriot models. *Journal of the Royal Statistical Society: Series B*, **68***:* 239-257.

18. Wang, J., & Fuller, W. A. (2003). The mean squared error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, **98:** 716-723.

19. Wang, J., Fuller, W. A., & Qu, Y. (2008). Small area estimation under a restriction. *Survey methodology*, **34***:* 29-36.

20. Rubin-Bleuer, S. and You, Y. (2013). A Positive Variance Estimator for the Fay-Herriot Small Area Model, SRID2-12-OOIE, Statistics Canada.

Reference (3/3)

21. Yoshimori, M., & Lahiri, P. (2014a). A new adjusted maximum likelihood method for the Fay–Herriot small area model. *Journal of Multivariate Analysis*, **124:** 281-294.

22. Yoshimori, M., & Lahiri, P. (2014b). Supplementary material to Yoshimori, M and Lahiri, P. (2014a). Unpublished note.

23. Yoshimori, M., & Lahiri, P. (2014c). A second-order efficient empirical Bayes confidence interval. *The Annals of Statistics*, **42***:* 1-29.

24. You, Y., & Rao, J. N. K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, **30**: 431-439.

正値地域データを解析するための変換モデルについて

菅澤翔之助

統計数理研究所
リスク解析戦略研究センター特任研究員

2016 年 1 月 29 日

概要

1. 基本となるモデル (Fay-Herriot モデル) について
2. 正値データのケースについて
3. unmatched sampling とリンク関数によるモデル
4. 今後の展開

1. Fay-Herriot モデル

目標: y_i を地域単位の集計データとしたとき $\theta_i = E[y_i|\theta_i]$ を推定したい.

難点: 集計数が少ないために y_i が θ_i の良い推定量になっていない.

$\Rightarrow$　モデルの力を利用して y_i よりも精度良い推定量を構成する.

そのために以下の 2 つの構造を仮定.

- $y_i|\theta_i \sim N(\theta_i, D_i)$

  y_i は真値 θ_i の周りで正規分布している.

- $\theta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i, \quad v_i \sim N(0, A).$

  θ_i は地域毎の共変量 $\boldsymbol{x}_i$ と地域特有の効果 v_i によって説明される.

このモデルを Fay-Herriot モデル (Fay and Herriot, 1979) という.

1. Fay-Herriot モデル

前スライドの構造をまとめて表現すると以下のようになる.

Fay-Herriot モデル

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i + \varepsilon_i, \quad i = 1, \ldots, m, \quad v_i \sim N(0, A), \quad \varepsilon_i \sim N(0, D_i).$$

- D_i は既知 (実際は何らかのデータから計算する).
- $v_1, \ldots, v_m, \varepsilon_1, \ldots, \varepsilon_m$ は全て独立.
- 未知パラメータは $\boldsymbol{\beta}$ および A.
- 興味の対象: $\theta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i$.

1. Fay-Herriot モデル

- θ_i の (2 乗誤差の意味で) 最良な推定量は

$$\widetilde{\theta}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \frac{A}{A + D_i}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})$$

  となる. この推定量は $E[(\widetilde{\theta}_i - \theta_i)^2] \leq E[(y_i - \theta_i)^2]$ を満たす.

  $\widetilde{\theta}_i$ を利用することで y_i よりも精度良い推定値を与えることができる.

- $\widetilde{\theta}_i$ は未知パラメータ $\boldsymbol{\beta}, A$ に依存しているので、実際に利用するためにはそれらを推定値 $\widehat{\boldsymbol{\beta}}, \widehat{A}$ で置き換えた $\widehat{\theta}_i$ を用いる必要がある.

  $\widehat{\boldsymbol{\beta}}, \widehat{A}$ は最尤推定法、モーメント法などの手法が提案されている.

2. 正値データのケース

現実のデータ解析では y_i が正値データのケースはよくある.

- FH モデルは $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i + \varepsilon_i$ だったので各地域データ y_i は平均 $\boldsymbol{x}_i'\boldsymbol{\beta}$ の正規分布に従っていると仮定している.

  正値データは対称に分布していないケースが多い.

- 正値データに対しては y_i の代わりに $\log y_i$ に FH モデルを当てはめる.

$$\log y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i + \varepsilon_i, \qquad i = 1, \ldots, m$$

  対数変換することは y_i の取りうる範囲を実数にすることと分布を対称にする２つの目的がある.

2. 正値データのケース

再掲

$$\log y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i + \varepsilon_i, \qquad i = 1, \ldots, m$$

- $\theta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i$ に対して地域パラメータ $\mu_i = \exp(\theta_i)$ の推定を考える.

  この枠組みでは Slud and Maiti (2006) によって推定量およびリスク評価法が与えられている.

- 対数変換以外の変換について.

  対数変換を含むクラスの変換を考えて柔軟に推定するモデルが Sugasawa and Kubokawa (2015) で提案されている.

2. 正値データのケース

再掲

$$\log y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i + \varepsilon_i, \qquad i = 1, \ldots, m$$

- このモデルは利用しやすいが実は問題がある．そもそもの目的は以下であった．

  目標: y_i を地域単位の集計データとしたとき $\theta_i = E[y_i|\theta_i]$ を推定したい．

  前スライドで興味あるパラメータを $\mu_i = \exp(\theta_i)$ と定義した．このモデルは $y_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + v_i + \varepsilon_i) = \mu_i \exp(\varepsilon_i)$ と表現できるので

  $$E[y_i|\mu_i] = \mu_i \exp(D_i/2) \neq \mu_i$$

  となる．

- 単純に y_i を変換したモデルではそもそもの目的からずれたものを推定している．

3. unmatched sampling とリンク関数によるモデル

You and Rao (2002) はリンク関数を用いたモデルを提案した.

$$y_i = \theta_i + \varepsilon_i, \qquad h(\theta_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i, \quad i = 1, \ldots, m$$

- $h(\theta_i) = \log \theta_i$ ととると対数変換モデルの代用になる.
- このモデルは $E[y_i|\theta_i] = \theta_i$ を満たす.
- You and Rao (2002) は一般のリンク関数 $h(\cdot)$ の設定でパラメータに事前分布を設定してベイズ推定を行っている.

  事前分布のパラメータを設定する必要があり、(モデル自体は有用であるが) 実用上あまり好まれない.

3. unmatched sampling とリンク関数によるモデル

本研究の目的: 頻繁に使用される対数リンク $h(\theta_i) = \log \theta_i$ のケースに限定して頻度論的な推定方法を考える.

$\Rightarrow$ 対数変換の手法に対する代替手法を提案する.

困難な点

- 以下のように y_i の周辺分布が解析的に得られない.

$$f(y_i) = \frac{1}{(2\pi D_i)^{1/2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(y_i - \exp(\sqrt{A}t_i + \boldsymbol{x}_i'\boldsymbol{\beta}))^2}{2D_i} \right) u(t_i) dt_i,$$

  $u(\cdot)$ は標準正規分布の密度関数.

  最尤推定を行うのは数値積分を含んだ繰り返し計算が必要.

3. unmatched sampling とリンク関数によるモデル

再掲

$$y_i = \theta_i + \varepsilon_i, \quad \log \theta_i = \boldsymbol{x}_i' \boldsymbol{\beta} + v_i, \quad i = 1, \ldots, m$$
$$v_i \sim N(0, A), \quad \varepsilon_i \sim N(0, D_i)$$

未知パラメータ $\boldsymbol{\phi} = (\boldsymbol{\beta}', A)'$.

パラメータ推定方法: Godambe and Thompson (1989) による推定方程式を利用する.

- この推定方程式は y_i の 4 次までのモーメントが必要.
- 周辺尤度は解析的に求まらないが y_i の周辺モーメントは解析的に得られる.

3. unmatched sampling とリンク関数によるモデル

$m_i \equiv \mathrm{E}[y_i] = \exp(x_i'\boldsymbol{\beta} + A/2)$ に対して $\mu_{ki} = \mathrm{E}[(y_i - m_i)^k]$ と定義. このとき

$$\mu_{2i} = m_i^2(e^A - 1) + D_i, \quad \mu_{3i} = m_i^3(e^A - 1)^2(e^A + 2),$$
$$\mu_{4i} = m_i^4(e^A - 1)^2(e^{4A} + 2e^{3A} + 3e^{2A} - 3) + 6m_i^2 D_i(e^A - 1) + 3D_i^2.$$

さらに $u_{1i} = y_i - m_i$, $u_{2i} = (y_i - m_i)^2 - \mu_{2i}$ に対して $\boldsymbol{u}_i(y_i, \boldsymbol{\phi}) = (u_{1i}, u_{2i})'$ とし, $\boldsymbol{\Sigma}_i(\boldsymbol{\phi})$, $\boldsymbol{P}_i(\boldsymbol{\phi})$ を以下のように定義.

$$\boldsymbol{\Sigma}_i(\boldsymbol{\phi}) = \begin{pmatrix} \mu_{2i} & \mu_{3i} \\ \mu_{3i} & \mu_{4i} - \mu_{2i}^2 \end{pmatrix}, \quad \boldsymbol{P}_i(\boldsymbol{\phi})' = m_i \begin{pmatrix} \boldsymbol{x}_i & 2m_i\boldsymbol{x}_i(e^A - 1) \\ 1/2 & m_i(2e^A - 1) \end{pmatrix}.$$

このとき $\boldsymbol{\phi}$ の推定方程式は

$$\sum_{i=1}^{m} \boldsymbol{P}_i(\boldsymbol{\phi})'\boldsymbol{\Sigma}_i(\boldsymbol{\phi})^{-1}\boldsymbol{u}_i(y_i, \boldsymbol{\phi}) = \boldsymbol{0}.$$

3. unmatched sampling とリンク関数によるモデル

θ_i の推定について

- θ_i のベイス推定量は以下のようになる.

$$\widetilde{\theta}_i(y_i, \boldsymbol{\phi}) = E[\theta_i | y_i]$$
$$= \frac{\mathrm{E}_z\left[\exp\left\{\sqrt{A}z + \boldsymbol{x}_i'\boldsymbol{\beta} - (2D_i)^{-1}(y_i - \exp(\sqrt{A}z + \boldsymbol{x}_i'\boldsymbol{\beta}))^2\right\}\right]}{\mathrm{E}_z\left[\exp\left\{-(2D_i)^{-1}(y_i - \exp(\sqrt{A}z + \boldsymbol{x}_i'\boldsymbol{\beta}))^2\right\}\right]},$$

  $E_z[\cdot]$ は $z \sim N(0,1)$ に対する期待値を表す.

- 推定量を代入することで最終的に $\widehat{\theta}_i = \widetilde{\theta}_i(y_i, \widehat{\boldsymbol{\phi}})$ を得る.(積分の部分は解析的に計算できないので数値積分で評価する必要がある.)

3. unmatched sampling とリンク関数によるモデル

$\widehat{\theta}_i$ のリスク評価

$\widehat{\theta}_i$ のリスク評価のために $\widehat{\theta}_i$ の MSE を考える.

$$\mathrm{MSE}_i = E[(\widehat{\theta}_i - \theta_i)^2]$$

- 一般に MSE_i は未知パラメータ $\boldsymbol{\phi}$ に依存するので、MSE_i の精度良い推定量を用いる.

  具体的には MSE の推定量を $\widehat{\mathrm{MSE}}_i$ が $E\left[\widehat{\mathrm{MSE}}_i\right] = \mathrm{MSE}_i + o(m^{-1})$ を満たすように構成する.

- 推定方程式で定義した $\widehat{\boldsymbol{\phi}}$ の漸近分散、漸近バイアスを評価して解析的に求めることができる. またパラメトリックブートストラップを用いて構成することもできる.

3. unmatched sampling とリンク関数によるモデル

数値実験

対数リンクが真のとき、対数変換モデルはどのくらい機能するのか.

- データ生成過程

$$
\begin{aligned}
y_i =& \theta_i + \varepsilon_i, \quad \log \theta_i = \beta_0 + \beta_1 x_i + v_i, \quad i = 1, \ldots, 30 \\
& v_i \sim N(0, A), \quad \varepsilon_i \sim N(0, D_i)
\end{aligned}
$$

  $\beta_0 = 0, \beta_1 = 0.6,\ x_i \sim N(0, 5),\ D_i \sim U(5, 15),\ A = 0.5, 1.$

- θ_i の予測量を対数リンク、対数変換、FH のそれぞれから計算し、真の θ_i との MSE および bias を 1,000 回の繰り返から計算.

  また direct estimator y_i の MSE および bias も同様に計算.

3. unmatched sampling とリンク関数によるモデル

数値実験 (結果)

		対数リンク	FH	対数変換	y_i
$A = 1$	MSE	2.62	3.01	3.02	3.05
	bias	0.145	0.325	0.112	0.332
$A = 0.5$	MSE	2.63	2.76	2.78	2.86
	bias	-0.477	0.178	-0.046	0.192

3. unmatched sampling とリンク関数によるモデル

実データへの適用

- y_i: 2014 年の都道府県ごとの家計調査 (教育費), $\quad i = 1, \ldots, 47$
- D_i は 2006 年から 2013 年のデータから計算.
- 共変量 x_i として 2011 年の大規模家計調査の結果を利用する.

対数リンクモデルを当てはめる.

$$y_i = \theta_i + \varepsilon_i, \qquad \log \theta_i = \beta_0 + \beta_1 x_i + v_i, \quad i = 1, \ldots, 47$$

3. unmatched sampling とリンク関数によるモデル

実データへの適用

$$y_i = \theta_i + \varepsilon_i, \quad \log \theta_i = \beta_0 + \beta_1 x_i + v_i, \quad i = 1, \ldots, 47$$

推定値: $\widehat{\beta} = 0.984,\ \widehat{\beta} = 0.843,\ \widehat{A} = 3.16$

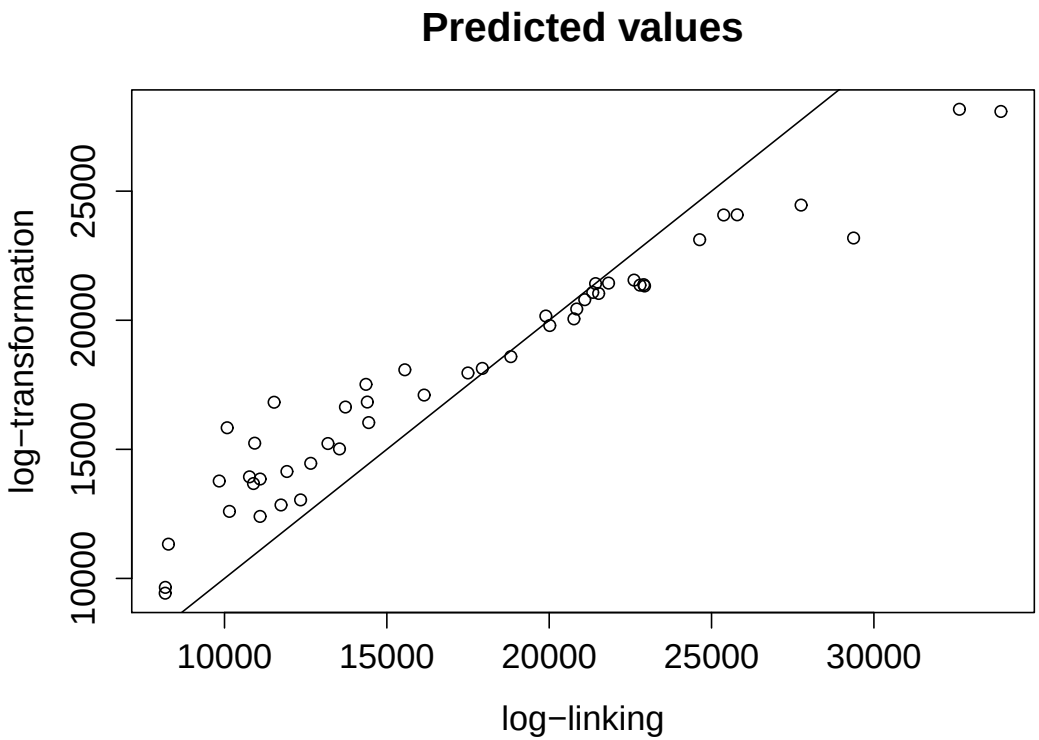


Figure: 対数リンクモデルと対数変換モデルの予測値

4. 今後の展開

考察

- unmatched sampling と対数リンクを用いたモデルは正値データに対して有用.
- 現実には割合などの有界連続値をとるデータもある. また実数値データでも非対称に分布しているケースもあるかもしれない.

  logistic リンク? $\mathbb{R}$ から $\mathbb{R}$ のリンク?
- そもそも対数リンクなど決め打ちしたリンクが適当とは限らない.

  リンクもデータから推定できると良い.

4. 今後の展開

ノンパラメトリックリンクを用いた unmatched-sampling モデル

$$y_i = \theta_i + \varepsilon_i, \qquad \theta_i = L(\boldsymbol{x}_i^t \boldsymbol{\beta} + v_i), \quad i = 1, \ldots, m.$$
$$v_i \sim N(0, 1), \quad \varepsilon_i \sim N(0, D_i)$$

リンク $L(\cdot)$ は以下のように P-spline を用いて表現する.

$$L(u) = \gamma_0 + \gamma_1 u + \ldots + \gamma_q u^q + \sum_{k=1}^{K} \gamma_{q+k} (u - t_k)_+^q.$$

- 関連する統計モデル: single index モデル (ただし既存のものは random effect を含まないもののみ)
- パラメータを頻度論的に推定するのは難しいが, 事前分布を設定してベイズ推定するのは比較的容易.(事後分布からのサンプリングは簡単に実行可能.)

  客観性を保つために $\boldsymbol{\beta}, \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{q+K})'$ に uniform prior を入れる.

4. 今後の展開

再掲

$$y_i = \theta_i + \varepsilon_i, \qquad \theta_i = L(\boldsymbol{x}_i^t \boldsymbol{\beta} + v_i), \quad i = 1, \ldots, m.$$

$$L(u) = \gamma_0 + \gamma_1 u + \ldots + \gamma_q u^q + \sum_{k=1}^{K} \gamma_{q+k} (u - t_k)_+^q.$$

- uniform prior を設定した場合, 事後分布は必ずしも proper にならない.

  適当な条件のもと事後分布は proper になることを示した.

- どんな連続値データに対しても利用可能.

  数値実験では対数リンクが misspecify されたケースで対数リンクに対する優越性なども観察された.

まとめ

- データを変換するモデルの代用として unmatched sampling とリンク関数によるモデルを導入した.
- 対数変換のケースに焦点を当て, 頻度論の枠組みでモデルパラメータの推定や小地域パラメータの推定, およびそのリスク評価について提案した.
- 今後の主展開: ノンパラメトリックリンクを用いたモデル

参考文献

- 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.
- Godambe, V. P. and Thompson, M. E. (1989). An extension of quasi-likelihood estimation (with Discussion). *Journal of Statistical Planning and Inference*. **22**, 137-152.
- Slud, E.V. and Maiti, T. (2006). Mean-squared error estimation in transformed Fay-Herriot models. *Journal of Royal Statistical Society: Series B*, **68**, 239-257.
- Sugasawa, S. and Kubokawa, T. (2015). Parametric transformed Fay-Herriot model for small area estimation. *Journal of Multivariate Analysis*, **139**, 295-311.
- You, Y. and Rao, J. N. K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, **30**, 3-15.

空間重み付き経験ベイズ推定と死亡データへの応用

川久保友超

千葉大学・法政経学部

2016 年 1 月 29 日

科研コンファレンス「経済統計・政府統計の理論と応用」

導入

要約

- 地理情報を組み込んだ小地域推定のモデルを提案.
- 提案モデルは，カウントデータや二値データなどの離散データにも適用できる.
- 死亡データへの応用を行い，標準化死亡比（SMR）と呼ばれるリスク指標を小地域推定する.
- 本発表の内容は，以下の working paper にもとづいている.

  Sugasawa, S., Kawakubo, Y. and Ogasawara, K.
  Geographically weighted empirical Bayes estimation via natural exponential family.
  arXiv preprint, arXiv:1508.01641.

導入

小地域推定

- 地域 i の所得の平均 μ_i を知りたいとき，サンプルサイズが小さい地域では，集計データ（標本平均）y_i は推定誤差が大きく信頼できない.
- μ_i を小地域母数，y_i をその direct estimator という.
- μ_i に周辺地域の情報を入れた事前分布を仮定すると，μ_i の経験ベイズ推定量は安定した推定量となる（borrowing strength）

  → model based estimator
- 線形混合モデル（Linear Mixed Model, LMM）と呼ばれるモデルのクラスが最も広く使われている.

導入

- カウントデータや二値データなどの離散データに対しては LMM は適切でないため，より広いクラスの一般化線形混合モデル（GLMM）がしばしば用いられる.
- しかしながら，GLMM は推定に数値積分が必要で，実行上煩雑.
- GLMM に代わり，小地域母数に共役事前分布を仮定したモデルを用いる．このモデルは，解析的にベイズ推定量が導出できる等のメリットがある.

既存のモデル

- Ghosh and Maiti (2004, *Biometrika*) によって次のモデルが提案された.

  $y_1, \ldots, y_m$ は独立. $y_i|\theta_i$ および θ_i の分布を以下のように設定.

$$y_i|\theta_i \sim f(y_i|\theta_i) = \exp[n_i(\theta_i y_i - \psi(\theta_i)) + c(y_i, n_i)],$$
$$\theta_i \sim \pi(\theta_i|\nu, m_i) = \exp[\nu(m_i\theta_i - \psi(\theta_i))]C(\nu, m_i),$$

  n_i: 既知

  $m_i = \psi'(\boldsymbol{x}_i^t\boldsymbol{\beta})$　　ハイパーパラメータ: $\boldsymbol{\phi} = (\boldsymbol{\beta}^t, \nu)^t$

  興味の対象: $\mu_i = E[y_i|\theta_i]$.

- 分散構造として $\mathrm{Var}(y_i|\theta_i) = n_i^{-1}Q(\mu_i) = n_i^{-1}(v_0 + v_1\mu_i + v_2\mu_i^2)$ を仮定.

  これらは正規分布、二項分布、ポアソン分布を含むため実用上十分広いクラス.

既存のモデル

- θ_i には共役事前分布を入れているので周辺尤度および事後分布は解析的に求まり，μ_i のベイズ推定量として以下を得る.

$$\widetilde{\mu}_i = \widetilde{\mu}_i(y_i, \boldsymbol{\phi}) = \frac{n_i y_i + \nu m_i}{n_i + \nu}.$$

この推定量は未知のパラメータ $\boldsymbol{\phi}$ に依存するので infeasible

- $\boldsymbol{\phi}$ は周辺尤度から以下のように推定できる.

$$\widehat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmax}} \sum_{i=1}^{m} \left[\log C(\nu, m_i) - \log C(n_i + \nu, \widetilde{\mu}_i(y_i, \boldsymbol{\phi})) \right].$$

この $\widehat{\boldsymbol{\phi}}$ をベイズ推定量 $\widetilde{\mu}_i$ に代入することで経験ベイズ推定量 $\widehat{\mu}_i = \widetilde{\mu}_i(y_i, \widehat{\boldsymbol{\phi}})$ を得る. → model based estimator

動機

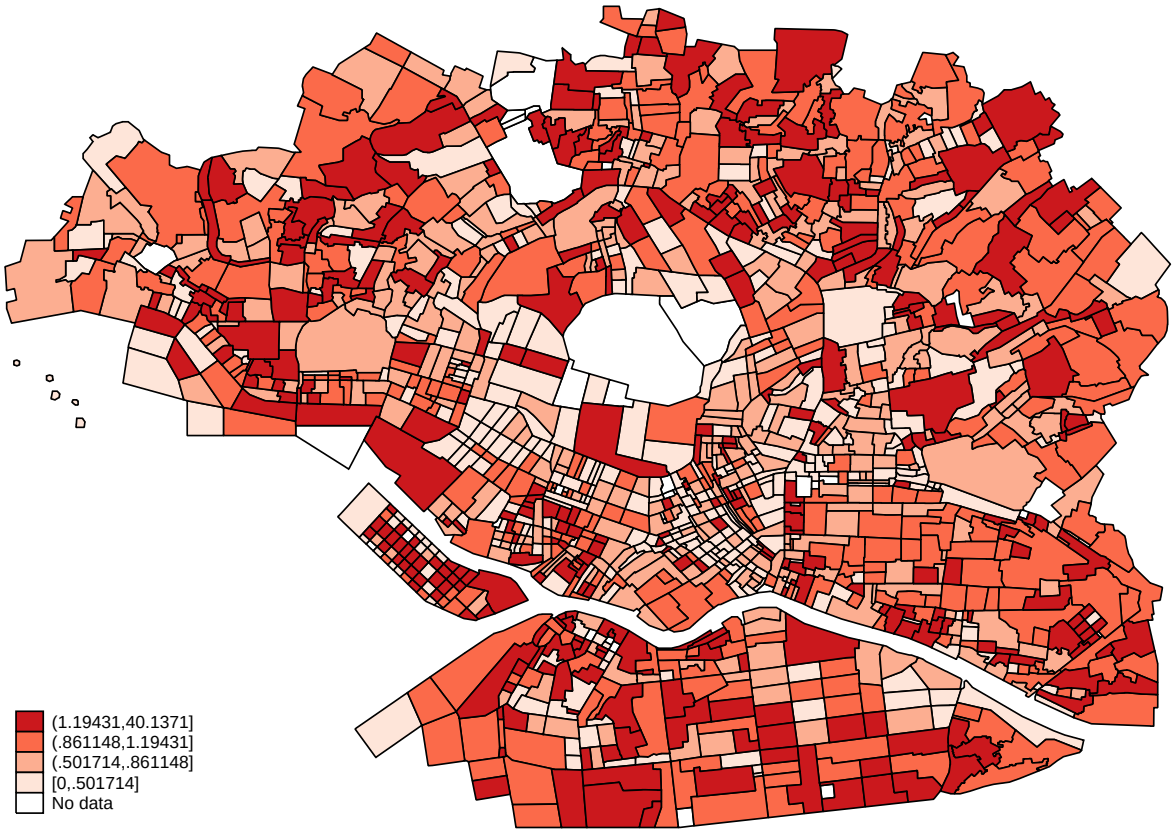


Figure : 1930 年東京市における男性の SMR の分布

問題提起

- 既存のモデルではハイパーパラメータが各地域で共通であることを仮定していた.

  ハイパーパラメータは各地域で異なっている (空間非定常) と考えるのが自然.

- 各地域の地理的関係性が情報として得られている場合，それもモデルに組み込むことができた方が精度良い推定ができそう.

空間重み付き経験ベイズ推定

- 既存モデルにおいて $\boldsymbol{\phi} = \boldsymbol{\phi}_i$ としたモデルを提案する.

  $\boldsymbol{\phi}_i,\ i = 1, \ldots, m$ を以下のように推定

$$\widehat{\boldsymbol{\phi}}_i = \operatorname*{argmax}_{\boldsymbol{\phi}} \sum_{k=1}^{m} w_{ik} \left[\log C(\nu, m_k) - \log C(n_k + \nu, \widetilde{\mu}_k(y_k, \boldsymbol{\phi})) \right],$$

$$m_k = \psi'(\boldsymbol{x}_k^t \boldsymbol{\beta}), \quad \widetilde{\mu}_k(y_k, \boldsymbol{\phi}) = (n_k y_k + \nu m_k)/(n_i + \nu).$$

  w_{ik} は２つの地域 i, k 間のウェイトで、d_{ik} を地域間の距離，b をバンド幅としたときに，カーネル関数 $K(x)$ を用いて $w_{ik} = K(d_{ik}/b)$ と定義する.

- カーネル関数として対称かつ有界サポートをもつカーネルを用いる．今回は以下の 4 次カーネルを用いる.

$$K(x) = \begin{cases} \frac{15}{16}(1 - x^2)^2, & (0 \leq x \leq 1), \\ 0, & (x > 1). \end{cases}$$

空間重み付き経験ベイズ推定

- 地域毎に推定された $\widehat{\phi}_i$ を用いて空間重み付き経験ベイズ（GWEB）推定量を

$$\widehat{\mu}_i^{\mathrm{GWEB}} = \frac{n_i y_i + \widehat{\nu}_i \widehat{m}_i}{n_i + \widehat{\nu}_i},$$

  と定義.

- バンド幅 b は以下のような Cross Validation による基準を用いて選択する.

$$\mathrm{CV}(b) = \sum_{i=1}^{m} \left\{ y_i - \widehat{\mu}_{(-i)}^{\mathrm{GWEB}}(b) \right\}^2,$$

  ただし $\widehat{\mu}_{(-i)}^{\mathrm{GWEB}}(b)$ は、バンド幅 b のもとで $y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m$ から求めた $\widehat{\phi}_i$ を用いた μ_i の GWEB 推定量.

推定手順

推定ステップ

1. 各地域の距離 d_{ik} を計算する.
2. CV 基準を最小化するような最適なバンド幅 b^* を求める.
3. バンド幅 b^* のもとで各地域間のウェイト w_{ik} を計算し、各 i に対して $\widehat{\boldsymbol{\phi}}_i$ を求める.
4. 推定量 $\widehat{\boldsymbol{\phi}}_i$ をベイズ推定量に代入して $\widehat{\mu}_i^{\mathrm{GWEB}}$ を得る.

MSE

- $\widehat{\mu}_i^{\mathrm{GWEB}}$ のリスク評価のために MSE の推定量を求める.
- MSE を以下のように分解.

$$\begin{aligned}
\mathrm{MSE}_i &= E\left[(\widehat{\mu}_i^{\mathrm{GWEB}} - \mu_i)^2\right] \\
&= E\left[(\widetilde{\mu}_i - \mu_i)^2\right] + E\left[(\widehat{\mu}_i^{\mathrm{GWEB}} - \widetilde{\mu}_i)^2\right] \\
&= \frac{\nu_i Q(m_i)}{(n_i + \nu_i)(\nu_i - v_2)} + E\left[(\widehat{\mu}_i^{\mathrm{GWEB}} - \widetilde{\mu}_i)^2\right].
\end{aligned}$$

- 第 1 項は $O(1)$, 第 2 項は $O(n^{-1})$ $(n = mb)$.
- parametric bootstrap を用いて MSE の 2 次漸近不偏推定量を求めることができる.

死亡データへの応用

データセット

- 1930 年東京市の死亡データを解析する.
- z_i: 地域 i の男性の死亡数
- N_i: 地域 i に住んでいる男性の総数
- 地域数は $m = 1372$ $(i = 1, \ldots, m)$.
- 男性の総死亡数は $L = \sum_{i=1}^{m} z_i = 13656$.
- 地域 i の男性の期待死亡数は, $n_i = L \times (N_i / \sum_{j=1}^{m} N_j)$

標準化死亡比（Standardized Mortality Ratio, SMR）

- SMR はある地域における死亡の潜在的な危険性をあらわす指標で, 「実際の死亡数と期待死亡数の比」で定義される.
- 地域 i における SMR の direct estimator は, $y_i = z_i / n_i$.
- しかしながら, 小地域では y_i は信頼できないので, SMR の model based estimator（GWEB 推定量）を求めたい.

死亡データへの応用

空間ポアソン・ガンマ混合モデル

- $z_1, \ldots, z_m$ は独立に以下の分布に従っていると仮定

$$z_i|\mu_i \sim \mathrm{Po}(n_i\mu_i), \quad \mu_i \sim \mathrm{Ga}(\nu_i m_i, \nu_i)$$

- “真の”SMR は μ_i，その direct estimator は $y_i = z_i/n_i$.
- 共変量は $\boldsymbol{x}_i = (x_{0i}, x_{1i})^t$, $x_{0i} = 1$, x_{1i} は地域 i の女性の死亡数,
  $E(\mu_i) = m_i = \exp(\boldsymbol{x}_i^t\boldsymbol{\beta}_i)$.
- 地理情報として隣接行列が得られているので，隣接行列によって定義されるグラフの最短距離として地域間の距離 d_{ik} を計算．これにもとづき
  $\boldsymbol{\phi}_i = (\beta_{0i}, \beta_{1i}, \nu_i)^t$ を推定し，$\widehat{\mu}_i^{\mathrm{GWEB}}$ を導出.

$$\widehat{\mu}_i^{\mathrm{GWEB}} = \frac{n_i}{n_i + \widehat{\nu}_i} y_i + \frac{\widehat{\nu}_i}{n_i + \widehat{\nu}_i} \exp(\boldsymbol{x}_i^t \widehat{\boldsymbol{\beta}}_i)$$

- 比較のため，既存のポアソン・ガンマ混合モデル（$\boldsymbol{\phi}_i = \boldsymbol{\phi}$）も適用.

推定結果

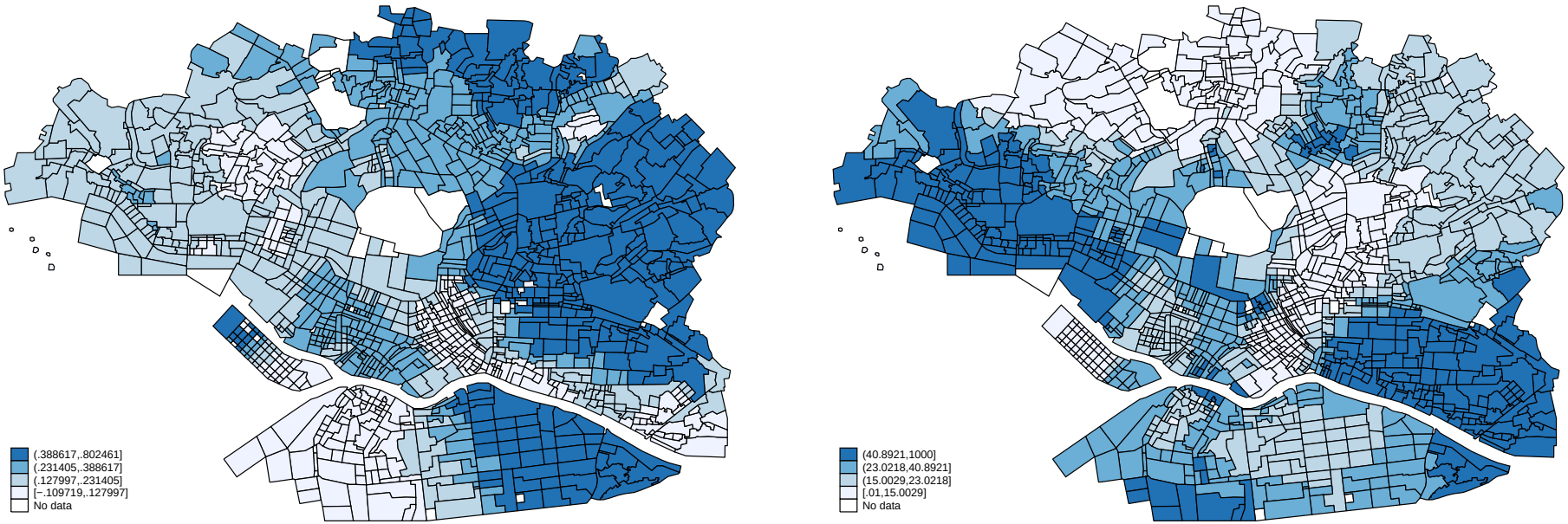


Figure : 推定されたハイパーパラメータの空間分布. 左 : β_{1i}, 右 : ν_i.

推定結果

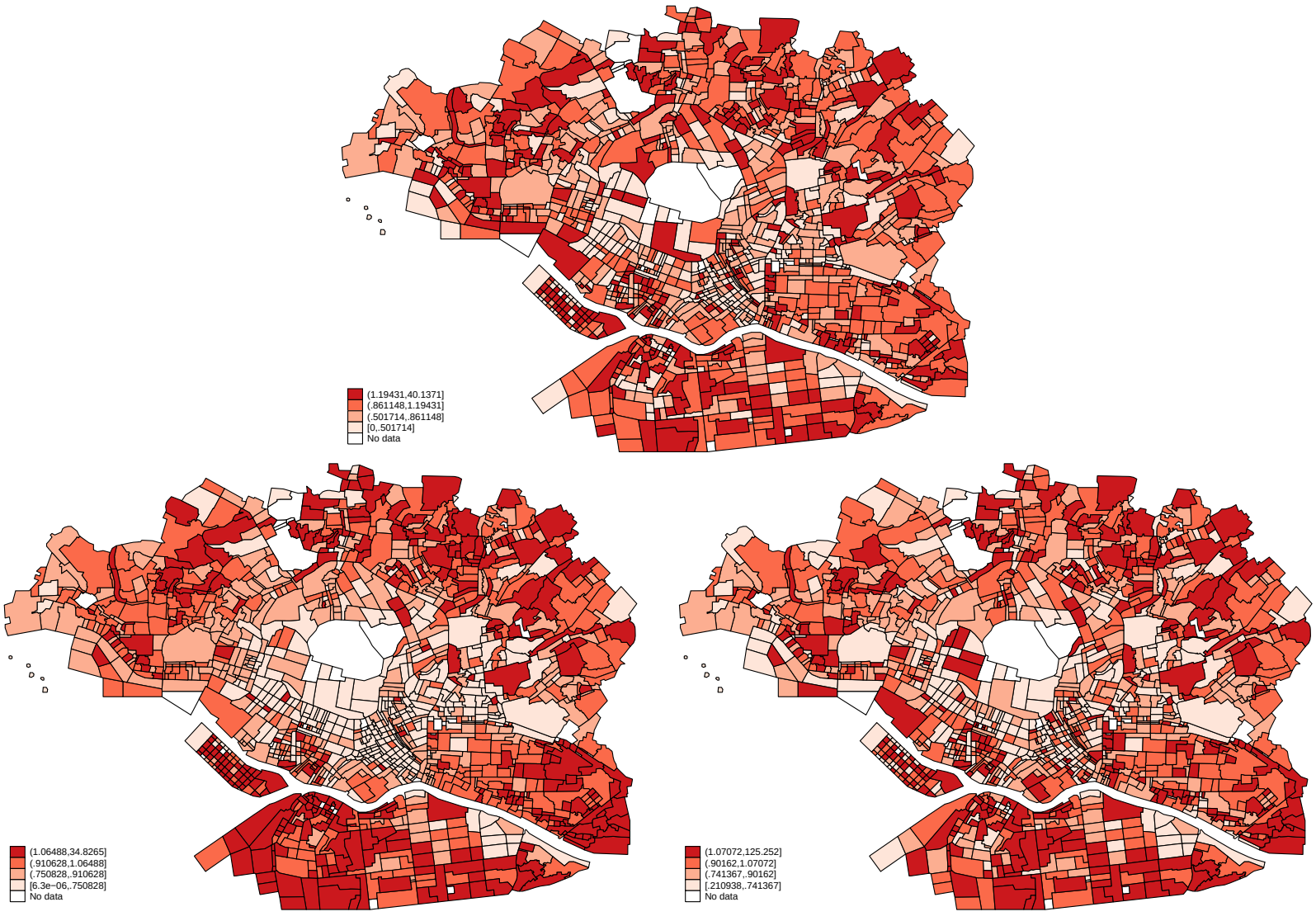


Figure : y_i（上），GWEB（左下），EB (右下)

MSE の比較

既存のポアソン・ガンマ混合モデルと比較するため，提案モデルと既存モデルの MSE の推定値を計算し，direct estimator y_i の MSE との比をそれぞれ計算.

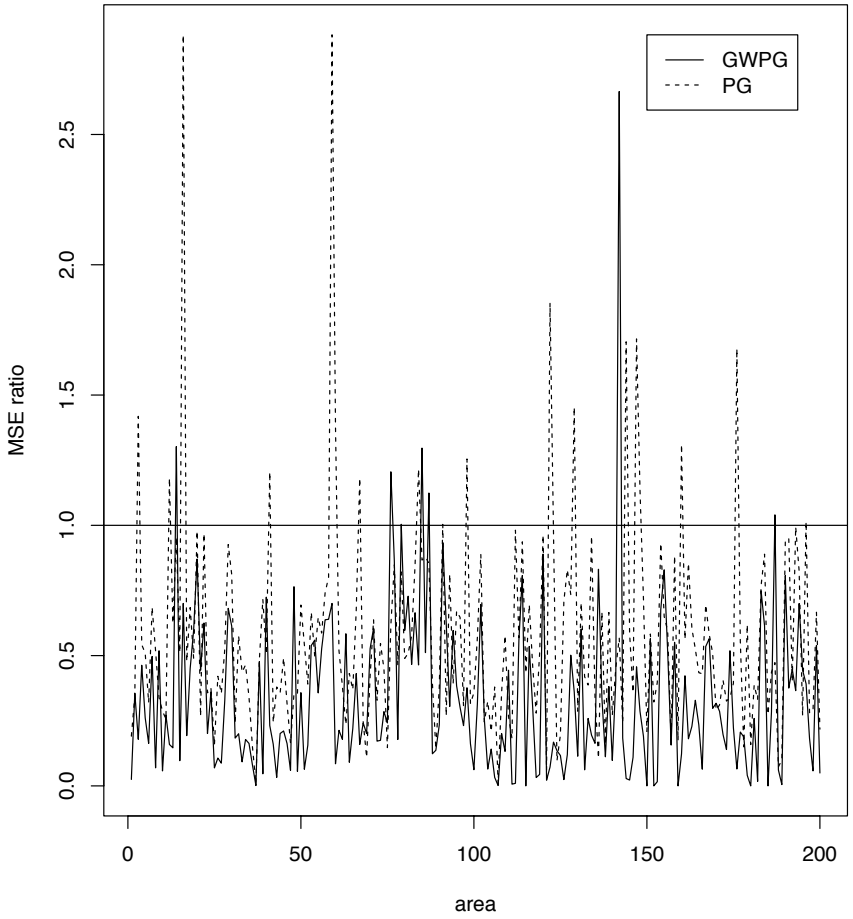


Figure : 200 地域における MSE 比.

まとめ

- 空間非定常性を考慮した小地域推定モデルを考え，そこから得られる model based estimator として GWEB 推定量を提案した.
- 提案モデルは離散データへの応用が可能.
- 例として，空間ポアソン・ガンマ混合モデルを用いて，SMR の小地域推定を行った.