CIRJE-J-171

線形混合モデルと小地域の推定

東京大学大学院経済学研究科
久保川　達也

2006 年 12 月

応用統計学 Vol. 35, No. 3 (2006), 1–24

総 合 報 告

線形混合モデルと小地域の推定

東京大学大学院経済学研究科 久保川達也

要 旨 標本調査では，調査区全体からデータがとられ全体の母集団特性が調べられる。同じデータを用いて市町村レベルの母集団特性を求めようとすると，市町村によっては取られたデータが少ないため標本平均などの推定値は推定誤差が大きくなってしまうという問題が生ずる。これを小地域問題といい，注目している地域（市町村）の周辺地域からデータを上手に取り込むことによって推定精度を高めることができる。そのための代表的なモデルが線形混合モデルであり，そこから導かれる経験最良線形予測量が小地域問題を解決する手法になっている。本稿では，線形混合モデルを利用した小地域推定について解説する。特に，線形混合モデルのもつ（共通母数）＋（変量効果）という構造が推定精度を高めるためにどのように働くのかについて説明し，実際どの程度誤差が抑えられているのかに関して平均２乗誤差の推定と信頼区間の構成についてまとめる。最後に，空間データ等を分析するための様々なモデリングの方法を紹介し，一般化線形混合モデルと死亡率推定への応用についても説明する。

1. はじめに

本稿で扱う問題は，標本調査における小地域の特性値の推定問題に関係している。通常は調査区全体の特性を調べるために標本調査が行われるが，そのデータを利用して地域ごとの特性値を推定したい状況がしばしば生ずる。例えば，得られたデータから各地域への予算配分の仕方を決めたり，政策を決定したりする場合がある。そのとき，狭い地域や人口が粗な地域に対しては十分なデータがとられていないため，その地域のデータだけでは特性値の十分な推測ができない。このような状況での推定問題を，特に小地域推定という。

小地域推定の問題を解決する方法は，周辺地域のデータを組み込んで推定精度を高めることであり，どのような形でデータを取り込むのかが重要な問題となる。そのために利用されるのが線形混合モデル (Linear Mixed Model, LMM) であり，そのモデルから導かれる経験最良線形不偏予測量 (Empirical Best Linear Unibiased Predictor, EBLUP) が小地域の安定した推定値を与えるのに役立つ。LMM は，基本的に共通母数に基づいて回帰する項と地域の差異を表す変量効果の項及び誤差項とから構成されている。すべての地域を通して回帰係数を共通に設定することによってすべてのデータをプールして安定した推定値を与えることができる。しかし，これだけでは地域の特徴や地域による差異を引き出すことができない。そこで地域効果を変量効果としてモデルに取り込む。この効果を予測してやることにより，標本平均を縮小する作用が生ずること

になる。LMM は，母数の共通化によるデータのプーリングと変動効果による標本平均の縮小作用を生み出すことのできるモデルであり，その結果生ずる予測量が EBLUP となる。したがって，EBLUP は，各々の地域の標本平均とプールされた回帰推定量との加重平均になっており，データ数が少ないときには標本平均をプールされた推定値の方向へ縮小することにより，推定精度の改善を図っている。

本稿では，LMM を利用した小地域推定について解説する。まず 2 節では，Battese, Harter and Fuller (1988) で取り上げられた小地域推定問題を通して，LMM の簡単な説明と EBLUP の導出を与え，有限母集団モデルへの拡張について述べる。また，LMM を構成する（共通母数による回帰項）＋（変量効果）という構造が小地域推定にどのように役立つのかについて，上述したような説明を与える。3 節では，予測精度を高めるために導出された EBLUP が実際どの程度推定誤差を改善しているのかについて，平均 2 乗誤差とその推定方法について説明する。また信頼区間の構成を行い，これらを用いた応用例を与える。4 節では，LMM を拡張したり変形することによって様々な問題に適合することができることを説明し，空間モデルや時系列・クロスセクションモデルなどを紹介する。最後に，離散分布への拡張として一般化線形混合モデル (Generalized Linear Mixed Model, GLMM) について紹介し，死亡率推定への応用について簡単に説明する。

小地域ということばは，市，郡，町，村などの地理的に小さい地域を指すばかりではなく，実は広い地域の母集団から特定の年齢，性，人種などに分割されたところの部分母集団を意味することもできる。したがって，論理的には本稿で解説する小地域推定の内容は利用できるデータ数が少ないという様々な応用分野に適応できることに注意しておく。

なお，小地域推定を扱った解説論文や解説書としては Ghosh and Rao (1994), Rao (2003) があり，LMM については，Robinson (1991), McCulloch and Searle (2000), Searle, Casella and McCulloch (1992) があるので，詳しい内容についてはそれらを参照してほしい。

2. 線形混合モデルと小地域推定

2.1. 線形混合モデル

まず，この分野の啓蒙的な論文として知られる Battese *et al*. (1988) が取り上げた小地域推定問題について紹介しよう。アイオワ州の北部中央の 12 の郡について穀物 (とうもろこし及び大豆) の作付面積の調査が以下のようにしてなされた。12 の郡それぞれについてさらに約 250h の農作区画 (segment) に細分し，i 番目の郡における農作区画の個数を N_i とする。連邦農務省は，N_i 個の区画の中から n_i 個の区画をランダムに抽出し，直接農家にインタヴュー調査を行うことによってその区画におけるとうもろこし (または大豆) の作付面積についてのデータを得た。この論文では，i 番目の郡における j 番目の区画のデータを y_{ij} で表すことにする。実際には，n_i は 1∼5 程度で, 12 の郡全体での総数は 36 程度である。他方, 人工衛星 LANDSAT からの観測により, 約 0.45h のピクセル (picture element, 画像から識別する単位) に対していずれの穀物が作付けされているのかが識別され, 12 の郡すべてにわたってこうした衛星データが補助情報として利用可能である。

各郡におけるとうもろこし（または大豆）の平均的作付面積（農作区画単位）を μ_i で表し，

これを推定することが解析の目的である。各郡の標本平均 $\overline{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ を用いて推定するのが最も簡単な方法であるが，n_i が 1〜5 程度であるため推定誤差が大きいという問題がある。Battese *et al.* (1988) は，小地域での推定精度を高めるために，連邦農務省の調査データと衛星データを利用した線形混合モデル (LMM) を用いることを考えた。

各々の区画について衛星データによりとうもろこし及び大豆と識別されたピクセルの個数を x_{1ij}, x_{2ij} で表すと，調査データ y_{ij} と衛星データ x_{1ij}, x_{2ij} との間にはほぼ線形関係が認められるため，

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i \tag{1}$$

なるモデルが想定できる。また誤差項 u_{ij} は，郡によって異なっており，郡に依存する項 v_i と郡の差異に依らない項 e_{ij} に加法的に分解されて

$$u_{ij} = v_i + e_{ij} \tag{2}$$

と表されるとする。$\boldsymbol{x}_{ij} = (1, x_{1ij}, x_{2ij})'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ とおくと，モデル (1) は，

$$y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + v_i + e_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i \tag{3}$$

と表現できる。v_i は地域の差異を表しているので地域効果と呼ばれる。これは未知母数もしくは確率変数として扱われ，それぞれ母数効果，変量効果と呼ばれる。どちらで扱うかは実際の問題に依存しており，特定の地域を扱うのであれば母数効果とみなすのが自然である。しかし，地域効果 v_i の背後に共通な分布が想定できる場合には，変量効果として扱うことができる。小地域の安定した推定は，v_i を変量として扱うことから導かれるので，ここでは，v_i が変量効果とみなせる場合を考える。すなわち，v_i, e_{ij} はすべて互いに独立な確率変数とし，それぞれ正規分布

$$v_i \sim \mathcal{N}(0, \sigma_v^2), \quad e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$$

に従うとして，今後話を進める。

モデルをより簡便に記述するために，行列を用いて表現しておくことにする。上の例では $\boldsymbol{x}_{ij}$, $\boldsymbol{\beta}$ は 3×1 のベクトルであるが，これ以降はより一般的に $m \times 1$ のベクトルとして扱うことにする。ベクトルと行列の大きさを適当に揃えて $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})'$, $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_k')'$, $\boldsymbol{x}_i' = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})$, $\boldsymbol{X} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_k')'$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{m-1})'$ とし，$\boldsymbol{e}$ も $\boldsymbol{y}$ と同様に $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{in_i})'$, $\boldsymbol{e} = (\boldsymbol{e}_1', \ldots, \boldsymbol{e}_k')'$, と定義する。また，すべての成分が 1 の $n_i \times 1$ ベクトルを $\boldsymbol{j}_{n_i}$ で表し，ブロック対角行列 block diag$(\cdot)$ を用いて $\boldsymbol{Z} = \text{block diag}(\boldsymbol{j}_{n_1}, \ldots, \boldsymbol{j}_{n_k})$ とおき，$\boldsymbol{v} = (v_1, \ldots, v_k)'$ とおく。このときモデル (3) は，

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{e} \tag{4}$$

と表すことができる。これを線形混合モデル (LMM) もしくは分散成分モデル (Variance Component Model) という。$\rho = \sigma_v^2/\sigma_e^2$ とおくとき，$\boldsymbol{y}_i$ の共分散行列は $\mathbf{Cov}(\boldsymbol{y}_i) = \sigma_e^2 \boldsymbol{V}_i(\rho)$, $\boldsymbol{V}_i(\rho) = \rho \boldsymbol{J}_{n_i} + \boldsymbol{I}_{n_i}$ と表される。ここで，$\boldsymbol{J}_{n_i} = \boldsymbol{j}_{n_i}\boldsymbol{j}_{n_i}'$ はすべての要素が 1 の $n_i \times n_i$ 行列，$\boldsymbol{I}_{n_i}$ は $n_i \times n_i$ の単位行列である。従って，$\boldsymbol{y}$ の分散共分散行列は

$$\mathbf{Cov}(\boldsymbol{y}) = \sigma_e^2 \boldsymbol{V}(\rho), \quad \boldsymbol{V}(\rho) = \text{block diag}(\boldsymbol{V}_1(\rho), \ldots, \boldsymbol{V}_k(\rho))$$

と書ける。

2.2. 最良線形不偏予測量 (BLUP) と分散成分の推定

それぞれの郡の穀物の平均的作付面積を推定したいので，i 番目の郡については $\overline{\boldsymbol{x}}_i = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$ とおくとき $\mu_i = \overline{\boldsymbol{x}}_i'\boldsymbol{\beta} + v_i$ が推定したい値となる。これは，（母数）＋（変量）の形をしているので，推定というよりは予測という言い方をするのが普通である。

分散成分の比 $\rho = \sigma_v^2/\sigma_e^2$ と $\boldsymbol{\beta}$ が既知の場合, 変量効果 v_i の最適な予測量は $\boldsymbol{y}$ を与えたときの条件付期待値で与えられる。これを計算すると，

$$E\left[v_i|\boldsymbol{y}\right] = \frac{\rho n_i}{1+\rho n_i}(\overline{y}_i - \overline{\boldsymbol{x}}_i'\boldsymbol{\beta})$$

となり，$\tilde{\mu}_i(\boldsymbol{\beta},\rho) = \overline{\boldsymbol{x}}_i'\boldsymbol{\beta} + \rho n_i(1+\rho n_i)^{-1}(\overline{y}_i - \overline{\boldsymbol{x}}_i'\boldsymbol{\beta})$ が μ_i の最適な予測量になる。ここで，母係数 $\boldsymbol{\beta}$ を一般化最小 2 乗 (Generalized Least Squares, GLS) 推定量

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}(\rho) =& (\boldsymbol{X}'\boldsymbol{V}(\rho)^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}(\rho)^{-1}\boldsymbol{y} \\ =& \Big(\sum_{i=1}^{k}\frac{n_i\overline{\boldsymbol{x}}_i\overline{\boldsymbol{x}}_i'}{1+n_i\rho}\Big)^{-1}\sum_{i=1}^{k}\frac{n_i\overline{\boldsymbol{x}}_i\overline{y}_i}{1+n_i\rho} \end{aligned} \tag{5}$$

で推定し，$\tilde{\mu}_i(\boldsymbol{\beta},\rho)$ に代入すると，

$$\widehat{\mu}_i(\rho) = \bar{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}}(\rho) + \frac{\rho n_i}{1+\rho n_i}(\bar{y}_i - \bar{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}}(\rho)) \tag{6}$$

なる形の線形不偏な予測量が得られる。これは，線形不偏な予測量の中で最適なものになっていることが示されるので， μ_i の最良線形不偏予測量 (Best Linear Unbiased Predictor, BLUP) と呼ばれる。

次に，分散成分 σ_v^2, σ_e^2 もしくはそれらの比 ρ を推定しよう。繰り返し数が異なる場合には最尤推定量 (MLE) や制限付最尤推定量 (Restricted ML, REML) を明示的に表現できないので，ここでは不偏推定量に修正を施した推定量を用いることにする。まず，σ_e^2 の不偏推定量は，

$$\hat{\sigma}_e^{2UB} = \frac{S_1}{N-k-m+\lambda}, \quad S_1 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big\{(y_{ij}-\overline{y}_i) - (\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i)'\widehat{\boldsymbol{\beta}}_1\Big\}^2 \tag{7}$$

で与えられる。ただし，$m-\lambda$ は行列 $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i)'$ のランクを表しており，通常は線形モデル (4) が定数項をもつときには $\lambda = 1$, もたないときには $\lambda = 0$ となる。また $\widehat{\boldsymbol{\beta}}_1$ は $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\{(y_{ij}-\overline{y}_i) - (\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i)'\boldsymbol{\beta}\}^2$ における $\boldsymbol{\beta}$ の最小 2 乗推定量を表している。$\hat{\sigma}_e^{2UB}$ は $\overline{y}_1, \ldots, \overline{y}_k$ と独立に分布し，

$$\begin{aligned} S_1/\sigma_e^2 &\sim \chi^2_{N-k-m+\lambda}, \\ \overline{y}_i &\sim \mathcal{N}(\overline{\boldsymbol{x}}_i'\boldsymbol{\beta}, \sigma_e^2/n_i + \sigma_v^2), \quad i = 1, \ldots, k, \end{aligned} \tag{8}$$

に従うことがわかる。σ_v^2 については，2 乗和 $\sum_{i=1}^{k} n_i(\overline{y}_i - \overline{\boldsymbol{x}}_i'\boldsymbol{\beta})^2$ の残差平方和に基づいて推定することが考えられる。この 2 乗和における $\boldsymbol{\beta}$ の最小 2 乗推定量は $\widehat{\boldsymbol{\beta}}_2 = (\sum_{j=1}^{k} n_j\overline{\boldsymbol{x}}_j\overline{\boldsymbol{x}}_j')^{-1}\sum_{j=1}^{k} n_j\overline{\boldsymbol{x}}_j\overline{y}_j$ で与えられるので，

$$S_2 = \sum_{i=1}^{k} n_i (\overline{y}_i - \overline{\boldsymbol{x}}_i' \widehat{\boldsymbol{\beta}}_2)^2 \tag{9}$$

と書ける。S_2 の期待値は,

$$N = \sum_{i=1}^{k} n_i, \quad N_* = N - \mathrm{tr}\left(\sum_{i=1}^{k} n_i^2 \overline{\boldsymbol{x}}_i \overline{\boldsymbol{x}}_i'\right)\left(\sum_{i=1}^{k} n_i \overline{\boldsymbol{x}}_i \overline{\boldsymbol{x}}_i'\right)^{-1}$$

とおくと，$E[S_2] = (k-m)\sigma_e^2 + N_*\sigma_v^2$ と書けるので，σ_v^2 の 1 つの不偏推定量は,

$$\hat{\sigma}_v^{2UB} = \frac{1}{N_*}\left\{S_2 - (k-m)\hat{\sigma}_e^{2UB}\right\} \tag{10}$$

で与えられることがわかる。

さて，分散成分 σ_v^2 の不偏推定量 $\hat{\sigma}_v^{2UB}$ の欠点は正の確率で負の値をとってしまうことである。そこで，Kubokawa (2000), Kubokawa, Saleh and Konno (2000) で提案された σ_e^2, σ_v^2 の打ち切り推定量を用いることにする。具体的には,

$$\hat{\sigma}_e^2 = \min\left\{\hat{\sigma}_e^{2UB}, \frac{(N-k)\hat{\sigma}_e^{2UB} + (k-m)(k-m+2)^{-1}S_2}{N-m}\right\}, \tag{11}$$

$$\begin{aligned}\hat{\sigma}_v^2 =& \frac{k-m}{N_*}\max\left\{\frac{S_2}{(k-m)\hat{\sigma}_e^{2UB}} - 1, \frac{2}{k-m}\right\} \\ & \times \min\left\{\hat{\sigma}_e^{2UB}, \frac{(N-k)\hat{\sigma}_e^{2UB} + (k-m)(k-m+2)^{-1}S_2}{N-m}\right\}\end{aligned} \tag{12}$$

で与えられており，不偏推定量に対して理論的優越性が保証されている。$\hat{\sigma}_v^2$ はほとんど至るところで正の推定値を与えることができる。これらを用いると，分散成分の比 ρ の推定量は,

$$\hat{\rho} = \hat{\sigma}_v^2/\hat{\sigma}_e^2 = \frac{k-m}{N_*}\max\left\{\frac{S_2}{(k-m)\hat{\sigma}_e^{2UB}} - 1, \frac{2}{k-m}\right\} \tag{13}$$

となることがわかる。上述の推定量は若干複雑な形をしているようにみえるが，実は場合分けは 1 通りしかなく，$S_2/\hat{\sigma}_e^{2UB} > k-m+2$ のときには，$(\hat{\sigma}_e^2, \hat{\sigma}_v^2)$ は不偏推定量 $(\hat{\sigma}_e^{2UB}, \hat{\sigma}_v^{2UB})$ に一致し，$S_2/\hat{\sigma}_e^{2UB} \le k-m+2$ のときには，$\hat{\sigma}_e^2 = \{(N-k)\hat{\sigma}_e^{2UB} + (k-m)(k-m+2)^{-1}S_2\}/(N-m)$, $\hat{\sigma}_v^2 = \hat{\sigma}_e^2\hat{\rho}$, $\hat{\rho} = 2/N_*$ となっている。結局，この $\hat{\rho}$ を (6) に代入することにより，予測量

$$\widehat{\mu}_i(\hat{\rho}) = \overline{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}}(\hat{\rho}) + \frac{\hat{\rho}n_i}{1+\hat{\rho}n_i}(\overline{y}_i - \overline{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}}(\hat{\rho})) \tag{14}$$

が得られる。これを，経験最良線形不偏予測量 (EBLUP) という。

2.3. 有限母集団モデルの枠組み

小地域の推定問題は，官庁統計の分野で利用されることが多く，その場合には有限母集団の枠組みで議論される。ここでは，Battese *et al.* (1988) に沿って超母集団の設定のもとで小地域推定の方法について説明する。

2.1 節の始めで紹介した問題設定を思い出そう。第 i 郡の農作区画の総数を N_i，第 i 郡，第 j

区画の穀物の作付け面積を Y_{ij} とし，Y_{ij} に線形混合モデル

$$Y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + v_i + e_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, N_i$$

を当てはめる。有限母集団の枠組みでは，i 番目の郡について，N_i 個の母集団の中から n_i 個の標本をランダムに抽出して，母集団平均 $\mu_i = \overline{Y}_i = \sum_{j=1}^{N_i} Y_{ij}/N_i$ を推定することが目的となる。いま i 番目の郡に対して，N_i 個のデータ $Y_{i1}, \ldots, Y_{iN_i}$ から抽出された n_i 個のデータの組を，簡単のために $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})'$ とし，抽出されなかったデータの組を $\boldsymbol{Y}_i^* = (Y_{i,n_i+1}, \ldots, Y_{i,N_i})'$ と書き，$\boldsymbol{Y}_i^P = (\boldsymbol{y}_i', \boldsymbol{Y}_i^{*\prime})'$ と置く。$\boldsymbol{y}_i$ を与えたときの $\boldsymbol{Y}_i^P$ の条件付期待値 $E[\boldsymbol{Y}_i^P|\boldsymbol{y}_i]$ を用いて $\boldsymbol{Y}_i^P$ を推定するのが最も自然と考えられる。この条件付期待値を具体的に求めるために $(\boldsymbol{y}_i', \boldsymbol{Y}_i^{*\prime})'$ の同時密度関数を書き表すと，

$$\begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{Y}_i^* \end{pmatrix} \sim \mathcal{N}_{N_i}\left(\begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{x}_i^* \end{pmatrix}\boldsymbol{\beta}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right)$$

となる。ただし，$\boldsymbol{x}_i^* = (\boldsymbol{x}_{i,n_i+1}', \ldots, \boldsymbol{x}_{i,N_i}')'$ とし，また n_i 次元ベクトル $\boldsymbol{j}_i = (1, \ldots, 1)'$, $N_i - n_i$ 次元ベクトル $\boldsymbol{j}_i^* = (1, \ldots, 1)'$ を用いると，分散共分散行列の各行列要素は

$$\boldsymbol{\Sigma}_{11} = \sigma_v^2 \boldsymbol{j}_i \boldsymbol{j}_i' + \sigma_e^2 \boldsymbol{I}_{n_i},\ \ \boldsymbol{\Sigma}_{12} = \sigma_v^2 \boldsymbol{j}_i \boldsymbol{j}_i^{*\prime},\ \ \boldsymbol{\Sigma}_{22} = \sigma_v^2 \boldsymbol{j}_i^* \boldsymbol{j}_i^{*\prime} + \sigma_e^2 \boldsymbol{I}_{N_i - n_i}$$

と表される。したがって $\boldsymbol{y}_i$ を与えたときの $\boldsymbol{Y}_i^*$ の条件付分布は

$$\boldsymbol{Y}_i^*|\boldsymbol{y}_i \sim \mathcal{N}_{N_i - n_i}\left(\boldsymbol{x}_i^{*\prime}\boldsymbol{\beta} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i'\boldsymbol{\beta}), \boldsymbol{\Sigma}_{22} - \Sigma_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right)$$

となる。ここで，簡単な計算から，

$$\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} = \frac{\rho}{1 + n_i\rho}\boldsymbol{j}_i^* \boldsymbol{j}_i'$$

となるので，

$$E[\boldsymbol{Y}_i^*|\boldsymbol{y}_i] = \boldsymbol{x}_i^{*\prime}\boldsymbol{\beta} + \frac{n_i\rho}{1 + n_i\rho}\boldsymbol{j}_i^*(\bar{y}_i - \bar{\boldsymbol{x}}_i'\boldsymbol{\beta})$$

となる。したがって，

$$E\Big[\sum_{j=n_i+1}^{N_i} Y_{ij}\Big|\boldsymbol{y}_i\Big] = E\left[\boldsymbol{j}_i^{*\prime}\boldsymbol{Y}_i^*|\boldsymbol{y}_i\right] = (N_i - n_i)\left\{\bar{\boldsymbol{x}}_i^{*\prime}\boldsymbol{\beta} + \frac{n_i\rho}{1 + n_i\rho}(\bar{y}_i - \bar{\boldsymbol{x}}_i'\boldsymbol{\beta})\right\}$$

と表される。ただし，$\bar{\boldsymbol{x}}_i^* = (N_i - n_i)^{-1}\sum_{j=n_i+1}^{N_i}\boldsymbol{x}_{ij}$ である。この $\boldsymbol{\beta}$ に一般化最小２乗推定量 $\widetilde{\boldsymbol{\beta}}(\hat{\rho})$ を代入すると，母集団平均 $\mu_i = \bar{Y}_i$ の予測量は，

$$\begin{aligned}\tilde{\mu}_i(\hat{\rho}) =& E[\bar{Y}_i|\boldsymbol{y}_i] = \frac{n_i}{N_i}\bar{y}_i + \frac{1}{N_i}E\Big[\sum_{j=n_i+1}^{N_i} Y_{ij}\Big|\boldsymbol{y}_i\Big] \\ =& \frac{n_i}{N_i}\bar{y}_i + \frac{N_i - n_i}{N_i}\left\{\bar{\boldsymbol{x}}_i^{*\prime}\widetilde{\boldsymbol{\beta}}(\hat{\rho}) + \frac{n_i\hat{\rho}}{1 + n_i\hat{\rho}}(\bar{y}_i - \bar{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}}(\hat{\rho}))\right\} \qquad (15)\end{aligned}$$

で与えられる。$\overline{\boldsymbol{x}}_{i(p)} = \sum_{j=1}^{N_i}\boldsymbol{x}_{ij}$ とおくと，この予測量は，

$$\tilde{\mu}_i(\hat{\rho}) = \overline{\boldsymbol{x}}'_{i(p)}\widetilde{\boldsymbol{\beta}}(\hat{\rho}) + \left\{\frac{n_i\hat{\rho}}{1+n_i\hat{\rho}} + \frac{n_i}{N_i}\frac{1}{1+n_i\hat{\rho}}\right\}\left(\overline{y}_i - \overline{\boldsymbol{x}}'_i\widetilde{\boldsymbol{\beta}}(\hat{\rho})\right)$$

と表されるので，n_i に比べて N_i がかなり大きいときには，$\overline{\boldsymbol{x}}'_{i(p)}\widetilde{\boldsymbol{\beta}}(\hat{\rho})+n_i\hat{\rho}(1+n_i\hat{\rho})^{-1}(\overline{y}_i-\overline{\boldsymbol{x}}'_i\widetilde{\boldsymbol{\beta}}(\hat{\rho}))$ となり，n_i が N_i に近いときには，(14) に近づくことがわかる。

有限母集団モデルについての議論は現実的であり重要であるがかなり複雑な問題設定になるため，我が国ではあまり活発には研究されていない。Ghosh and Meeden (1997) は有限母集団モデルについてベイズ法に基づいた現代的なアプローチを解説しており，わかりやすい文献である。

2.4. 予測精度を高めるためのモデリング

(14) で与えられる経験最良線形不偏予測量 (EBLUP) を

$$\widehat{\mu}_i(\hat{\rho}) = \frac{\hat{\rho}n_i}{1+\hat{\rho}n_i}\bar{y}_i + \frac{1}{1+\hat{\rho}n_i}\bar{\boldsymbol{x}}'_i\widetilde{\boldsymbol{\beta}}(\hat{\rho})$$

と変形すると，$\overline{y}_i$ と $\overline{\boldsymbol{x}}'_i\widetilde{\boldsymbol{\beta}}(\hat{\rho})$ との加重平均になっていることがわかる。$\overline{y}_i$ は直接取られたデータに基づいた平均値であるので，個々の郡の特徴を反映していると考えられる。しかし n_i が小さいときには，$\overline{y}_i$ の予測誤差が問題となる。他方，$\overline{\boldsymbol{x}}'_i\widetilde{\boldsymbol{\beta}}(\hat{\rho})$ は全データに基づいて構成されているので安定しているが，郡の特徴は $\overline{y}_i$ ほど強くは現れないと思われる。EBLUP は，これらの点を考慮した方法であり，n_i もしくは $\hat{\rho}$ が小さければ $\overline{y}_i$ を $\overline{\boldsymbol{x}}'_i\widetilde{\boldsymbol{\beta}}(\hat{\rho})$ の方向へ縮小することによって安定化を図っている。すなわち，n_i が小さければ，データの不足を周辺もしくは全体のデータで補うことによって予測精度を高めていると解釈される。このようにして EBLUP が小地域の予測問題に役立つことがわかる。言い換えれば，EBLUP を生み出すところのモデルの形に，小地域の予測を効果的に行う仕組みが備わっていることになる。この論文で取り上げている線形混合モデル (3), (4) は，

（データ）＝（共通母数）＋（変量効果）＋（誤差項）

の形をしており，(共通母数）と（変量効果）が，安定した推定量を得るためそれぞれ次のような役目を演じていることになる。

[1] 変量効果と縮小推定. 説明を簡単にするために，モデル (3) において $\boldsymbol{\beta}=\mathbf{0}$ としてみよう。もし v_i が母数効果であるときには，v_i はそれぞれの郡のみに基づいた標本平均 $\overline{y}_i$ で推定されるので，推定誤差が問題になるだけでなく，郡の個数 k が大きいときには未知母数が多くなってしまい，情報量規準などのモデル選択の観点からも好ましくない。そこで v_i を変量にしてみると，2.2 節で説明したように，v_i は条件付期待値 $E[v_i|\boldsymbol{y}] = \rho n_i(1+\rho n_i)^{-1}\overline{y}_i = \overline{y}_i-(1+\rho n_i)^{-1}\overline{y}_i$ によって予測できる。これに ρ の推定値を代入したものは，$\overline{y}_i$ を縮小する推定量の形をしており，いわゆる Stein 効果によって $\overline{y}_i$ のリスクの改善がなされる。$\boldsymbol{\beta}\neq\mathbf{0}$ のときには，$E[v_i|\boldsymbol{y}] = \rho n_i(1+\rho n_i)^{-1}(\overline{y}_i-\overline{\boldsymbol{x}}'_i\boldsymbol{\beta})$ となるので，$\overline{y}_i$ が安定した推定量の方向へ縮小されることになる。こうして，線形混合モデルにおいて変量効果が，$\overline{y}_i$ を縮小する作用を生むことがわかる。

[2] 共通母数によるデータのプーリング. モデル (3) からわかるように y_{ij} の期待値は $E[y_{ij}] = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$ であり，これは i に依存しない共通な母数 $\boldsymbol{\beta}$ に基づいている。$\boldsymbol{\beta}$ は全データ $\boldsymbol{y}$

の加重平均 $\widetilde{\boldsymbol{\beta}}(\hat{\rho})$ により推定されるので，y_{ij} の期待値は $\boldsymbol{x}'_{ij}\widetilde{\boldsymbol{\beta}}(\hat{\rho})$ で推定されることになる。すなわち，母数を共通にとることによってデータをプーリングする作用が働き，結果として安定した推定が可能になる。このように，主要項の母数に等号制約や順序制約などの制約をいれることによって安定した推定がなされ，また変量効果と併せると $\overline{y}_i$ をその安定化された推定量の方向へ縮小することができ，推定精度を高めることができる。

以上の考え方は，経験ベイズ法の枠組みで Efron and Morris (1975) の一連の論文の中で示されてきたものであり，ベイズ的アプローチの現実的な有用性は基本的には上述の考え方に基づいている。

3. 予測誤差の評価

2 節で説明されたように EBLUP は予測精度を高める手法として用いられてきた。しかし，予測誤差がどの程度に押さえられているのかを見積もることは，データ解析の現場では重要なことであろう。この節では，EBLUP の平均 2 乗誤差の推定と EBLUP に基づいた信頼区間の構成を行い，実際の応用例に当てはめることにする。簡単のために，ここでは誤差分散 σ_e^2 を既知と仮定した Fay-Herriot モデルを用いて説明する。σ_e^2 が未知母数のときの平均 2 乗誤差の推定及び信頼区間の構成については，それぞれ Prasad and Rao (1990), 笹瀬-久保川 (2005) を参照されたい。また，ここではテーラー展開に基づいたアプローチを採用しているが，ブートストラップ法やジャックナイフ法を用いた方法も提案されている (Lahiri (2003))。

3.1. 平均 2 乗誤差の推定

説明をわかりやすくするために，この節では，誤差分散 σ_e^2 を既知とする Fay and Herriot (1979) のモデル

$$\overline{y}_i = \overline{\boldsymbol{x}}'_i\boldsymbol{\beta} + v_i + e_i, \quad i = 1, \ldots, k, \tag{16}$$

を扱うことにする。ここで $e_i \sim \mathcal{N}(0, \sigma_e^2/n_i)$ である。これは標本平均のモデルとして (3) から導かれるが，官庁から発行される数値には集計データが多いことから，むしろ Fay-Herriot モデル (16) を用いる方がよい場合もある。(16) は地域レベルのデータに基づいているので地域レベルモデルといい，(3) を個体レベルモデルという。地域レベルモデルについては，σ_e^2 を何らかの方法で推定してやる必要があることに注意する。モデル (16) において σ_v^2 及び 分散成分比 $\rho = \sigma_v^2/\sigma_e^2$ は $(\overline{y}_1, \ldots, \overline{y}_k)$ に基づいて推定される。ρ の推定量を $\hat{\rho} = \hat{\rho}(\overline{y}_1, \ldots, \overline{y}_k)$ で表すことにする。(14) で与えられる予測量 EBLUP は，この $\hat{\rho}$ を用いて

$$\widehat{\mu}_i(\hat{\rho}) = \overline{y}_i - \hat{\gamma}_i(\bar{y}_i - \bar{\boldsymbol{x}}'_i\widetilde{\boldsymbol{\beta}}(\hat{\rho})), \quad \hat{\gamma}_i = \gamma_i(\hat{\rho}) = (1 + n_i\hat{\rho})^{-1}$$

なる形で表現できる。EBLUP の $\mu_i = \overline{\boldsymbol{x}}'_i\boldsymbol{\beta} + v_i$ に対する推定誤差として，平均 2 乗誤差を σ_e^2 で割ったもの

$$M_i(\rho, \widehat{\mu}_i(\hat{\rho})) = E\left[\{\widehat{\mu}_i(\hat{\rho}) - \mu_i\}^2\right]/\sigma_e^2 \tag{17}$$

を用いる。これは標準化平均 2 乗誤差と呼ぶべきものであるが，ここでは混乱がない限りこの誤

差を平均2乗誤差 (Mean Squared Error, MSE) と呼ぶことにする。この MSE を推定することによって予測量 EBLUP がどの程度の誤差があるのかを見積もることができる。この節では，Datta, Kubokawa and Rao (2002) で導かれた結果を紹介し，詳しい内容についてはその論文を参照されたい。

まず，$\overline{y}_i$ が $\mathcal{N}(\overline{\boldsymbol{x}}_i'\boldsymbol{\beta}, \sigma_e^2/n_i + \sigma_v^2)$ に従うことに注意し，縮小推定の理論において知られている Stein の等式を用いると，EBLUP の MSE に対して正確な不偏推定量が得られる。それは，簡単のために $\hat{\gamma}_i = \gamma_i(\hat{\rho})$, $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}(\hat{\rho})$ とおくと，

$$\begin{aligned}\widehat{M}_i^E(\hat{\rho}) =& n_i^{-1} - 2n_i^{-1}\hat{\gamma}_i + 2\hat{\gamma}_i^2\left\{\overline{y}_i - \overline{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}}\right\}\frac{\partial\hat{\rho}}{\partial\overline{y}_i} + \hat{\gamma}_i^2(\overline{y}_i - \overline{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}})^2 \\ &+ 2\hat{\gamma}_i\overline{\boldsymbol{x}}_i^t\Big\{\sum_{j=1}^k \hat{\gamma}_j n_j \overline{\boldsymbol{x}}_j\overline{\boldsymbol{x}}_j'\Big\}^{-1}\Big\{\hat{\gamma}_i\overline{\boldsymbol{x}}_i - n_i^{-1}\sum_{j=1}^k \hat{\gamma}_j^2 n_j\overline{\boldsymbol{x}}_j\overline{y}_j + n_i^{-1}\sum_{j=1}^k \hat{\gamma}_j^2 n_j\overline{\boldsymbol{x}}_j\overline{\boldsymbol{x}}_j'\widetilde{\boldsymbol{\beta}}\frac{\partial\hat{\rho}}{\partial\overline{y}_i}\Big\}\end{aligned}$$

で与えられる。いま σ_e^2 が既知であることに注意すると，(8) から (10) と同様にして σ_v^2 の不偏推定量 $\hat{\sigma}_v^{2UB} = \{S_2 - (k-m)\sigma_e^2\}/N_*$ を得ることができる。2.2 節の議論をふまえると，ρ の推定量として

$$\hat{\rho} = \frac{k-m}{N_*}\max\left\{\frac{S_2}{(k-m)\sigma_e^2} - 1, \frac{2}{k-m}\right\} \tag{18}$$

を考えることができる。この場合には，$\partial\hat{\rho}/\partial\overline{y}_i$ は

$$\frac{\partial\hat{\rho}}{\partial\overline{y}_i} = \begin{cases} 0 & \text{if} \quad \hat{\rho} = 2/N_*, \\ 2n_i(\overline{y}_i - \overline{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}_2)/(N_*\sigma_e^2) & \text{if} \quad \hat{\rho} > 2/N_* \end{cases}$$

で与えられるので，これを $\widehat{M}_i^E(\hat{\rho})$ へ代入すると，正確な不偏推定値が計算できる。

正値な関数を推定するとき正確な不偏推定量が必ずしもよいとは限らないことは一般によくあることであるが，このことは MSE の正確な不偏推定量に対しても当てはまる。すなわち，その不偏推定量は正の確率で負の値を取ってしまうだけでなく，実際の推定値のバラツキは比較的大きくなることが知られている。Datta *et al.* (2002) でも $\widehat{M}_i^E(\hat{\rho})$ の推定誤差が大きいことを数値実験を通して検証している。そこで，MSE を漸近的に近似した推定量を求めることにしよう。小地域推定の問題を扱っているので各郡の標本サイズ n_i は小さいため，郡の個数 k が大きい場合を考えて k についての2次漸近近似を求めることにする。$\hat{\rho}$ の分散 $V_{\hat{\rho}}(\rho) = E[(\hat{\rho} - \rho)^2]$ に対して，

$$\begin{aligned} g_{1i}(\rho) =& n_i^{-1} - n_i^{-1}\gamma_i(\rho), \\ g_{2i}(\rho) =& \{\gamma_i(\rho)\}^2\overline{\boldsymbol{x}}_i'\Big\{\sum_{j=1}^k \gamma_j n_j\overline{\boldsymbol{x}}_j\overline{\boldsymbol{x}}_j'\Big\}^{-1}\overline{\boldsymbol{x}}_i, \\ g_{3i}(\rho) =& n_i^{-1}\gamma_i(\rho)V_{\hat{\rho}}(\rho), \end{aligned}$$

とおくと，EBLUP の MSE の k に関する2次近似は

$$M_i(\rho, \widehat{\mu}_i(\hat{\rho})) = g_{1i}(\rho) + g_{2i}(\rho) + g_{3i}(\rho) + o(k^{-1})$$

となる。$\hat{\rho}$ のバイアスを $b_{\hat{\rho}}(\rho) = E[\hat{\rho} - \rho]$ とすると，MSE の 2 次近似に基づいて MSE の 2 次漸近不偏推定量を構成することができ，

$$\widehat{M}_i^U(\hat{\rho}) = g_{1i}(\hat{\rho}) + g_{2i}(\hat{\rho}) + 2g_{3i}(\hat{\rho}) - b_{\hat{\rho}}(\hat{\rho})\left\{\gamma_i(\hat{\rho})\right\}^2 \tag{19}$$

で与えられる。実際，$E[\widehat{M}_i^U(\hat{\rho})] = M_i(\rho, \widehat{\mu}_i(\hat{\rho})) + o(k^{-1})$ が成り立つ。(18) で与えられる $\hat{\rho}$ に対しては，$b_{\hat{\rho}}(\rho) = 0 + o(k^{-1})$,

$$V_{\hat{\rho}}(\rho) = 2\sum_{i=1}^{k}(1 + n_i\rho)^2/N^2 + o(k^{-1})$$

となるので，これらを $\widehat{M}_i^U(\hat{\rho})$ に代入すると，MSE の 2 次漸近不偏推定値が計算できる。

実際の場面で推定誤差を評価する際に，通常の MSE ではなく条件付 MSE を用いた方がよいという議論が近年なされている。例えば，i 番目の郡のデータ $y_{i1}, \ldots, y_{in_i}$ が得られ，その小地域に対する平均値 $\overline{y}_i$ が与えられた時点で，EBLUP を使用したときにどの程度の推定誤差があるのかを見積もりたい場合がある。Booth and Hobert (1998) は，特に離散的な一般化線形混合モデルに対しては，通常の MSE よりも条件付 MSE を用いた方がよいことを指摘している。$\overline{y}_i$ を所与としたときの EBLUP の条件付 MSE は，

$$M_i^C(\rho, \widehat{\mu}_i(\hat{\rho}) \,|\, \overline{y}_i) = E\left[\{\widehat{\mu}_i(\hat{\rho}) - \mu_i\}^2 \,|\, \overline{y}_i\right]/\sigma_e^2$$

で定義される。$g_{3i}^C(\rho \,|\, \overline{y}_i - \overline{\boldsymbol{x}}_i'\boldsymbol{\beta}) = \{\gamma_i(\rho)\}^4 n_i^2(\overline{y}_i - \overline{\boldsymbol{x}}_i'\boldsymbol{\beta})^2 V_{\hat{\rho}}(\rho)$ とし，

$$E\left[\frac{\partial\hat{\rho}}{\partial\overline{y}_j}\Big|\overline{y}_i\right] = o_p(k^{-1}) \quad \text{for} \quad j \neq i, \tag{20}$$

を仮定すると，条件付 MSE の 2 次近似は

$$M_i^C(\rho, \widehat{\mu}_i(\hat{\rho}) \,|\, \overline{y}_i) = g_{1i}(\rho) + g_{2i}(\rho) + g_{3i}^C(\rho \,|\, \overline{y}_i - \overline{\boldsymbol{x}}_i'\boldsymbol{\beta}) + o_p(k^{-1}). \tag{21}$$

で与えられる。さらに，$\hat{\rho}$ の条件付バイアスを $b_{\hat{\rho}}(\rho \,|\, \overline{y}_i) = E[\hat{\rho} - \rho \,|\, \overline{y}_i]$ とおくと，条件付 MSE の 2 次漸近不偏推定量は，

$$\widehat{M}_i^{CU}(\hat{\rho} \,|\, \overline{y}_i) = g_{1i}(\hat{\rho}) + g_{2i}(\hat{\rho}) + \left\{\gamma_i(\hat{\rho})n_i\{\overline{y}_i - \overline{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}}(\hat{\rho})\}^2 + 1\right\} g_{3i}(\hat{\rho}) - \{\gamma_i(\hat{\rho})\}^2 b_{\hat{\rho}}(\hat{\rho} \,|\, \overline{y}_i), \tag{22}$$

となる。実際，条件 (20) のもとで $E[\widehat{M}_i^{CU}(\hat{\rho} \,|\, \overline{y}_i) \,|\, \overline{y}_i] = M_i^C(\rho, \widehat{\mu}_i(\hat{\rho})) \,|\, \overline{y}_i) + o_p(k^{-1})$ が成り立つ。(18) で与えられる $\hat{\rho}$ に対しては，

$$b_{\hat{\rho}}(\rho \,|\, \overline{y}_i) = \frac{1}{N}\left\{n_i(\overline{y}_i - \overline{\boldsymbol{x}}_i'\boldsymbol{\beta})^2 - (1 + n_i\rho)\right\} + o_p(k^{-1})$$

となっている。

様々な線形混合モデルに対して EBLUP の MSE, 条件付 MSE の推定が議論されており，詳しくは，Prasad and Rao (1990), Booth and Hobert (1998), Datta and Lahiri (2000), Datta *et al*. (2002), Rao (2003) とその中で引用されている文献が参照される。

3.2. 信頼区間の構成

経験最良線形不偏予測量 EBLUP の誤差評価に関するもう 1 つの方向性は，EBLUP に基づいた信頼区間を構成することである。この小節でも σ_e^2 は既知として話を進める。まず，i 番目の小地域の平均 $\mu_i = \overline{\boldsymbol{x}}_i'\boldsymbol{\beta} + v_i$ に対する簡単な信頼区間は，μ_i を与えたときの $\overline{y}_i$ の条件付分布が $\overline{y}_i|\mu_i \sim \mathcal{N}(\mu_i, \sigma_e^2/n_i)$ であることに注意すると，

$$I_i^* \ : \ \overline{y}_i \pm z_{\alpha/2}\sqrt{\sigma_e^2/n_i} \tag{23}$$

で与えられる。ただし，$z_{\alpha/2}$ は標準正規分布の上側 $\alpha/2$ 点を表す。これは信頼係数 $1-\alpha$ の信頼区間となるが，n_i が小さいときには $\overline{y}_i$ のバラツキが大きくなってしまうとともに信頼区間の長さが長くなってしまう。

そこで，μ_i のベイズ的信頼区間を導くことが考えられる。μ_i のベイズ推定量は，$\gamma_i = (1+n_i\rho)^{-1}$ に対して $\widehat{\mu}_i^B(\boldsymbol{\beta},\rho) = \overline{\boldsymbol{x}}_i'\boldsymbol{\beta} + (1-\gamma_i)(\overline{y}_i - \overline{\boldsymbol{x}}_i'\boldsymbol{\beta})$ と表される。$\overline{y}_i$ を所与としたときの μ_i の条件付分布は，

$$\mu_i|\overline{y}_i \sim \mathcal{N}\left(\widehat{\mu}_i^B(\boldsymbol{\beta},\rho), (\sigma_e^2/n_i)(1-\gamma_i)\right)$$

で与えられるので，信頼係数 $1-\alpha$ のベイズ的信頼区間は，

$$I_i^B(\boldsymbol{\beta},\rho) : \ \widehat{\mu}_i^B(\boldsymbol{\beta},\rho) \pm z_{\alpha/2}\sqrt{(\sigma_e^2/n_i)(1-\gamma_i)}$$

と書ける。これは未知母数 $\boldsymbol{\beta}$, ρ を含んでいるので，それらの推定量を代入した信頼区間が考えられる。一般に $\hat{\rho}$ を $\overline{y}_1, \ldots, \overline{y}_k$ に基づいた ρ の推定量とし，$\boldsymbol{\beta}$ を (5) で与えられる一般化最小 2 乗推定量 $\widetilde{\boldsymbol{\beta}}(\hat{\rho})$ で推定すると，μ_i の経験ベイズ推定量は

$$\widehat{\mu}_i^{EB}(\hat{\rho}) = \overline{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}}(\hat{\rho}) + (1-\hat{\gamma}_i)\left(\overline{y}_i - \overline{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}}(\hat{\rho})\right), \quad \hat{\gamma}_i = (1+n_i\hat{\rho})^{-1}$$

と書けるので，経験ベイズ的信頼区間は $I_i^B(\boldsymbol{\beta},\rho)$ から

$$I_i^{EB}(\hat{\rho}) : \ \widehat{\mu}_i^{EB}(\hat{\rho}) \pm z_{\alpha/2}\sqrt{(\sigma_e^2/n_i)(1-\hat{\gamma}_i)}$$

となる。経験ベイズ推定量 $\widehat{\mu}_i^{EB}(\hat{\rho})$ は $\overline{y}_i$ に比べて推定精度が高いだけでなく，信頼区間の長さも短くなっている。しかし，被覆確率 $P[\mu_i \in I_i^{EB}(\hat{\rho})]$ は $1-\alpha$ に一致しないという欠点をもつ。

そこで，3.1 節の議論を用いて，被覆確率が k に関して 2 次漸近的に $1-\alpha$ で近似できるような信頼区間を構成する。このような信頼区間は，$n_1 = \cdots = n_k$ で σ_e^2 が既知のときには Basu, Ghosh and Mukerjee (2003) などによって求められ，$n_1, \ldots, n_k$ が等しいことを仮定せず，σ_e^2 が未知というより一般的な設定の下では笹瀬-久保川 (2005) によって得られた。Basu *et al.* (2003) の論法に従って，$z_{\alpha/2}$ の代わりに $z_{\alpha/2}\{1+(2k)^{-1}h(\hat{\rho})\}$ を用いた信頼区間

$$I_i^{AEB} : \ \widehat{\mu}_i^{EB}(\hat{\rho}) \pm z_{\alpha/2}\left[1+(2k)^{-1}h(\hat{\rho})\right]\sqrt{(\sigma_e^2/n_i)(1-\hat{\gamma}_i)} \tag{24}$$

を考える。ここで，補正関数 $h(\hat{\rho})$ は，$V_{\hat{\rho}}(\rho) = E[(\hat{\rho}-\rho)^2]$ を用いて

$$h(\hat{\rho}) = (1+z_{\alpha/2}^2)\frac{kn_i^2\hat{\gamma}_i^4}{4(1-\hat{\gamma}_i)^2}V_{\hat{\rho}}(\hat{\rho}) + \frac{kn_i\hat{\gamma}_i^2}{1-\hat{\gamma}_i}\left\{\overline{\boldsymbol{x}}_i'\Big(\sum_{j=1}^{k}\frac{n_j\overline{\boldsymbol{x}}_j\overline{\boldsymbol{x}}_j'}{1+n_j\hat{\rho}}\Big)^{-1}\overline{\boldsymbol{x}}_i + 2n_i\hat{\gamma}_iV_{\hat{\rho}}(\hat{\rho})\right\} \tag{25}$$

で与えられる。このとき，$k \to \infty$ のとき

$$P[\mu_i \in I_i^{AEB}] = 1 - \alpha + O(k^{-3/2})$$

が成り立つことが示される。(18) で与えられた $\hat{\rho}$ に対しては，$V_{\hat{\rho}}(\rho) = 2\sum_{i=1}^{k}(1+n_i\rho)^2/N^2 + o(k^{-1})$ となるので，これを (25) に代入すれば補正項が得られる。特に $n_1 = \cdots = n_k = n$ のときには，

$$h(\hat{\rho}) = \frac{1+z_{\alpha/2}^2}{2n^2\hat{\rho}^2} + \frac{1}{n\hat{\rho}}\left\{\overline{\boldsymbol{x}}_i'\Big(\sum_{j=1}^{k}\overline{\boldsymbol{x}}_j\overline{\boldsymbol{x}}_j'\Big)^{-1}\overline{\boldsymbol{x}}_i + 4\right\}$$

と表される。

3.3. 応用例：地価公示価格の小地域推定

さて，具体的な小地域に関するデータを用いて EBLUP の特徴，平均２乗誤差推定及び EBLUP に基づいた信頼区間を調べてみよう。ここで用いるデータは，神奈川県における京浜急行線沿いの宅地物件について 2001 年に公表された $1m^2$ 当たりの地価公示価格である。各駅を１つの小地域と考え，また i 番目の駅を最寄り駅とする物件のデータをその小地域からとられたデータと考えて，その個数を n_i で表す。小地域の総数は $k = 32$ であり，n_i は 1 から 11 まで不均一な値をとっているが平均 4 程度である。各地価公示価格を対数変換したものを y_{ij} とし，(3) に対応するモデル

$$y_{ij} = \beta_0 + x_{1i}\beta_1 + x_{2ij}\beta_2 + v_i + e_{ij}$$

を想定してみる。ここで，共変量 x_{1i} は i 番目の駅から品川駅に到着するのに要する時間，x_{2ij} は物件 (i,j) から最寄り駅 (i) までの距離を表している。$\boldsymbol{x}_{ij} = (1, x_{1i}, x_{2ij})'$, $\overline{\boldsymbol{x}}_i = (1, x_{1i}, \overline{x}_{2i})'$ とおくと, 行列 $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_i)'$ のランクは 1 となるので，(7) の $\hat{\sigma}_e^2$ における λ は $\lambda = 2$ となることに注意する。このモデルにおいて，駅 (小地域) ごとに平均的な地価公示価格

$$\mu_i = \beta_0 + x_{1i}\beta_1 + \overline{x}_{2i}\beta_2 + v_i$$

の予測に関して，これまで説明してきた手法の挙動を調べてみよう。

データから推定値 $\widetilde{\boldsymbol{\beta}}(\hat{\rho}), \hat{\sigma}_e^2, \hat{\rho}$ を計算すると，$\widetilde{\boldsymbol{\beta}}(\hat{\rho}) = (13.2702, -0.0138229, -6.24894\times10^{-5})$, $\hat{\sigma}_e^2 = 0.055070$, $\hat{\rho} = 0.231126$ となる。β_1 の推定値が負の値であることから，東京駅から遠くなるにつれて地価公示価格は低くなる傾向にあり，合理的な符号を示していることがわかる。表 1 は，京浜急行沿線の物件 $1m^2$ 当たりの駅ごとの平均価格の予測値と予測誤差を与えている。No.1 京急川崎から No.32 津久井浜までの 32 の駅が東京に近い順に番号付けられている。n_i は利用可能なデータ数，$\overline{y}_i$ は標本平均値，EBLUP$_i$ は (14) から計算される予測値，$\overline{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}}$ はプールされた推定値に基づいた回帰推定値を示している。表 1 の左側にそれらの予測値が与えられている。その表をながめてみると，EBLUP$_i$ は $\overline{y}_i$ を $\overline{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}}$ の方向へ縮小しており，特に n_i が小さいときに縮小の度合いが大きくなり n_i が大きくなるにつれて縮小の程度は小さくなることがわかる。表 1 の右側には，(17) で与えられた平均２乗誤差（標準化されたもの）の推定値が与えられている。$1/n_i$ は，v_i を母数としたときの $\overline{y}_i$ の平均２乗誤差であり，$\widehat{M}_i^U$, $\widehat{M}_i^{CU}$ はそれぞれ EBLUP$_i$ の MSE と条件付 MSE の２次漸近不偏推定値を表しており，(19), (21) から計算

表 1. 京浜急行線沿線の物件の $1m^2$ 当たりの駅ごとの平均価格の予測値と予測誤差

			予測値			予測誤差		
No.	最寄り駅	n_i	$\overline{y}_i$	EBLUP$_i$	$\overline{\boldsymbol{x}}_i'\widetilde{\boldsymbol{\beta}}$	$1/n_i$	$\widehat{M}_i^U$	$\widehat{M}_i^{CU}$
1	京急川崎	5	554983	487388	419461	0.200	0.129	0.129
2	鶴見市場	7	319928	333513	356725	0.143	0.098	0.101
3	京急鶴見	2	570920	437352	386659	0.500	0.194	0.193
4	花月園前	2	283069	320221	339006	0.500	0.185	0.187
5	生麦	7	293188	304061	322512	0.143	0.096	0.099
6	京急子安	2	321134	351243	366099	0.500	0.191	0.192
7	子安	1	525000	347976	316423	1.000	0.232	0.226
8	神奈川新町	1	270000	322382	335868	1.000	0.237	0.232
9	神奈川	1	331000	339691	341732	1.000	0.239	0.233
10	戸部	3	375575	354298	340257	0.333	0.155	0.156
11	日ノ出町	4	322805	324860	326771	0.250	0.133	0.135
12	黄金町	3	305847	315063	321615	0.333	0.153	0.155
13	南太田	2	371510	334768	319035	0.500	0.185	0.184
14	屏風浦	2	244499	266831	277832	0.500	0.178	0.179
15	杉田	2	236169	260133	272017	0.500	0.177	0.178
16	京急富岡	8	228446	237685	255764	0.125	0.084	0.087
17	能見台	11	246572	245929	244303	0.091	0.067	0.069
18	金沢文庫	6	257464	259956	263451	0.167	0.103	0.105
19	追浜	7	189859	204841	231626	0.143	0.093	0.096
20	京急田浦	3	186865	206629	221548	0.333	0.153	0.156
21	安針塚	3	163998	192669	215441	0.333	0.163	0.167
22	逸見	4	178816	190617	202218	0.250	0.131	0.134
23	汐入	2	174379	200149	213316	0.500	0.180	0.181
24	横須賀中央	3	258351	228923	210511	0.333	0.158	0.158
25	京急安浦	6	208107	204624	199889	0.167	0.105	0.107
26	堀ノ内	2	212941	207532	205079	0.500	0.209	0.208
27	新大津	2	189447	194258	196523	0.500	0.195	0.196
28	北久里浜	6	201243	194083	184575	0.167	0.105	0.106
29	京急久里浜	6	240698	212929	179640	0.167	0.107	0.107
30	ＹＲＰ野比	7	197581	187313	171819	0.143	0.097	0.098
31	京急長沢	5	165187	162517	159486	0.200	0.129	0.131
32	津久井浜	6	148490	148449	148391	0.167	0.143	0.145

されたものである。これらの予測誤差の推定値をながめてみると，EBLUP$_i$ の予測誤差は $\overline{y}_i$ よりもかなり小さく，特に n_i が小さいときには著しい改善がなされていることがわかる。条件付 MSE の推定値 $\widehat{M}_i^{CU}$ と MSE の推定値 $\widehat{M}_i^{U}$ の間にはあまり差がみられない。このことは，正規分布に基づいた LMM を扱う限り両者はわずかな差でしかないことを意味している。Booth and Hobert (1998) が主張するように，GLMM を扱うときには両者に重大な差が生ずるのかもしれない。表 1 では与えられていないが，MSE の正確な不偏推定量 $\widehat{M}_i^{E}(\hat{\rho})$ の値を計算したところ，すべて負の値になってしまった。全体的にかなり縮小がなされているため MSE の正確な不偏推定値が 0 を越えて負の値を取ってしまったのかもしれないが，正確な不偏推定値が好ましくないことを意味しており，文献で指摘されてきた見解と符合する結果である。

図 1 は，2 次補正した経験ベイズ的信頼区間 I_i^{AEB} と $\overline{y}_i$ に基づいた信頼区間 I_i^* の両端の値を，駅を No.1 から No.32 まで横軸にとって描いた図である。I_i^{AEB} の値は (24) から計算できる。I_i^* の信頼区間の動きが n_i が小さいときに不安定になるのに比べ，I_i^{AEB} は安定した信頼区間を与えている。I_i^{AEB} の動きをながめてみると,「快特」「特急」電車が停車する駅では土地価格が高くなることを反映して駅ごとに微妙に変動しながら，全体として東京から遠くなるにつれて価格が減少するという合理的な傾向がみられる。図 2 は，信頼区間の長さとデータ数 n_i との関係を示したものである。I_i^* の長さは，n_i が大きいときには I_i^{AEB} と同程度であるものの, n_i が小さいときには I_i^{AEB} に比べてかなり大きくなってしまうことがわかる。

最後に，この節で計算した MSE の推定値と信頼区間は, 3 節において σ_e^2 を既知として得られた結果を利用したもので，具体的には $\hat{\sigma}_e^2 = 0.055070$ の値を既知として代入している。正確には，σ_e^2 を未知とするモデルのもとで導かれた MSE 推定量と信頼区間を用いるべきであるが，これらはもっと複雑な形をして修正項の数がかなり多くなってしまう。どちらがよいのかについて実用面からまた数値的側面から検討するのがよいと思われる。σ_e^2 が未知の場合の MSE 推定量及び信頼区間についてはそれぞれ Prasad and Rao (1990), 笹瀬-久保川 (2005) を参照されたい。

4. 様々な線形混合モデルと一般化線形混合モデル

2.4 節で説明したように，線形混合モデル (LMM) が小地域推定を行う上で優れた予測量を導くことのできる要因は，共通母数と変量効果を組み込んでいる点である。したがって，共通母数と変量効果を巧みに組み入れることによって，様々な小地域推定のための有効なモデルを構築することができる。この節では，LMM の代表的な一般化や変形について紹介するとともに，一般化線形混合モデル (GLMM) への拡張や死亡率推定のためのモデルなどを紹介する。

4.1. 様々な線形混合モデル

これまで取り上げてきたモデル (3), (16) は，より一般的な LMM

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{e} \tag{26}$$

の例となっている。ここで，$\boldsymbol{y}$ は観測データのベクトル，$\boldsymbol{X}$, $\boldsymbol{Z}$ は既知の行列である。また $\boldsymbol{v}$, $\boldsymbol{e}$ は独立にそれぞれ多次元正規分布 $\mathcal{N}(0, \boldsymbol{G})$, $\mathcal{N}(0, \boldsymbol{R})$ にしたがう確率変数で，共分散行列 $\boldsymbol{G}$, $\boldsymbol{R}$ はいくつかの分散成分に依存している。$\boldsymbol{y}$ の共分散行列は

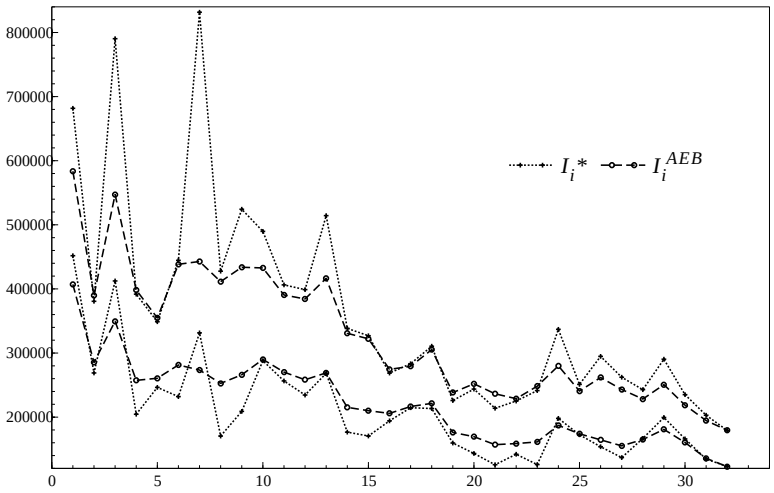

図 1. 補正後の信頼区間 I_i^{AEB} と 正規分布に基づく信頼区間 I_i^* の両端の値の比較 (No.1 から No.32 までの各駅を横軸に並べている。)

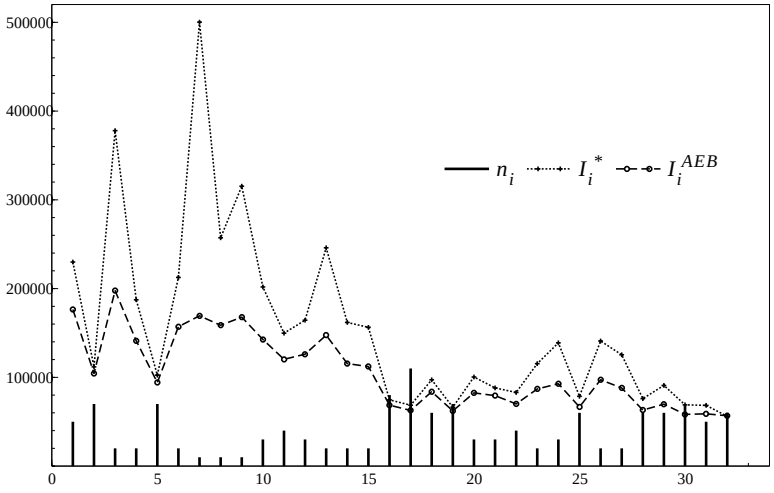

図 2. 補正後の信頼区間 I_i^{AEB} と 正規分布に基づく信頼区間 I_i^* の長さの比較とデータ数 n_i との関係 (n_i のスケールは縦軸 1000 が 1 個のデータを表している。No.1 から No.32 までの各駅を横軸に並べている。)

$$\boldsymbol{V} = \boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$$

と表されるので，$\boldsymbol{\beta}$ の一般化最小 2 乗推定量は

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$$

となる。このとき, たとえばベクトル $\boldsymbol{a}$, $\boldsymbol{b}$ を既知として $\mu = \boldsymbol{a}'\boldsymbol{\beta} + \boldsymbol{b}'\boldsymbol{v}$ を推定したいときには, その BLUP は

$$\widehat{\mu} = \boldsymbol{a}'\widehat{\boldsymbol{\beta}} + \boldsymbol{b}'\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

で与えられる。分散成分を最尤法やモーメント法を用いて推定して代入すると EBLUP が得られる。

データベクトル $\boldsymbol{y} = (y_i)$ の成分 y_i は，(3) のような個体レベルモデルに対しては各個体データ y_{ij} に対応している一方，(16) のような地域レベルモデルに対しては $\overline{y}_i$ に対応している。その意味で，モデル (26) は様々な LMM の変形を含んでいる。以下にいくつかの典型的なモデルを紹介しよう。なお，各モデルでの具体的な推定方法など詳しい内容については，Rao (2003), Searle *et al*. (1992), McCulloch and Searle (2000) が参照される。

[1] 多変量線形混合モデル. Fay-Herriot モデル (16) では一変量データに対して線形モデルが考えられたが，経時測定データや時系列データなどは一般に多変量データとして扱うことができる。i 番目の地域に対して p-変量データ $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{ip})'$ が与えられており，

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{v}_i + \boldsymbol{e}_i, \quad i = 1, \ldots, k,$$

なるモデルが考えられるとき，これを多変量 Fay-Herriot モデルという。ここで，$\boldsymbol{X}_i$ は $p \times m$ の共変量，$\boldsymbol{\beta}$ は $m \times 1$ の回帰係数ベクトル，$\boldsymbol{v}_i$ と $\boldsymbol{e}_i$ は互いに独立な確率変数で，$\boldsymbol{v}_i \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}_v)$, $\boldsymbol{e}_i \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Psi}_i)$ に従う。

また (3) を多変量への拡張したモデルは

$$\boldsymbol{y}_{ij} = \boldsymbol{B}\boldsymbol{x}_{ij} + \boldsymbol{v}_i + \boldsymbol{e}_{ij}, \quad j = 1, \ldots, n_i;\ i = 1, \ldots, k,$$

で表される。ここで，$\boldsymbol{y}_{ij}$, $\boldsymbol{v}_i$, $\boldsymbol{e}_{ij}$ は $p \times 1$ ベクトル，$\boldsymbol{B}$ は $p \times m$ の回帰係数行列であり，$\boldsymbol{v}_i \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}_v)$, $\boldsymbol{e}_{ij} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}_e)$ に従う。このモデルに対する EBLUP の性質は Kubokawa and Srivastava (2003) などで議論されている。また $\boldsymbol{\Sigma}_e$, $\boldsymbol{\Sigma}_v$ は共分散成分と呼ばれ，それらの優れた推定量の導出と性質が Srivastava and Kubokawa (1999), Kubokawa and Tsai (2006) で示されている。

[2] 時系列-クロス・セクション モデル. 計量経済学などの研究分野では時系列データに対する分散成分モデルがパネルデータ解析として研究されてきた。i 番目の地域の集計データ $\overline{y}_i$ が時系列的に得られているときの代表的なモデルは

$$\overline{y}_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_i + u_{it} + e_{it}, \quad t = 1, \ldots, T;\ i = 1, \ldots, k,$$

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}, \quad |\rho| < 1,$$

と表現される。ただし，$v_i \sim \mathcal{N}(0, \sigma_v^2)$, $e_{it} \sim \mathcal{N}(0, \sigma_e^2)$, $\varepsilon_{it} \sim \mathcal{N}(0, \sigma^2)$ であり，また確率変数列 $\{v_i\}, \{e_{it}\}, \{\varepsilon_{it}\}$ は互いに独立である。実は，3.3 節で扱った地下公示価格のデータもクロス・セクションのデータが時系列的に取られており，そのモデルを用いて解析することもできる。経時的に得られたデータのモデリングについては，Diggle, Liang and Zeger (1994), Verbeke and Molenberghs (2000) などが参照される。

[3] 空間モデル. 基本的な Fay-Herriot モデル (16) については，地域効果 $v_1, \ldots, v_k$ を独立同一分布に従う変量と仮定したが，空間モデルを作る際には v_i は隣接する地域効果 v_j の影響を考慮するのが現実的である。A_i を i 番目の地域に隣接する地域のインデックスの集合とすると，$\{v_\ell | \ell \in A_i\}$ を所与としたときの v_i の条件付分布が

$$v_i \,|\, \{v_\ell | \ell \in A_i\} \sim \mathcal{N}\left(\rho \sum_{\ell \in A_i} q_{i\ell} v_\ell, \sigma_v^2\right)$$

で与えられるモデルや重み w_{ij} を用いて

$$v_i \,|\, \{v_\ell | \ell \in A_i\} \sim \mathcal{N}\left(\frac{\sum_{\ell \in A_i} w_{i\ell} v_\ell}{\sum_{\ell \in A_i} w_{i\ell}}, \frac{\sigma_v^2}{\sum_{\ell \in A_i} w_{i\ell}}\right)$$

なるモデルを考えるのが自然である。これを条件付自己回帰 (CAR, Conditional Autoregression) 空間モデルという。最初のモデルについては，$q_{i\ell}$ は $q_{\ell i} = q_{i\ell}$ をみたす既知の値とし，$\boldsymbol{Q} = (q_{i\ell})$ を $k \times k$ の行列で，$q_{ii} = 0$, $\ell \notin A_i$ に対して $q_{i\ell} = 0$ をみたすものとする。このとき，最初の CAR モデルは

$$\boldsymbol{v} \sim \mathcal{N}_k(\boldsymbol{0}, \sigma_v^2 \boldsymbol{\Gamma}(\rho)), \quad \boldsymbol{\Gamma}(\rho) = (\boldsymbol{I} - \rho \boldsymbol{Q})^{-1}$$

で表される。$\boldsymbol{v}$ の共分散行列 $\boldsymbol{\Gamma}(\rho)$ としては，その他 $\boldsymbol{\Gamma}(\rho_1, \rho_2) = \rho_1 \boldsymbol{I} + \rho_2 \boldsymbol{D}$ などの取り方が知られている。ここで $\boldsymbol{D} = (e^{-d_{i\ell}})$ であり，$d_{i\ell}$ は地域 i と ℓ の距離のようなものをとることになる。空間モデルについては，Banerjee, Carlin and Gelfand (2004), Schabenberger and Gotway (2005) などが参照される。

[4] 変量係数モデル. モデル (3), (16) は，定数項が地域によって変動する形をしているが，一般に変動を回帰係数に組み込んだモデルも考えられる。y_{ij} は個体レベルのモデルとして

$$y_{ij} = \boldsymbol{x}_{ij}' \boldsymbol{\beta}_i + e_{ij}, \quad j = 1, \ldots, n_i;\ i = 1, \ldots, k,$$

の形で与えられ，さらに $\boldsymbol{\beta}_i = (\beta_1, \ldots, \beta_m)'$ が地域レベルの共変量 $\boldsymbol{W}_i$ を用いて

$$\boldsymbol{\beta}_i = \boldsymbol{W}_i \boldsymbol{\alpha} + \boldsymbol{v}_i, \quad i = 1, \ldots, m$$

のようなモデルに従っているとする。ここで，$\boldsymbol{v}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_v)$, $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ に従う。このとき，モデルは $y_{ij} = \boldsymbol{x}_{ij}' \boldsymbol{W}_i \boldsymbol{\alpha} + \boldsymbol{x}_{ij}' \boldsymbol{v}_i + e_{ij}$ と表されて，これを変量係数モデルという。たとえば，$\boldsymbol{x}_{ij}$ の第 1 成分 $\boldsymbol{x}_{ij1}$ がある処置を施したとき 1 そうでないとき 0 をとるダミー変数であるならば，処置を施したモデルには変量 $\boldsymbol{v}_i$ の第 1 成分 v_{i1} が含まれるのに対して，もう一方のモデルには v_{i1} が含まれない。従って，処置を施したモデルは，そうでない場合に比べて分散が大きくなる。こうした状況は現実にしばしば起こりうることであり，変量係数モデルも考慮に値す

るモデルになるであろう。

[5] 2元配置混合モデル. (3) をさらに複雑にしたモデルに2元配置の線形混合モデルがある。i 番目の地域が M_i 個のクラスターに分けられ，j 番目のクラスターが N_{ij} 個の個体から成ると仮定する。各地域から，m_i 個のクラスターが抽出され，j 番目のクラスターについて さらに n_{ij} 個の個体が抽出されるとすると，2元配置線形混合モデルは

$$y_{ij\ell} = \boldsymbol{x}'_{ij\ell}\boldsymbol{\beta} + v_i + u_{ij} + e_{ij\ell}, \quad \ell = 1, \ldots, n_{ij};\ j = 1, \ldots, m_i;\ i = 1, \ldots, k$$

と表される。ここで，v_i, u_{ij} はそれぞれ地域効果，クラスター効果を表し，$e_{ij\ell}$ は誤差項で，すべて互いに独立に分布し，$v_i \sim \mathcal{N}(0, \sigma_v^2)$, $u_{ij} \sim \mathcal{N}(0, \sigma_u^2)$, $e_{ij\ell} \sim \mathcal{N}(0, \sigma_e^2)$ に従うと仮定する。$\overline{\boldsymbol{x}}_i = \sum_{j=1}^{m_i}\sum_{\ell=1}^{n_{ij}} \boldsymbol{x}_{ij\ell} / \sum_{j=1}^{m_i} n_{ij}$ とおくとき，$\mu_i = \overline{\boldsymbol{x}}'_i\boldsymbol{\beta} + v_i$ が推定したい量になる。このモデルは，有限母集団の枠組みで考えた方が現実的であり，EBLUP とその MSE の性質などが論じられている。

[6] 分散変動モデル. 以上で紹介してきたモデルは誤差分散が等しいことを仮定している。この仮定を緩めるモデルを扱う必要がある場合は，分散変動モデルを考えることができる。たとえば，(3) においては誤差項 e_{ij} は $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ に従うと仮定しているが，これを，σ_i^2 が所与のときの e_{ij} の条件付分布

$$e_{ij}|\sigma_i^2 \sim \mathcal{N}(0, \sigma_i^2)$$

で置き換え，$\sigma_1^2, \ldots, \sigma_k^2$ を互いに独立な確率変数として σ_i^{-2} に適当な分布を想定する。たとえば，ガンマ分布や逆ガウス分布などが考えられる。こうして，分散の異質性を考慮したモデルが得られる。

[7] 経験ベイズモデルと階層ベイズモデル. モデル (3) を

$$y_{ij} = \theta_{ij} + e_{ij}, \quad \theta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + v_i,$$

と分解すると，条件付分布は $y_{ij}|(\theta_{ij}, \sigma_e^2) \sim \mathcal{N}(\theta_{ij}, \sigma_e^2)$ となり，θ_{ij} の事前分布が $\theta_{ij} \sim \mathcal{N}(\boldsymbol{x}'_{ij}\boldsymbol{\beta}, \sigma_v^2)$ で与えられるベイズモデルの形で捉えることができる。母数 $\boldsymbol{\beta}$, σ_v^2, σ_e^2 が既知のときを主観的ベイズという。これに対して客観性を持たせるために $\boldsymbol{\beta}$, σ_v^2, σ_e^2 を未知母数として扱ったり，変量として扱い適当な分布を仮定したりする。未知母数として扱ったものは経験ベイズモデルと呼ばれるが，線形混合モデル (3) は経験ベイズモデルに対応している。

一方，変量として扱ったモデルを階層ベイズモデルという。この場合，(i,j)-成分に θ_{ij} をもつ行列を $\boldsymbol{\theta}$ とおくと，$(\boldsymbol{\theta}, \sigma_e^2)|(\boldsymbol{\beta}, \sigma_v^2) \sim \pi_1(\boldsymbol{\theta}, \sigma_e^2|\boldsymbol{\beta}, \sigma_v^2)$ を第1段階事前分布，$(\boldsymbol{\beta}, \sigma_v^2) \sim \pi_2(\boldsymbol{\beta}, \sigma_v^2)$ を第2段階事前分布という。一般に，第1段階事前分布は正確に，第2段階事前分布は無情報的に設定すると，客観的なベイズ推定を構成できることが知られている。例えば，第1段階事前分布 $\pi_1(\boldsymbol{\theta}, \sigma_e^2|\boldsymbol{\beta}, \sigma_v^2)$ として

$$\begin{aligned}\theta_{ij}|(\boldsymbol{\beta}, \sigma_v^2) &\sim \mathcal{N}(\boldsymbol{x}'_i\boldsymbol{\beta}, \sigma_v^2),\\ \sigma_e^2 &\sim \sigma_e^{-2}\mathrm{d}\sigma_e^2\end{aligned}$$

がとられる。$\sigma_e^{-2}\mathrm{d}\sigma_e^2$ は尺度変換に関して不変な測度で無情報事前分布を表している。また第2

段階事前分布 $\pi_2(\boldsymbol{\beta}, \sigma_v^2)$ としては σ_v^2 には無情報事前分布 $\sigma_v^{-2}\mathrm{d}\sigma_v^2$ を想定し，$\boldsymbol{\beta}$ に対しては，(1) 一様分布 $\mathrm{d}\boldsymbol{\beta}$, (2) $\boldsymbol{\beta}|\sigma_v^2 \sim \mathcal{N}(\boldsymbol{\beta}_0, \sigma_v^2\boldsymbol{A})$, (3) $\boldsymbol{\beta}|(\sigma_v^2, \lambda) \sim \mathcal{N}(\boldsymbol{\beta}_0, \lambda\sigma_v^2\boldsymbol{A})$, $\lambda \sim \pi_3(\lambda)$, などの場合が考えられる。ここで $\boldsymbol{\beta}_0$, $\boldsymbol{A}$ は既知の値とする。このような階層ベイズ推定量の理論的な性質については Kubokawa and Strawderman (2007) とその中の参考文献が参照される。また階層ベイズを用いた空間データの解析については，Banerjee *et al*. (2004) が参照される。

4.2. 一般化線形混合モデル

これまでは，小地域から得られたデータが連続変量の場合を取り上げ正規分布に基づいた LMM の性質，様々なモデルの紹介や推定法の説明を行ってきた。しかしデータが死亡数など離散的に変動するときには，２項分布やポアソン分布などに基づいたモデルを考える必要がある。こうした離散分布に回帰項と変量効果を組み入れて小地域推定に適したモデルを構築することができ，それを統一的に扱ったモデルが一般化線形混合モデル (GLMM) である。

全体で k 個の地域があり，i 番目の地域から n_i 個のデータ（もしくは集計データ）$y_{i1}, \ldots, y_{in_i}$ が取られており，v_i を所与としたときの y_{ij} の条件付分布が

$$f(y_{ij}|v_i) = \exp\left\{[y_{ij}\theta_{ij} - b(\theta_{ij})]/\tau_{ij} + c(y_{ij}, \tau_{ij})\right\},$$

$j = 1, \ldots, n_i;\ i = 1, \ldots, k,$ で与えられるとする。このようなモデルは一般化線形モデルと呼ばれる (McCullagh and Nelder (1989))。この密度関数は自然母数 θ_{ij} と尺度母数 $\tau_{ij}(> 0)$ を用いて表現されており，τ_{ij} は既知と仮定されている。y_{ij} の条件付期待値を $E[y_{ij}|v_i] = \mu_{ij}$ と書くとき，μ_{ij} がリンク関数 $g(\cdot)$ を通して共変量 $\boldsymbol{x}_{ij}$ と関係づけることができることを仮定して

$$g(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + v_i$$

なる形で表現できるとする。ここで，$v_i \sim \mathcal{N}(0, \sigma_v^2)$, $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ であり，すべて互いに独立であると仮定する。このようなモデルが GLMM で，4.1 節で列挙された様々な構造を組み込むことによって，現実のデータに適合するモデルを構築することができる。

GLMM の代表的な例として２項分布とポアソン分布の場合を取り上げよう。

(1) ２項分布．$y_{ij}|v_i$ が２項分布 $Bin(n_{ij}, p_{ij})$ に従うときには，$\theta_{ij} = \log\{p_{ij}/(1-p_{ij})\}$, $\tau_{ij} = 1$, $b(\theta_{ij}) = n_{ij}\log(1 + e^{\theta_{ij}})$, $c(y_{ij}, \tau_{ij}) = \log n_{ij}!/\{y_{ij}!(n_{ij} - y_{ij})!\}$ に対応している。リンク関数の代表的な取り方にはロジットとプロビットがある。

ロジット：$\mathrm{logit}(p_{ij}) = \log\{p_{ij}/(1-p_{ij})\} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + v_i$

プロビット：$\mathrm{probit}(p_{ij}) = \Phi^{-1}(p_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + v_i$

(2) ポアソン分布．$y_{ij}|v_i$ がポアソン分布 $Po(\lambda_{ij})$ に従うときには，$\theta_{ij} = \log\lambda_{ij}$, $\tau_{ij} = 1$, $b(\theta_{ij}) = \exp\{\theta_{ij}\}$, $c(y_{ij}, \tau_{ij}) = -\log y_{ij}!$ に対応している。リンク関数として対数関数をとり，

$$\log(\lambda_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + v_i$$

という形でモデル化する。

2.4 節で説明されたように，共通母数と変量効果を組み入れた GLMM は，地域の特性値に対して精度の高い安定した推定値を与えることができる。その代表的な例が疾病地図の作成に用い

られる死亡率及び死亡指標の推定である。ここでは，次のような多項データの小地域推定について説明しよう。地域が $i = 1, \ldots, k$ に，また年齢階級が $j = 1, \ldots, J$ に分割されているとき，i 地域 j 年齢階級の (特定の病気による) 死亡指標を推定することを考える。

y_{ij} : i 地域，j 年齢階級の観測死亡数
n_{ij} : i 地域，j 年齢階級の人口
N_{0j} : 標準人口集団の j 年齢階級人口，　$N_0 = N_{01} + N_{02} + \cdots + N_{0J}$
Y_{0j} : 標準人口集団の j 年齢階級の死亡数，　$Y_0 = Y_{01} + Y_{02} + \cdots + Y_{0J}$

死亡率の代表的な指標として用いられるのが標準化死亡率 (Standardized Mortality Rate, SMR) で, (観測死亡数) ／ (期待死亡数) で定義される。i 地域, j 年齢階級に対して標準人口集団から期待される死亡数は，$E_{ij} = n_{ij}Y_{0j}/N_{0j}$ で与えられるので，このカテゴリーの SMR は

$$SMR_{ij} = y_{ij}/E_{ij}$$

となる。また i 地域の SMR は

$$SMR_i = \frac{\sum_{j=1}^{J} y_{ij}}{\sum_{j=1}^{J} E_{ij}} = \frac{\sum_{j=1}^{J} E_{ij} SMR_{ij}}{\sum_{j=1}^{J} E_{ij}}$$

となる。標準人口集団の人口と死亡数が利用できないときには，E_{ij} の代わりに $n_{ij} \sum_{i=1}^{k} y_{ij} / \sum_{i=1}^{k} n_{ij}$ を用いる。

n_{ij} が小さいときには SMR_{ij} はバラツキが大きくなってしまう。また SMR_i についても J が小さいか $\sum_{j=1}^{J} n_{ij}$ が小さいときには同様な問題が生ずる。この問題に対して Ghosh, Natarajan, Stroud and Carlin (1998) は階層ベイズ的な GLMM を考え， Gibbs sampler を用いて Missouri 州の肺ガンの死亡率地図を作成した。また丹後 (1988) はポアソン・ガンマモデルを当てはめて疾病地図の作成を行った。以下でこれらのモデルを説明しよう。

[1] 一般化線形混合モデルの適用. 第 i 地域に対して $t_i = \sum_{j=1}^{J} y_{ij}$ とおくと，$(y_{i1}, \ldots, y_{iJ})$ は多項分布 $Mult(t_i; p_{i1} \ldots, p_{iJ})$ に従う。ここで $\sum_{j=1}^{J} p_{ij} = 1$ である。いま $y_{i1}, \ldots, y_{iJ}$ が互いに独立に分布し，y_{ij} がポアソン分布 $Po(\lambda_{ij})$ に従っているとすると，離散分布のよく知られた性質から $\sum_{j=1}^{J} y_{ij} = t_i$ を与えたときの $(y_{i1}, \ldots, y_{iJ})$ の条件付分布は多項分布 $Mult(t_i; p_{i1} \ldots, p_{iJ})$ に従う。ここで，$p_{ij} = \lambda_{ij} / \sum_{j=1}^{J} \lambda_{ij}$ である。$\log \lambda_{ij}/E_{ij}$ に線形混合関係を仮定したモデル

$$\log \frac{\lambda_{ij}}{E_{ij}} = x_j \beta + v_i$$

を考えると，λ_{ij}/E_{ij} の予測量が SMR_{ij} に対応する死亡指標になる。ここで，x_j は年齢階級に対応して設定される値で，たとえば $J = 3$ で，54 才までは $x_1 = -1$, 55 才〜64 才に対しては $x_2 = 0$, 65 才以上は $x_3 = 1$ などと設定される。また交互作用項 γ_{ij} を $\gamma_{ij} \sim \mathcal{N}(0, \sigma_\gamma^2)$ として組み入れたモデルも考えられる。

一般化線形混合モデルは，共変量が利用できる場合には有用であるが，ロジットモデルやプロビットモデルを用いた場合には推定量を明示的に書き下すことができないため，どのような操作によって平滑化がなされるのか，そのイメージを描きにくい。この点，次のポアソン・ガンマモ

デルについては推定量がわかりやすい形で表現できる。

[2] ポアソン・ガンマモデル. i 地域，j 年齢階級の死亡率を p_{ij} とし，観測死亡数 y_{ij} が平均 $n_{ij}p_{ij}$ のポアソン分布 $y_{ij} \sim Po(n_{ij}p_{ij})$ にしたがうとする。ポアソン分布の共役事前分布はガンマ分布であることから，p_{ij} にガンマ分布 $Ga(a,b)$ を仮定するようなベイズモデルが考えられる。丹後 (1988) は p_{ij} にいくつかのモデルを想定して死亡率推定に有用な経験ベイズ推定手法を導いた。まず，地域効果 α_i と年齢効果 β_j に対して M_1： $p_{ij} = \alpha_i\beta_j$ と仮定し，$\beta_j = Y_{0j}/N_{0j}$ とおくと α_i の最尤推定量は SMR_i が出てくるという興味深い結果を導いている。すなわち，SMR_i はベイズモデルを仮定せず年齢階級死亡率が地域によらない定数という仮定のもとで導出された指標であることを意味する。さらに丹後 (1988) では，α_i がガンマ分布に従い β_j を未知母数とするモデルを考えて経験ベイズ推定値の導出と有用性を議論している。これは，

$$\log p_{ij} = \log \beta_j + \log \alpha_i, \quad \alpha_i \sim Ga(a,b)$$

なる線形混合モデルの形で表現できる。また M_2： $p_{ij} = \beta_j\gamma_{ij}$ なる構造を入れ，γ_{ij} にガンマ分布 $Ga(a_j,b_j)$ を仮定するモデルも考察し，死亡率推定の有用性を示している。p_{ij} の経験ベイズ推定量は，a_j, b_j, β_j を周辺分布の最尤推定量 $\hat{a}_j$, $\hat{b}_j$, $\widehat{\beta}_j$ を用いて

$$\hat{p}_{ij}^{EB} = \widehat{\beta}_j \frac{\hat{a}_j + y_{ij}}{\hat{b}_j + n_{ij}\widehat{\beta}_j} \tag{27}$$

なる形で表される。モデル M_2 を組み入れたポアソン-ガンマモデルは，

$$\log p_{ij} = \log \beta_j + \log \gamma_{ij}, \quad \gamma_{ij} \sim Ga(a_j,b_j)$$

として GLMM の形で表されるが，経験ベイズ推定量が (27) の形で表されており，推定量が平滑化される方法が理解しやすい。すなわち，$\hat{a}_j = \hat{b}_j = 0$ なら $\hat{p}_{ij}^{EB} = \widehat{\beta}_j y_{ij}/n_{ij}$ であり，$\hat{a}_j > 0$, $\hat{b}_j > 0$ が組み込まれることによって，$\hat{p}_{ij}^{EB}$ はより安定した推定値を与えることができる。

GLMM の詳しい解説については，McCullagh and Nelder (1989, 14.5 節)，Fahrmeir and Tutz (2001), McCulloch (2003) が参照される。また，McCulloch and Searle (2000) はわかりやすく書かれた本であり，疾病地図など空間モデルを扱った本としては，Lawson, Browne and Vidal Rodeiro (2003), Lawson (2006) などがあるので参照してもらいたい。

謝　辞　２人の査読者の方には貴重なコメントを頂きまして深く感謝申し上げます。本研究は，科学研究費補助金 15200021, 15200022, 16500172 及び東京大学大学院経済学研究科２１世紀ＣＯＥプログラムから研究助成を受けております。

参　考　文　献

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall, New York.

Basu, R., Ghosh, J.K., and Mukerjee, R. (2003). Empirical Bayes prediction intervals in a normal regression model: higher order asymptotics. *Statist. Prob. Letters*, **63**, 197-203.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.*, **83**, 28-36.

Booth, J.G. and Hobert, J.P. (1998). Standard errors of prediction in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **93**, 262-272.

Datta, G.S., Kubokawa, T. and Rao, J.N.K. (2002). Estimation of MSE in small area estimation. Unpublished manuscript.

Datta, G.S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statist. Sinica*, **10**, 613-627.

Diggle, P., Liang, K.-Y., and Zeger, S.L. (1994). *Longitudinal Data Analysis*. Oxford Univ. Press.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.*, **70**, 311-319.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. 2nd ed. Springer, New York.

Fay, R.E. and Herriot, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.*, **74**, 269-277.

Ghosh, M. and Meeden, G. (1997). *Baysian Methods for Finite Population Sampling*. Chapman and Hall, New York.

Ghosh, M., Natarajan, K., Stroud, T.W. and Carlin, B.P. (1998). Generalized linear models for small-area estimation. *J. Amer. Statist. Assoc.*, **93**, 273-282.

Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statist. Science*, **9**, 55-93.

Kubokawa, T. (2000). Estimation of variance and covariance components in elliptically contoured distributions. *J. Japan Statist. Soc.*, **30**, 143-176.

Kubokawa, T., Saleh, A.K.Md.E., and Konno, Y. (2000). Bayes, minimax and nonnegative estimators of variance components under Kullback-Leibler loss. *J. Statist. Plan. Inf.*, **86**, 201-214.

Kubokawa, T. and Srivastava, M.S. (2003). Prediction in multivariate mixed linear models. *J. Japan Statist. Soc.*, **33**, 245-270.

Kubokawa, T. and Strawderman, W.E. (2007). On minimaxity and admissibility of hierarchical Bayes estimators. *J. Multivariate Analysis*, to appear.

Kubokawa, T. and Tsai, M.-T. (2006). Estimation of covariance matrices in fixed and mixed effects linear models. *J. Multivariate Analysis*, **97**, 2242-2261.

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statist. Science*, **18**, 199-210.

Lawson, A.B. (2006). *Statistical Methods in Spacial Epidemiology*. 2nd ed. Wiley, England.

Lawson, A.B., Browne, W.J. and Vidal Rodeiro, C.L. (2003). *Disease Mapping with WinBUGS and MLwiN*. Wiley, England.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. 2nd ed. Chapman and Hall, London.

McCulloch, C.E. (2003). *Generalized Linear Mixed Models*. NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 7. IMS, USA.

McCulloch, C.E. and Searle, S.R. (2000). *Generalized, Linear and Mixed Models*. Wiley, New York.

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *J. Amer. Statist. Assoc.*, **85**, 163-171.

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

Robinson, G.K. (1991). That BLUP is a good thing: The estimation of random effects. *Statist. Science*, **6**, 15-51.

Searle, S.R., Casella, G., and McCulloch, C.E. (1992). *Variance Components*, Wiley, New York.

Schabenberger, O. and Gotway, C.A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman and Hall, New York.

Srivastava, M.S. and Kubokawa, T. (1999). Improved nonnegative estimation of multivariate components of variance. *Ann. Statist.*, **27**, 2008-2032.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

笹瀬吉隆，久保川達也 (2005). 経験ベイズ信頼区間の漸近補正と小地域推定への応用. 日本統計学会誌（和文誌), **35**, 27-54.

丹後俊郎 (1988). 死亡指標の経験的ベイズ推定量について. 応用統計学, **17**, 81-96.

（2006 年 9 月 27 日受付　11 月 29 日最終修正　12 月 3 日採択）

応用統計学 Vol. 35, No. 3 (2006)

著者連絡先：久保川達也
113-0033 東京都文京区本郷 7-3-1 東京大学大学院経済学研究科
FAX: 03-5841-5521
tatsuya@e.u-tokyo.ac.jp

Japanese J. Appl. Statist. 35 (3) (2006), 1–24

Linear Mixed Models and Small Area Estimation

Tatsuya Kubokawa

Graduate School of Economics, University of Tokyo

Abstract

Sample survey data can be used to derive a reliable estimate of a total mean for a large area. When the same data are used to estimate means of small areas like city, county or town belonging to the large area, the usual direct estimators like the sample mean have unacceptably large standard errors due to the small sizes of the samples in the small areas. This is called a small area problem. To find more accurate estimates for given small areas, one needs to "borrow strength" from the related areas. The linear mixed model (LMM) is recognized as an appropriate model for handling such a problem, and the resulting empirical best linear unbiased predictor (EBLUP) can yield a smaller standard error.

This article gives a review of the small area estimation based on LMM. Especially, the article explains how the structure of (common parameters)+(random effects) in LMM works to get accurate estimates. The estimators of the mean squared errors of EBLUP and the confidence interval based on EBLUP are derived to evaluate accuracy of EBLUP. Finally, some generalizations and various variants of LMM are described for analyzing spatial data, and the generalized linear mixed model (GLMM) and its application to estimation of mortality rates are explained.

Key words: confidence interval, empirical Bayes method, finite population, generalized linear mixed model, linear mixed model, mean squared error, random effects, small area estimation, variance components model.

E-mail address: tatsuya@e.u-tokyo.ac.jp (Tatsuya Kubokawa)

Received September 27, 2006; Received in final form November 29, 2006; Accepted December 3, 2006.