線形混合モデルと小地域推定
－特に変数選択規準の導入を巡って－

久保川 達也

（東京大学 大学院経済学研究科, E-mail: tatsuya@e.u-tokyo.ac.jp）

1　線形混合モデルを利用した小地域推定

$\boldsymbol{X}_i$ を $n_i \times p$ 行列，$\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})'$，$\boldsymbol{Z}_i$ を $n_i \times m_i$ フルランク行列，$\boldsymbol{G}_i$ を $m_i \times m_i$ 正定値行列とする一般的な線形混合モデル

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

について考察した。ここで，変量効果 $\boldsymbol{v}_i$ と誤差項 $\boldsymbol{\epsilon}_i$ は互いに独立に分布し，

$$\boldsymbol{v}_i \sim \mathcal{N}_{m_i}(\boldsymbol{0}, \sigma^2\boldsymbol{G}_i), \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i})$$

とする。$N = \sum_{i=1}^k n_i$, $M = \sum_{i=1}^k m_i$, $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_k')'$, $\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_k')'$ とし，$\boldsymbol{v}$, $\boldsymbol{\epsilon}$ も同様に定義し，$\boldsymbol{Z} = \text{block diagonal}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_k)$ とおくと，$\boldsymbol{R} = \boldsymbol{I}_N$ に対して

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{\epsilon}, \quad \boldsymbol{v} \sim \mathcal{N}_M(\boldsymbol{0}, \sigma^2\boldsymbol{G}),\ \boldsymbol{\epsilon} \sim \mathcal{N}_N(\boldsymbol{0}, \sigma^2\boldsymbol{R}),$$

と表される。ただし，$\boldsymbol{G} = \text{block diag}(\boldsymbol{G}_1, \ldots, \boldsymbol{G}_k)$ である。枝分かれ誤差回帰モデル (NERM), Fay-Herriot モデル, 経時測定 線形混合モデルはこのモデルに含まれる。

Henderson (1950) の混合モデル方程式を解くことにより，$\boldsymbol{\beta}$, $\boldsymbol{v}$ の推定量

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}, \quad \widehat{\boldsymbol{v}} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

が得られる。共分散行列 $\boldsymbol{G}$, $\boldsymbol{R}$ が既知の場合に，既知のベクトル $\boldsymbol{a} \in \boldsymbol{R}^p$, $\boldsymbol{b} \in \boldsymbol{R}^q$ に対して $\mu = \boldsymbol{a}'\boldsymbol{\beta} + \boldsymbol{b}'\boldsymbol{v}$ を推定したいときには，μ の最良線形不偏予測量 (BLUP) は，

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

で与えられる。$\boldsymbol{G}$, $\boldsymbol{R}$ が未知の場合には，ML, REML などにより推定量 $\widehat{\boldsymbol{G}}$, $\widehat{R}$ が求められ，これらを代入することにより，経験最良線形不偏予測量 (EBLUP) が得られる。

集計データに基づいた地域レベルモデルとして用いられる Fay-Herriot モデル

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

については，$\mu_i = \overline{\boldsymbol{x}}_i'\boldsymbol{\beta} + v_i$ の EBLUP は，

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

と表される。ここで，$\psi = \sigma_v^2/\sigma^2$, $\boldsymbol{V}(\psi) = \psi \boldsymbol{I}_k + \boldsymbol{D}$ に対して，$\widehat{\boldsymbol{\beta}}(\psi) = (\boldsymbol{X}'\boldsymbol{V}(\widehat{\psi})^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{V}(\widehat{\psi})^{-1}\boldsymbol{y}$ である。

$\widehat{\psi}$ の推定方法について紹介し，小地域の推定精度を高めるために EBLUP $\widehat{\mu}_i(\widehat{\psi})$ が有用であること，特に線形混合モデルのどのような仕組みが推定精度を高めるのに役立つのかについて説明した。また，EBLUP の平均 2 乗誤差の推定，EBLUP に基づいた信頼区間の構成と補正方法，$\boldsymbol{\beta}$ に関する線形仮説に関する GLS 検定の Bartlett 型補正について説明した。

2　線形混合モデルにおけるベイズ情報量規準 (BIC)

次に，$\boldsymbol{X}$ と $\boldsymbol{v}$ についての変数選択規準として BIC, 修正 BIC を導出した。$\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{I}_N$, $\|\boldsymbol{u}\|_{\boldsymbol{A}}^2 = \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}$ に対して, $\widehat{\boldsymbol{\beta}}(\boldsymbol{V}) = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$, $\hat{\sigma}_0^2(\boldsymbol{V}) = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{V})\|_{\boldsymbol{V}^{-1}}^2/N$ とおくと，BIC は，

$$BIC(\widehat{\boldsymbol{\beta}}(\boldsymbol{V}), \hat{\sigma}_0^2(\boldsymbol{V})) = N[\log(2\pi\hat{\sigma}_0^2(\boldsymbol{V})) + 1] + \log|\boldsymbol{V}| + r_{(\boldsymbol{X})}\log(N)$$

で与えられる。$\boldsymbol{V}$ もくしは $\boldsymbol{G}$ が未知母数 $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_d)$ の関数で表されるときには，BIC は

$$BIC(\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{V}}), \hat{\sigma}_0^2(\widehat{\boldsymbol{V}}), \widehat{\boldsymbol{V}}) = N[\log(2\pi\hat{\sigma}_0^2(\widehat{\boldsymbol{V}})) + 1] + \log|\widehat{\boldsymbol{V}}| + (r_{(\boldsymbol{X})} + d)\log(N)$$

と書かれる。ここで，$\widehat{\boldsymbol{\psi}} = (\widehat{\psi}_1, \ldots, \widehat{\psi}_d)$ は，$\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi} = O_p(N^{-1/2})$ なる ψ の推定量で，$\boldsymbol{G}$, $\boldsymbol{V}$ は $\widehat{\boldsymbol{G}} = \boldsymbol{G}(\widehat{\boldsymbol{\psi}})$, $\widehat{\boldsymbol{V}} = \boldsymbol{V}(\widehat{\boldsymbol{\psi}})$ で推定されるものとする。

修正 BIC は，尤度関数を群内分散と群間分散の項に分割し，それぞれに BIC を当てはめることにより導かれ，

$$\begin{aligned} mBIC(\widehat{\boldsymbol{G}}) =& N[\log(2\pi\tilde{\sigma}_1^2) + 1] + \log|\widehat{\boldsymbol{V}}| + \|\tilde{\boldsymbol{y}}_2 - \widetilde{\boldsymbol{X}}_2\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{V}})\|_{\widehat{\boldsymbol{W}}_2^{-1}}^2/\tilde{\sigma}_1^2 \\ & - M + r_{(\widetilde{\boldsymbol{X}}_1)}\log(N - M) + (r_{(\widetilde{\boldsymbol{X}}_2)} + d)\log(M) \end{aligned}$$

で与えられる。記号の具体的な意味については，論文を参照してほしい。

NERM と経時測定 線形混合モデルについて，BIC, 修正 BIC を導出した。数値実験により，真の変数の組を選択する意味において，BIC, 修正 BIC は AIC, CAIC より優れていることがわかった。特に，k が小さいときには，修正 BIC は BIC よりよいようである。真のモデルが変量効果を含む場合と含まない場合を調べたが，いずれの場合も修正 BIC の挙動がよいことが示された。

最後に，京浜急行電鉄本線及び久里浜線沿いの宅地物件について 1997 年から 2001 年までの 5 年間に公表された $1m^2$ 当たりの地価公示価格のデータを利用して，修正 BIC が選択する説明変数を調べてみた。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，NERM と経時測定 線形混合モデルに修正 BIC を適用してみたところ，合理的な線形混合モデルが選択された。