A variable selection criterion for GEE with large cluster sizes

広島大・理　佐藤 倫治

医療, 経済などのあらゆる分野において, 相関のあるデータに関する解析が議論されている. その中でも経時データは, 観測対象者を時間経過に伴って観測した繰り返し測定データのことで, 同一個体間のデータは相関を持ち, 異個体間のデータは独立であるという特徴を持つ. この相関のある経時データを解析する手法の一つに Liang & Zeger (1986) で提案された一般化推定方程式 (GEE) がある. 周辺モデルは以下のように表される.

目的変数ベクトル $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})'$ と説明変数行列 $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{im})'$ の組 $(\boldsymbol{y}_i, \boldsymbol{X}_i)$ が与えられているとする. ただし, $i = 1, \ldots, n$ である. また, $\boldsymbol{x}_{ij}$ は $(p \times 1)$ 説明変数ベクトルであり, $i \neq k$ ならば $\boldsymbol{y}_i$と $\boldsymbol{y}_k$ は独立である. ただし, $\boldsymbol{y}_i$ 内には相関があるとする. また, y_{ij} の密度関数は,

$$f(y_{ij}; \theta_{ij}, \phi) = \exp\left[\{y_{ij}\theta_{ij} - a(\theta_{ij})\}/\phi + b(y_{ij}, \phi)\right],$$

であると仮定する. 平均は $\mathrm{E}[Y_{ij}] = \dot{a}(\theta_{ij})$, 分散は $\mathrm{Var}[Y_{ij}] = \ddot{a}(\theta_{ij})\phi$, と表せる. このとき, 次のモデル, $\eta_{ij} = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}, \theta_{ij} = u(\eta_{ij})$ を考える. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ は未知パラメータ, ϕ は尺度パラメータである. $\boldsymbol{\beta}$ は以下で定義される GEE により推定される.

$$\sum_{i=1}^{n} \boldsymbol{D}_i^{'}(\boldsymbol{\beta})\boldsymbol{V}_i^{-1}(\boldsymbol{\beta})\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\} = \boldsymbol{0}_p.$$

ここで, $\boldsymbol{\mu}_i = (\dot{a}(\theta_{i1}), \ldots, \dot{a}(\theta_{in}))', \boldsymbol{D}_i(\boldsymbol{\beta}) = \boldsymbol{A}_i(\boldsymbol{\beta})\boldsymbol{\Delta}_i(\boldsymbol{\beta})\boldsymbol{X}_i, \boldsymbol{V}_i(\boldsymbol{\beta}) = \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta})\boldsymbol{R}(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta})\phi$, $\boldsymbol{A}_i = \mathrm{diag}\{\ddot{a}(\theta_{i1}), \ldots, \ddot{a}(\theta_{in})\}, \boldsymbol{\Delta}_i = \mathrm{diag}\{\partial\theta_{i1}/\partial\eta_{i1}, \ldots, \partial\theta_{in}/\partial\eta_{in}\}$ である. また, $\boldsymbol{R}(\boldsymbol{\alpha})$ は作業用相関行列であり, $\boldsymbol{\alpha}$ は相関パラメータである.

Inatsu & Imori (2013) では, GEE を用いたモデルに対する変数選択規準を提案した. $\boldsymbol{z}_i$ を $\boldsymbol{y}_i$ と独立に同一の分布に従うデータとしたとき, モデルの良さを測る尺度として以下のリスクを用いた.

$$\mathrm{E}_y\left[\mathrm{E}_z\left[\sum_{i=1}^{N}(\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_i)'\mathrm{Cov}[\boldsymbol{z}_i]^{-1}(\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_i)\right]\right]. \tag{1}$$

そして (1) の有効な推定量を求めることで, 相関構造の影響を反映させた変数選択規準として,

$$\mathrm{PMSEG} = \sum_{i=1}^{n}\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})\}'\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{R}_0^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})\} + 2p,$$

を提案した. ここで, $\hat{\boldsymbol{\beta}}$ と $\hat{\boldsymbol{\beta}}_f$ はそれぞれ候補モデルとフルモデルにおける GEE 推定量であり, $\boldsymbol{R}_0$ は真の相関行列の一致推定量である.

Inatsu & Imori (2013) では, 標本数 n だけが大きくなる大標本での変数選択規準を導出していた. 本発表では, 標本数 n だけでなく, 観測時点数 m も共に大きくなる大標本高次元での変数選択規準の導出について話す.

参考文献

[1] Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*. **73**, 13-22.

[2] Inatsu, Y. & Imori, S. (2013). Model selection criterion based on the prediction mean squared error in generalized estimating equations. *TR* 13-10, *Statistical Research Group* , Hiroshima University, Hiroshima.