説明変数と相関構造を同時に選択した場合の一般化推定方程式法による予測について

広島大・理　佐藤 倫治
大阪医大　伊藤 ゆり
大阪医大　福井 敬祐

実データ解析において, 同一個体内で相関を持つデータの解析は重要な問題である. 中でも経時測定データは, 各個体ごとに時間経過に伴い繰り返し測定されたデータであり, 同一個体内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. 臨床試験データや成長データも経時測定データの例として知られており, 医学研究や疫学研究において盛んに扱われている.

経時データを解析する手法の一つに, Liang and Zeger (1986) で提案された一般化推定方程式 (GEE) がある. 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 $(\boldsymbol{y}_i, \boldsymbol{X}_i)$ が観測されているとし, 目的変数 $\boldsymbol{y}$ を予測したいとする. ここで, $i = 1, \ldots, n$ であり, n は個体数, m は時点数を意味する. 経時データの特徴として, $i \neq k$ ならば $\boldsymbol{y}_i$ と $\boldsymbol{y}_k$ は独立であると仮定する. このとき, 一般化推定方程式法では, 目的変数に周辺モデルとして一般化線形モデルを仮定する. つまり, y_{ij} の密度関数は,

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

と仮定する. 平均は $\mathrm{E}[Y_{ij}] = \dot{a}(\theta_{ij}) = \mu_{ij}$, 分散は $\mathrm{Var}[Y_{ij}] = \ddot{a}(\theta_{ij})\phi$, と表せる. このとき, 各目的変数に対してモデル, $g(\mu_{ij}) = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}$ を考える. ここで $g(\cdot)$ はリンク関数と呼ばれる関数である. そしてこの回帰係数 $\boldsymbol{\beta}$ を以下の GEE で推定する.

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

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.

一般化推定方程式では, 真の相関行列の代わりに, 解析者が自由に選ぶことができる作業用相関行列を用いることでモデル化を行い推定を行う. 大標本モデルにおいては, 作業用相関構造として誤った構造を用いたとしても回帰係数の推定量が一致性や漸近正規性などのいい性質を持つことが知られているが, 大標本高次元データにおいては, 相関構造を誤った場合一致性が保証されなくなる. 一方, 作業用相関構造として真の相関構造を選択した場合, 大標本高次元データにおいても回帰係数の一致性が保証される.

Inatsu & Imori (2013) では, 大標本データにおける, 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) の漸近不偏な推定量を求めることで, 変数選択規準を提案した.

本発表では, 大標本高次元データにおける Inatsu and Imori (2013) で提案されたモデル選択規準の漸近性質について話す. Inatsu and Imori (2013) では言及されていなかったが, 説明変数と相関構造を同時に選択できる規準であることがわかったため, 数値シミュレーションを通して評価する. またこのモデル選択規準を用いた実用例を示す.