離散分布の経時測定データにおける線形な変化係数の推測について

広島大学原爆放射線医科学研究所 佐藤健一

広島大学理学研究科 柳原宏和

札幌医科大学医療人育成センター 加茂憲一

要　旨 回帰モデルにおいて, 変化係数は時間とともに変化する説明変数の効果を評価でき, 視覚的にも理解しやすいために広く利用されるようになってきた. その推定値は固定された時間ごとに近傍データを用いて回帰をすることで容易に得られる. しかしながら, 時間についての連続性がないために観測時間を通した曲線としての信頼区間の構成および検定に困難があった. 本稿では Satoh and Yanagihara (2008) で提案された連続分布の経時測定データにおける線形な変化係数に関する推測方法を, 成長曲線モデルおよび変量効果モデルとの関連を述べながら, 一般化推定方程式を利用し離散分布に適用することを試みる. この手法の利点のひとつとして汎用的なソフトウェアを利用して計算できることが挙げられる. 解析例ではロジスティック回帰モデルにおける変化係数の推定を紹介する.

1. はじめに

データの観測環境および測定環境の整備と電子化にともない医学・生物学分野を中心に繰り返し測定データおよび計測時間の情報も付与された経時測定データが記録される機会が増え, その解析の重要性が高まってきている. 従来の観測終了時点のみのデータを用いた評価に代わって途中経過データも利用することで, より詳細な評価が可能となる. Lindsey (1993) には経時測定

データの例が多数掲載されているので参考にされたい．ここでは仮に, 8 歳, 10 歳, 12 歳および 14 歳の 4 時点で観測された少年少女の身長の成長データを考え, 身長に性差があるかどうかに関心があるとする．今, ある年齢に着目すれば, 身長 y は少年ならば 1, 少女なら 0 を取る説明変数 a および観測誤差 ε を用いて,

$$y = \beta_1 + \beta_2 a + \varepsilon, \tag{1}$$

と記述でき, 少年の効果, もしくは性差の有無は β_2 の有意性によって評価される．しかしながら, このデータは経時測定データなので, 時点 $t \in \{8, 10, 12, 14\}$ における観測値 $y(t)$ は時間とともに変化する回帰係数, すなわち変化係数 $\beta_1(t)$ および $\beta_2(t)$ を用いて,

$$y(t) = \beta_1(t) + \beta_2(t)a + \varepsilon(t), \tag{2}$$

と表現するのが自然である．そして性差の有無は時間の関数である変化係数 $\beta_2(t) \equiv 0$ の検定によって評価されるであろう．変化係数は Hastie and Tibshirani (1993) によって提案され, 時間とともに変化する説明変数の効果を視覚的にも評価できるため, 経時測定データの解析に極めて有用である．

変化係数の推定は時間を固定するたびに近傍データを用いて回帰を繰り返すカーネル平滑化法が一般的であり, Hoover *et al.* (1998) および Satoh and Ohtaki (2006) などによって議論されている．しかし, この方法では時間の連続性がないために関数としての信頼区間の構築および測定時点間の相関構造の記述が極めて困難であった．そこで, Satoh and Yanagihara (2008) は変化係数に対して時間に関する線形性を仮定することを提案し, これが従来の成長曲線モデル (例えば, Potthoff and Roy (1964), Vonesh and Carter (1987)) で記述できることを示した．そして, 変化係数が陽な形で推定可能なことを示し, 時間の関数としての信頼区間の構成および検定方法を理論的に導出した．

成長曲線モデルは Laird and Ware (1982) によって提案された変量効果モデルにおいて, 特に時間と共変量の交互作用を考慮したモデルと考えることができるため, Patterson and Thompson (1971) によって提案された制限付最尤推定法を利用して推定可能である. 変量効果モデルの解析パッケージは主要な統計ソフトウェアにおいて実装されていることが多いため, 変量効果モデルのもとで線形な変化係数は容易に推定できる.

一方, 離散分布の経時測定データに対する変化係数の推定についての研究は一般化線形モデル (Dobson (1990)) を経時測定データに拡張した Liang and Zeger (1986) による一般化推定方程式を利用するのが基本となる. 一般化推定方程式は連続分布のみを対象とした Laird and Ware (1982) の変量効果モデルと異なり, 離散分布も扱えるため医学分野をはじめ広く利用されている (Nakashima *et al.* (2003), Nakashima, Neriishi and Minamoto (2008)). 一般化推定方程式における変化係数の推定もカーネル平滑化が利用されており, Wang, Carroll and Lin (2005) によって議論されている. また, Qu and Li (2005) では一般化推定方程式を改良した 2 次推定関数を利用した変化係数の推定を行っている. いずれの研究においても, 変化係数の信頼区間は各点ごとに構成されている.

そこで, 本稿では離散分布の経時測定データおいて変化係数を線形なものに限定し, 一般化推定方程式を用いて推定することで関数として信頼区間および検定を考える. 2 章では一般化線形モデルのもとで線形な変化係数を導入し, その推測方法を提案する. 3 章において, 繰り返し観測された二値反応データを用いて変化係数の解析例を示し, これに関連した数値実験を 4 章で行う.

2. 線形な変化係数の推測

Satoh and Yanagihara (2008) は成長曲線モデルにおいて線形な変化係数の推測問題を議論したが, ここでは経時離散データに対して一般化線形モデルの記述にしたがって線形な変化係数について説明し, その推定, 信頼区間の構成および検定方法を提案する.

個体 $i = 1, \cdots, n$ の時点 $t = t_{i1}, \cdots, t_{ip_i}$ における観測値を $y_i(t)$ とする. ただし, p_i は個体 i の観測時点数であり, 個体間の観測値は独立とする. 個体ごとの観測時点が揃っている場合, すなわち $p_1 = \cdots = p_n = p$ かつ $j = 1, \cdots, p$ に対して $t_{1j} = \cdots = t_{nj} = t_j$ が成り立つ場合, 測定時点はバランス型とよばれ, 成り立たない場合はアンバランス型とよばれる. バランス型の場合であっても Brockwell and Davis (1991) で議論されるような一般的な時系列データと異なり等間隔である必要はない. したがって, バランス型の場合は $\boldsymbol{y} = (y(t_1), \cdots, y(t_p))'$ とおけば観測時点間の相関 $\mathrm{Cor}(\boldsymbol{y})$ を考えることができる.

観測値の分布に指数型分布族を仮定し, 時点 t での期待値 $\mu_i(t) = E[y_i(t)]$ のリンク関数 g が k 個の時間に依存しない説明変数 $\boldsymbol{a}_i = (a_{i1}, \cdots, a_{ik})$ によって記述できるとする, すなわち,

$$g(\mu_i(t)) = \boldsymbol{a}_i' \boldsymbol{\beta}(t) = \sum_{j=1}^{k} a_{ij} \beta_j(t), \tag{3}$$

ただし, $\boldsymbol{\beta}(t) = (\beta_1(t), \cdots, \beta_k(t))'$ は時間とともに変化する効果をあらわす変化係数である. 指数型分布族は連続分布の正規分布をはじめ, ガンマ分布, 離散分布の 2 項分布, ポアソン分布など多くの分布を含んでいる. ここで, 時間に関する q 個の基底をあらわす説明変数 $\boldsymbol{x}(t) = (x_1(t), \cdots, x_q(t))'$ を考え, 次の線形構造を変化係数に仮定する.

$$\boldsymbol{\beta}(t) = \Theta \boldsymbol{x}(t), \tag{4}$$

ただし, $\Theta = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_k)' = (\theta_{ij})$ は $k \times q$ 未知パラメータ行列. 例えば, 変化係数に直線を適用するなら $\boldsymbol{x}(t) = (1, t)'$. 測定時点数が多く滑らかな曲線を仮定したい場合にはBスプライン基底またはガウス基底などの散布図平滑化に利用される基底を用いることも可能である. 基底については, 例えば, Satoh, Yanagihara and Ohtaki (2003) およびRuppert, Wand and Carroll (2003) の3章を参照されたい. したがって, 線形な変化係数であっても十分自由度が高い非線形の曲線群が表現できる.

まとめると, $g(\mu_i(t))$ の線形構造は次式のようにかける.

$$g(\mu_i(t)) = \boldsymbol{a}_i' \Theta \boldsymbol{x}(t). \tag{5}$$

変化係数に時間に関する線形性を仮定することで $g(\mu_i(t)) = \boldsymbol{a}_i' \Theta \boldsymbol{x}(t)$ を導いたが, $\boldsymbol{b}_i = \Theta' \boldsymbol{a}_i$ とおけば $g(\mu_i(t)) = \boldsymbol{b}_i' \boldsymbol{x}(t)$ とかけるので $\boldsymbol{b}_i$ は時間に関する曲線をあらわす個体ごとの回帰係数となることに注意する.

式 (5) で使われる共変量について成長曲線モデルの用語を使い補足する. $\boldsymbol{a}_i$ は個体間の差異を説明するので個体間共変量, $\boldsymbol{x}(t)$ は個体内の差異を説明するので個体内共変量とよばれる. よって, $\boldsymbol{a}_i$ は時間に依存せず, $\boldsymbol{x}(t)$ は時間に依存する共変量である. さらに時間に依存する個体間共変量がある場合, 2つの対応が考えられる. 一つは $\boldsymbol{x}(t)$ に含めて $\boldsymbol{x}_i(t)$ と記述する方法である. この場合は変化係数も個体ごとに異なり, $\boldsymbol{\beta}_i(t) = \Theta \boldsymbol{x}_i(t)$ となるが, 変化係数を記述する基底として使われるため, $\boldsymbol{x}_i(t)$ に含められた共変量の効果を変化係数として評価することはできない. 一方, $\boldsymbol{a}_i$ に含めて $\boldsymbol{a}_i(t)$ とする場合は後述するように交互作用として変数を展開することで推定可能である. どちらの方法をとるにしてもモデルが推定された後で新しい個体の予測曲線を考える場合, 時間に依存する個体間共変量がある場合には, それが測定されるまで, 観測値の予測値, もしくは理論値を得ることはできない. 逆に, 時間に依存す

る個体間共変量がなければ測定開始時点の情報 $\boldsymbol{a}$ によって, 個体ごとの予測曲線および変化係数が直ちに得られるという利点がある.

リンク関数に仮定した線形構造は説明変量と時間に関する基底の交互作用の 1 次結合として次式のように展開できる.

$$g(\mu_i(t)) = \sum_{j=1}^{k}\sum_{l=1}^{q}\theta_{jl}\{a_{ij}x_l(t)\}. \tag{6}$$

観測時点 t が与えられていれば, $\{a_{il}x_m(t)\}$, $l = 1, \cdots, k$, $m = 1, \cdots, q$ は既知であり, 説明変数は kq 個となる. そこで, 新たに未知パラメータと説明変数を $\boldsymbol{\xi} = (\boldsymbol{\theta}_1', \cdots, \boldsymbol{\theta}_k')'$ および $\boldsymbol{z}_i(t) = (a_{i1}\boldsymbol{x}(t)', \cdots, a_{ik}\boldsymbol{x}(t)')'$ とそれぞれおけば $g(\mu_i(t)) = \boldsymbol{z}_i(t)'\boldsymbol{\xi}$ とかけ, $\boldsymbol{\xi}$ の推定量 $\hat{\boldsymbol{\xi}}$ は Liang and Zeger (1986) によって提案された次の一般化推定方程式の解として求めることができる.

$$S(\boldsymbol{\xi}) = \sum_{i=1}^{n} D_i'V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}_{kq}, \tag{7}$$

ただし, $D_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\xi}'$, $\boldsymbol{\mu_i} = (\mu_i(t_1), \cdots, \mu_i(t_p))'$, $V_i = \Delta^{1/2}R\Delta^{1/2}$, $\Delta_i = \phi U_i$, $U_i = \mathrm{diag}(\mathrm{var}(y_i(t_1)), \cdots, \mathrm{var}(y_i(t_p)))$, R は $\mathrm{Cor}(\boldsymbol{y})$ に対して仮定された作業相関行列, ϕ は超過分散を考慮した尺度パラメータである. 一般化推定方程式は作業相関行列の与え方が正しくなくても平均構造が正しく記述されていれば, 不偏推定方程式 (例えば, Wilks (1962) の 12 章) となるため推定量の一致性が保障される. さらに推定量は漸近正規性を持ち, その漸近分散はモデル分散, もしくはナイーブ分散とよばれ,

$$\mathrm{var}(\hat{\boldsymbol{\xi}}) = \left(\sum_{i=1}^{n} D_i'V_i^{-1}D_i\right)^{-1} = \begin{pmatrix} \Omega_{11} & \cdots & \Omega_{1k} \\ \vdots & \ddots & \vdots \\ \Omega_{k1} & \cdots & \Omega_{kk} \end{pmatrix},$$

で与えられる. ただし, $\mathrm{cov}(\hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\theta}}_j) = \Omega_{ij}$, $\hat{\boldsymbol{\theta}}_j$ は $\boldsymbol{\theta}_j$ の推定量である. したがって, 漸近的に,

$$\hat{\boldsymbol{\theta}}_j \sim N_q(\boldsymbol{\theta}_j, \hat{\Omega}_{jj}), \qquad (n \to \infty), \tag{8}$$

とかける. また, 変化係数の推定量は $\hat{\beta}_j(t) = \hat{\boldsymbol{\theta}}_j' \boldsymbol{x}(t)$ となる.

観測値が連続分布の場合, 変量効果モデルおよび成長曲線モデルを用いた推定も可能である. 変量効果モデルを利用する場合はランダム効果 $\boldsymbol{\delta}$ を個体ごとの回帰係数に加え, $y_i(t) = \boldsymbol{b}_i' \boldsymbol{x}(t) + \boldsymbol{\delta}_i' \boldsymbol{x}(t) + \varepsilon(t)$, $\varepsilon_i(t) \sim_{i.i.d} N(0, \sigma^2)$, $\boldsymbol{\delta}_i \sim_{i.i.d} N_q(\mathbf{0}, \Delta)$ と仮定することが多い. よって, 統計ソフトウェアではランダム効果を導入する変数として $\boldsymbol{x}(t)$, もしくはその一部を指定する. 特に, 変量効果モデルの特別な場合としてバランス型であれば Potthoff and Roy (1964), アンバランス型であれば Vonesh and Carter (1987) の成長曲線モデルによって陽な形で推定量が得られる. また, 連続変量かつバランス型であれば一般化推定方程式でも変量効果モデルでも推定可能である. さらに, 展開式 (6) から分かるように, Θ または $\boldsymbol{\xi}$ のいくつかの成分を必要に応じて 0 としたい場合は対応する交互作用の説明変数をモデルに取り入れないことで同様に推定できる. 例えば, 変化係数に対して多項式曲線を仮定する場合であれば説明変数ごとに異なる次数を当てはめることも可能である. どの成分を 0 とすべきかという問題は一般的には変量選択の問題として議論され, 一般化推定方程式では Pan (2001) が同時密度関数による尤度を擬似尤度に置き換えることで変量選択規準 QIC を提案している. また, 変量効果モデルでは Vaida and Blanchard (2005) および Liang, Wu and Zou (2008), 成長曲線モデルでは Satoh, Kobayashi and Fujikoshi (1997) および Fujikoshi *et al.* (2003) などに議論されている.

次に, 変化係数の信頼区間について考える. $\boldsymbol{\xi}$ の分散に含まれる未知パラメータ $(\boldsymbol{\xi}, R, \phi)$ をそれぞれの推定量で置き換えることで Ω_{jj} の推定量, $\hat{\Omega}_{jj}$ が得られる. よって, 変化係数の分散の推定量は $\hat{\lambda}_j(t) = \boldsymbol{x}(t)' \hat{\Omega}_{jj} \boldsymbol{x}(t)$ で与えら

れるので, Rao (1973), 1 章 1f 節の Cauchy-Schwarz の不等式を用いると,

$$\sup_{t\in\mathbb{R}} \frac{\{\hat{\beta}_j(t)-\beta_j(t)\}^2}{\hat{\lambda}_j(t)} \leq \sup_{\boldsymbol{x}\in\mathbb{R}^q} \frac{\{(\hat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)'\boldsymbol{x}\}^2}{\boldsymbol{x}'\hat{\Omega}_{jj}\boldsymbol{x}} \leq (\hat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)'\hat{\Omega}_{jj}^{-1}(\hat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j), \quad (9)$$

が成り立つ. この不等式により, 変化係数 $\hat{\beta}_j(t)$ の信頼区間の構成問題は対応するパラメータベクトル $\hat{\boldsymbol{\theta}}_j$ の問題に置き換わる. 変化係数は時間の関数であるが, 右辺はすでに時間に依存しないことに注意する. $\hat{\boldsymbol{\theta}}_j$ の漸近正規性により, 右辺は自由度 q のカイ二乗分布に近似的に従う. したがって, $c_{q,\alpha}$ を χ_q^2 の上側 $\alpha\times 100$ パーセント点とすると, すなわち $\Pr(\chi_q^2 \geq c_{q,\alpha}) = \alpha$ とすると, 変化係数の $\alpha\times 100$ パーセント信頼区間は次式で与えられる.

$$\mathcal{I}_{j,\alpha}(t) = \left[\hat{\beta}_j(t) - \sqrt{\hat{\lambda}_j(t)c_{q,\alpha}}, \quad \hat{\beta}_j(t) + \sqrt{\hat{\lambda}_j(t)c_{q,\alpha}}\right]. \quad (10)$$

それゆえ, 信頼区間の被覆確率は漸近的に $\Pr(\beta_j(t)\in\mathcal{I}_{\alpha,j}(t) : {}^\forall t\in\mathbb{R}) \geq 1-\alpha$ を満たす. 一方, 固定された t ごとに $\{\hat{\beta}_j(t)-\beta_j(t)\}^2/\hat{\lambda}_j(t) \xrightarrow{d} \chi_1^2$ が成り立つので各測定時点ごとの信頼区間を構成することもできるが, $\mathcal{I}_{j,\alpha}(t)$ に比べて狭くなることに注意する.

最後に変化係数の有意性を評価する. 変化係数がすべての時点で 0 であるという帰無仮説は,

$$H_0 : \beta_j(t) = 0 \quad ({}^\forall t\in\mathbb{R}). \quad (11)$$

とかける. これは対応する説明変数 a_j が観測値に対して効果を持たないことを意味する. 素朴な検定統計量として $T_j(t) = \hat{\beta}_j(t)^2/\hat{\lambda}_j(t)$ $(j=1,\cdots,k)$ を考えると, $\Pr(T_j(t)\leq x : {}^\forall t\in\mathbb{R})$ を評価する必要がある. ここで,

$$\Pr(T_j(t)\leq x : {}^\forall t\in\mathbb{R}) = \Pr\Big(\max_{t\in\mathbb{R}} T_j(t) \leq x\Big), \quad (12)$$

が成り立つことに注意すれば信頼区間の構成と同様に $\max_{t\in\mathbb{R}} T_j(t) \leq T_j$ が成り立つことが分かる. ただし, $T_j = \hat{\boldsymbol{\theta}}_j'\hat{\Omega}_{jj}^{-1}\hat{\boldsymbol{\theta}}_j$. T_j は帰無仮説のもとで漸近

的に χ^2_q に従うので, 無仮説 H_0 は, $T_j > c_{q,\alpha}$ ならば有意水準 $100\alpha\%$ で棄却される, もしくは, $\Pr(\chi^2_q > T_j)$ から p 値を求めることが可能である.

3. 経時離散データの解析例

経時的に測定された離散分布の例として Koch *et al.* (1977) によって紹介された二値反応データを表 1 に示す. データは患者の診断群別 (重度, 軽度), 投薬群別 (標準薬, 新薬) の 3 時点 (1,2,4 週) における二値反応 (正常 N, 異常 A) である. ここでの関心は標準薬に対する新薬の効果であり, 時間とともに変化する効果を変化係数として評価することが目標となる.

まず, 2 章での記号にあわせてデータを表現し, 一般化線形モデルに含まれるロジスティック回帰モデルを構築する. 患者 i の第 t 週における観測値を,

$$y_i(t) = \begin{cases} 1 & (\text{時点 } t \text{ での反応が } N) \\ 0 & (\text{時点 } t \text{ での反応が } A) \end{cases}, \qquad i = 1, \cdots, 340, \qquad t \in \{1, 2, 4\},$$

$k = 3$ 個の説明変数 $\boldsymbol{a} = (a_1, a_2, a_3)'$ を,

$$a_1 = 1, \qquad a_2 = \begin{cases} 1 & (\text{軽度}) \\ 0 & (\text{重度}) \end{cases}, \qquad a_3 = \begin{cases} 1 & (\text{新薬}) \\ 0 & (\text{標準薬}) \end{cases},$$

とする. したがって, 定数項 a_1 は重度かつ標準薬に対応する. ここで, 観測時点はバランス型なので $\boldsymbol{y} = (y(1), y(2), y(4))'$ の相関行列 $\mathrm{Cor}(\boldsymbol{y})$ が定義できる. 観測値の期待値, もしくは反応確率 $p(t) = E[y(t)] = \Pr\{y(t) = 1\}$ に対して, リンク関数としてロジット関数 $g(s) = \log(s/(1-s))$, 測定時点数が 3 点なので変化係数に直線を仮定し $\boldsymbol{x}(t) = (1, t)'$, Θ を 3×2 の未知パラメータ行列とすれば次の線形構造が仮定できる.

$$\log \frac{p_i(t)}{1 - p_i(t)} = \boldsymbol{a}_i' \Theta \boldsymbol{x}(t). \tag{13}$$

線形構造の部分を時間と説明変数の交互作用に展開すると $a_{i1} = 1$ なので,

$$\boldsymbol{a}_i' \Theta \boldsymbol{x}(t) = \theta_{11} + \theta_{12} t + \theta_{21} a_{i2} + \theta_{22} a_{i2} t + \theta_{31} a_{i3} + \theta_{32} a_{i3} t. \tag{14}$$

したがって, 未知パラメータを推定するために必要なデータは交互作用の項も合わせて表 2 の形式で作成する必要がある. 注意であるが, 定数項に対応する a_{i1} と時間との交互作用をあらわす $a_{i1}t$ は常に $a_{i1}t = t$ となるため, 新たにデータ列を作成せず t を使えばよい. ロジスティック回帰の場合は (7) 式の一般化推定方程式において, $\mathrm{var}(y_i(t)) = \mu_i(t)(1-\mu_i(t))$, $D_i = U_i Z_i$ および $Z_i = (\boldsymbol{z}_i(1), \boldsymbol{z}_i(2), \boldsymbol{z}_i(4))'$ とかけることに注意すれば, 以下の推定値を得ることができる.

$$(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}_3) = \begin{pmatrix} -1.718 & 1.555 & -0.640 \\ 0.371 & -0.117 & 0.705 \end{pmatrix}, \hat{\Omega}_{33} = \begin{pmatrix} 0.102 & -0.038 \\ -0.038 & 0.018 \end{pmatrix}.$$

推定には統計ソフトウェア SPSS 15.0 を利用し, 被験者変数および被験者内変数として表 2 の i および t, 尺度パラメータ $\phi = 1$, $\mathrm{Cor}(\boldsymbol{y})$ に対する作業相関行列として無構造をそれぞれ指定した. 各測定時点間の相関係数は $\widehat{\mathrm{Cor}}(y(1), y(2)) = 0.067$, $\widehat{\mathrm{Cor}}(y(2), y(4)) = -0.056$ および $\widehat{\mathrm{Cor}}(y(4), y(1)) = -0.026$ と推定された. 図 1(a) に推定された反応確率の曲線を診断群別投与群別に示す. 投与群に依らず, 診断群としては重度よりも軽度の方が観測期間を通して反応確率が高く, 診断群を固定した場合は標準薬よりも新薬を投与した群が反応確率および上昇率が高いのが見て取れる. 新薬に対する変化係数の推定値は $\hat{\beta}_3(t) = \hat{\boldsymbol{\theta}}_3' \boldsymbol{x}(t) = -0.640 + 0.705t$ で与えられるが, 反応確率のロジット関数に対する寄与をあらわしているためこのままでは解釈が難しい. そこで, 医学分野でよく利用されるオッズ比を利用して変化係数を解釈することにする. 新薬の標準薬に対するオッズ比は,

$$\frac{p(t|a_3 = 1)/\{1 - p(t|a_3 = 1)\}}{p(t|a_3 = 0)/\{1 - p(t|a_3 = 0)\}} = \exp(\beta_3(t)) \tag{15}$$

とかけるので 95% 信頼区間は指数関数の単調性から,

$$\mathcal{I}_{3,\alpha}(t) = \left[\exp\left\{ \hat{\beta}_3(t) - \sqrt{\hat{\lambda}_3(t) c_{q,\alpha}} \right\}, \quad \exp\left\{ \hat{\beta}_3(t) + \sqrt{\hat{\lambda}_3(t) c_{q,\alpha}} \right\} \right]. \tag{16}$$

で与えられる. ただし, $q = 2$, $\alpha = 0.05$ および $c_{q,\alpha} = 5.991$ である. 図 1 の (b) に変化係数の推定値から得られる新薬の標準薬に対するオッズ比と信頼区間を示す. 新薬の効果がない場合のオッズ比は 1 になることに注意する. オッズ比の 95% 信頼区間の下限が観測期間を通してほぼ 1 より大きい. より厳密に帰無仮説 $\beta_3(t) \equiv 0$ の検定を行うと $T_3 = \hat{\boldsymbol{\theta}}_3' \hat{\Omega}_{33}^{-1} \hat{\boldsymbol{\theta}}_3 = 57.054$ となり, これが帰無仮説のもとで自由度 2 のカイ二乗分布に従うことから, 上側 5% 点の 5.991 と比較すれば帰無仮説は高水準で棄却できることが分かる. このように, $\boldsymbol{\theta}$ の各成分の有意性だけに着目するのではなく, 変化係数としてまとめてベクトルで評価することで対応する説明変数の効果が評価できる.

4. 数値実験

本稿で提案した変化係数の推測は (9) 式で与えられる不等式に基づいて構成されている. 変化係数の信頼区間および有意性を曲線として評価するためには (9) 式の左辺の分布を求める必要があるが, 一般的には解析的に導出することができない. しかし, 左辺の分布は右辺の時間によらない分布で上から評価でき, さらに漸近的にはカイ二乗分布にしたがうことが分かる. それゆえ, 我々は左辺の分布の代わりに右辺の分布を利用して変化係数の信頼区間などを評価する方法を提案した. ところが, 不等式の中間の式も考慮すれば変化係数の信頼区間は保守的になる. そこで, 数値実験によって (9) 式の左辺の分布を調べ, 理論分布との乖離を調べることにする.

簡単のため, 3 章で解析したデータの説明変数などを利用することを考える. まず, 説明変数および測定時点は表 1 または表 2 で示す値を使い, 個体間および測定時点間の反応を独立と仮定する. $\boldsymbol{x}(t) = (1, t)'$ とし, (13) 式で与

えられる平均構造において真のパラメータ行列を,

$$\Theta = \begin{pmatrix} -1.718 & 1.555 & 0 \\ 0.371 & -0.117 & 0 \end{pmatrix},$$

とする. よって, $\beta_3(t) = 0$ ($^\forall t \in \mathbb{R}$) となる. この帰無仮説のもとで, $T_3(t) = \hat{\beta}_3^2(t)/\hat{\lambda}_3(t)$ の上限分布 $\sup_{t\in\mathbb{R}} T_3(t)$ と, その近似分布 $T_3 = \hat{\boldsymbol{\theta}}_3' \hat{\Omega}_{33}^{-1} \hat{\boldsymbol{\theta}}_3$ の乖離を評価する. 理論的には常に $\sup_{t\in\mathbb{R}} T_3(t) \leq T_3$ が成り立つ. したがって, 近似分布の上側 $\alpha\%$ 点は上限分布のそれより必ず大きくなる. ここでは, 上側 95% 点について調べる.

帰無仮説のもとでの上限分布は以下のモンテカルロ法により求める. 1)340×3 個の反応データの乱数を (13) 式から生成する, 2) 一般化推定方程式を利用して $\hat{\beta}_3(t)$ と $\hat{\lambda}_3(t)$ を求める, 3)$\sup_{t\in\mathbb{R}} T_3(t)$ を求める, 1)-3) を $20,000$ 回繰り返す. 一般化推定方程式については, 作業相関行列として無構造を用い, 尺度パラメータを 1 に固定した. また, 3) の上限は変化係数に直線を仮定するので, 式 (9) の分子分母は高々t の 2 次関数となる. したがって, 極値は最大で 3 点あるので $t = \pm\infty$ における収束値と合わせて比較することで最大値もしくは上限を求めた. 図 2 に上限のヒストグラムと漸近的に近似分布がしたがう自由度 2 のカイ二乗分布の密度関数を示す. 近似分布の 95% 点は $c_{2,0.05} = 5.992$ であるが, 上限分布においてこれより大きい値は 310 個あるので 98.45% に該当した. それゆえ, 式 (9) に基づいて構成される信頼区間および検定はかなり保守的であることが数値実験によって示された. 一方, 固定された測定時点ごとに構成される信頼区間は自由度 1 のカイ 2 乗分布を利用する. このときの近似分布における 95% 点は $c_{1,0.05} = 3.841$ となるが, 上限分布では 94.76% に対応していたために僅かに過小評価であるが近い値となっていた. 参考のために図 2 に自由度 1 のカイ二乗分布の密度関数を示したところ, 上限分布のヒストグラムをよく近似していた.

5. おわりに

経時測定データにおける変化係数の推測方法を離散分布に適用した. 変化係数を線形なクラスに限定することで説明変数と時間に関する共変量の交互作用として扱え, 一般的な統計ソフトウェアを利用して容易に推定できるようになった. また, その信頼区間および検定についても線形性により従来の多変量解析の理論が利用できた. 変化係数が提案される以前から時間と共変量の交互作用を回帰に取り入れる試みはあったと思われるが, 変化係数という形で整理しなおすことで視覚的にも分かりやすくデータの解釈をする上で有益な情報に成り得る. 経時測定データが得られた場合には成長曲線モデル, 変量効果モデルおよび一般化推定方程式を利用し, 本稿で論じた線形な変化係数の推測を行うことで, 最終時点のみのデータ, もしくは測定時点ごとのデータ解析では見過ごされていた知見が得られる可能性もある.

謝　辞 この研究の一部は広島大学藤井研究助成基金, 文部科学省科学研究費若手研究 (B) 課題番号 19700265 および 18790398 の援助を受けています.

参　考　文　献

Brockwell, P. J. and Davis, R. A. (1991): *Time Series: Theory and Methods* (2nd ed.), Springer.

Dobson, A. (1990): *Introduction to Generalized Linear Models*, Chapman and Hall.

Fujikoshi, Y., Noguchi, T., Ohtaki, M. and Yanagihara, H. (2003): Corrected versions of cross-validation criteria for selecting multivariate regression and growth curve models, *Annals of the Institute of Statistical*

Mathematics **55**, 537–553.

Hastie, T. and Tibshirani, R. (1993): Varying-coefficient models, *Journals of the Royal Statistical Society Series* **B 55**, 757–796.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L. P. (1998): Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data, *Biometrika* **85**, 809–822.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H. and Lehnen, R. G. (1977): A general methodology for analysis of experiments with repeated measurements of categorical data, *Biometrics* **33**, 133–158.

Laird, N. M. and Ware, J. H. (1982): Random-effects models for longitudinal data, *Biometrika* **38**, 963–974.

Liang, H., Wu, H. and Zou, G. (2008): A note on conditional AIC for linear mixed-effects models, *Biometrika* **95**, 773–778.

Liang, K. Y. and Zeger, S. L. (1986): Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

Lindsey, J. K. (1993): *Models for Repeated Measurements*, Oxford University Press.

Nakashima, E., Tsuji, S., Fukuoka, H., Ohtaki, M. and Ito, K. (2003): Estimating the stenosis probabilities in arteriosclerosis obliterans using generalized estimating equations, *Statistics in Medicine* **22**, 2149–2160.

Nakashima, E., Neriishi, K. and Minamoto, A. (2008): Comparison of methods for ordinal lens opacity data from atomic-bomb survivors : Univariate worse-eye method and bivariate GEE method using global odds ratio, *Annals of the Institute of Statistical Mathematics* **60**, 465–482.

Pan, W. (2001): Akaike's information criterion in generalized estimating equations, *Biometrics* **57**, 120–125

Patterson, H. D. and Thompson, R. (1971): Recovery of inter-block information when block sizes are unequal, *Biometrika* **58**, 545–554.

Potthoff, R. F. and Roy, S. N. (1964): A generalized multivariate analysis of variance model useful especially for growth curve problems, *Biometrika* **51**, 313–326.

Qu, A. and Li, R. (2005): Quadratic inference functions for varying-coefficient models with longitudinal data, *Biometrics* **62**, 379–391.

Rao, C. R. (1973): *Linear Statistical Inference and Its Applications* (2nd ed.), John Wiley.

Ruppert, D., Wand, M. and Carroll, R. (2003): *Semiparametric Regression*, Cambridge University Press.

Satoh, K., Kobayashi, M. and Fujikoshi, Y. (1997): Variable selection for the growth curve model, *Journal of Multivariate Analysis* **60**, 277–292.

Satoh, K., Yanagihara, H. and Ohtaki, M. (2003): Bridging the gap between B-spline and polynomial regression model, *Communications in Statistics - Computation and Simulation* **32**, 179-190.

Satoh, K. and Ohtaki, M. (2006): Nonparametric growth curve model with local linear approximation, *Communications in Statistics - Theory and Methods* **35**, 641–648.

Satoh, K. and Yanagihara, H. (2008): Estimation of varying coefficients for a growth curve model, *Technical Report, No. 08-08, Statistical Research Group, Hiroshima University.*

Vaida, F. and Blanchard, S. (2005): Conditional Akaike 's information for mixed-effects models, *Biometrika* **92**, 351–370.

Vonesh, E. F. and Carter, R. L. (1987): Efficients inference for random-coefficient growth curve models with unbalanced data, *Biometrics* **43**, 617–628.

Wang, N., Carroll, R. J. and Lin, X. (2005): Efficient semiparametric marginal estimation for longitudinal/clustered data, *Journal of the American Statistical Association* **100**, 147–157.

Wilks, S. S. (1962): *Mathematical Statistics*, John Wiley.

著者連絡先: 佐藤健一

〒 734-8551 広島市南区霞 1-2-3 総合研究棟 4F 計量生物

TEL: 082-257-5857

E-mail: ksatoh@hiroshima-u.ac.jp

表 1. 患者 340 人の診断群別 (重度, 軽度), 投薬群別 (標準薬, 新薬) の 3 時点 (1,2,4 週) における反応 (正常 N, 異常 A) のパネルデータ. 出展: Koch *et al.* (1977). 例えば, 軽度かつ標準薬で, 1,2,4 週の反応がすべて N, プロファイル NNN の患者数は 16 人.

	1 週	N	N	N	N	A	A	A	A
	2 週	N	N	A	A	N	N	A	A
	4 週	N	A	N	A	N	A	N	A
診断群	投薬群								
軽度	標準薬	16	13	9	3	14	4	15	6
軽度	新薬	31	0	6	0	22	2	9	0
重度	標準薬	2	2	8	9	9	15	27	28
重度	新薬	7	2	5	2	31	5	32	6

表 2. 経時データの入力例. 表 1 において, 軽度かつ標準薬で, 1,2,4 週の症状がすべて N, プロファイル NNN の患者を番号 1, 重度かつ新薬で, プロファイル AAA の患者を番号 340 として, データの入力例を示す.

i	t	$y_i(t)$	a_{i1}	a_{i2}	a_{i3}	$a_{i2}t$	$a_{i3}t$
1	1	1	1	1	0	1	0
1	2	1	1	1	0	2	0
1	4	1	1	1	0	4	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
340	1	0	1	0	1	0	1
340	2	0	1	0	1	0	2
340	4	0	1	0	1	0	4

Captions

図 1. 診断群別投与群別の経時反応データに対する変化係数の推定. 表 1 のデータに対して, 直線を仮定した変化係数を推定した. (a) 反応確率 (N を観測する確率) の推定曲線を診断群別に, 新薬および標準薬を, それぞれ, 実線および点線で示す. (b) 新薬の標準薬に対するオッズ比, すなわち, $\exp(\boldsymbol{\beta}_3(t))$ を実線で, その 95% 信頼区間を点線で示す.

図 2. モンテカルロ法による上限分布のヒストグラムとその近似分布の比較. モンテカルロ法によって帰無仮説 $\beta_3(t) = 0$ のもとで式 (9) の左辺の上限分布 $\sup_{t\in\mathbb{R}} T_3(t)$ を調べ, そのヒストグラムと自由度 2 のカイ二乗分布にしたがう T_3 の密度関数の曲線を示す. 参考のために自由度 1 のカイ二乗分布の密度関数曲線も示した.

Statistical Inference on a Linear Varying Coefficient on Longitudinal Data of Discrete Distribution

Kenichi Satoh[1], Hirokazu Yanagihara[2] and Ken-ichi Kamo[3]

[1]Research Institute for Radiation Biology and Medicine, Hiroshima University

[2]Graduate School of Science, Hiroshima University

[3]Department of Liberal Arts, Sapporo Medical University

Abstract

Varying coefficients can be used for visualizations or interpretations of the covariate effects which might be varying on time axis. The estimator of varying coefficient is usually obtained by kernel smoothing methods. Scince it is essentially the linear regression around fixed time point, constructing a confidence interval or testing null hypothesis for a function of time is difficult. In this paper, we apply an estimating method proposed by Satoh and Yanagihara (2008) on the growth curve model to the discrete distributions using generalized estimating equations. Those new estimators of varying coefficients can be easily calculated by the ordinaly statistical software package. An example of logistic regression analysis with longitudinal data was illustrated.

Key words: generalized estimating equations, generalized linear model,

growth curve model, linear mixed effect model, logistic regression, varying coefficient

*Corresponding auther,
E-mail: ksatoh@hiroshima-u, ac, jp (Kenichi Satoh)

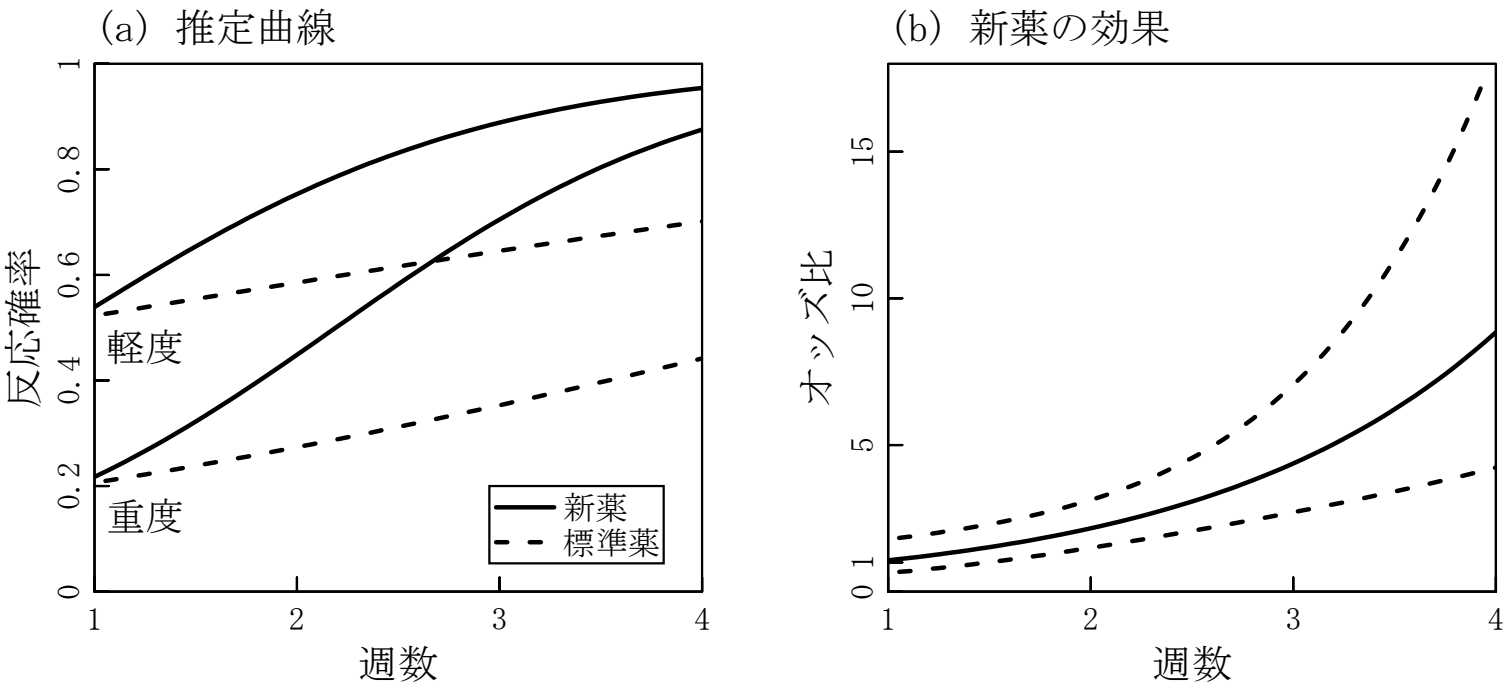
(a) 推定曲線
反応確率
0
0.2
0.4
0.6
0.8
1
軽度
重度
新薬
標準薬
1
2
3
4
週数
(b) 新薬の効果
オッズ比
0 1
5
10
15
1
2
3
4
週数

図1.

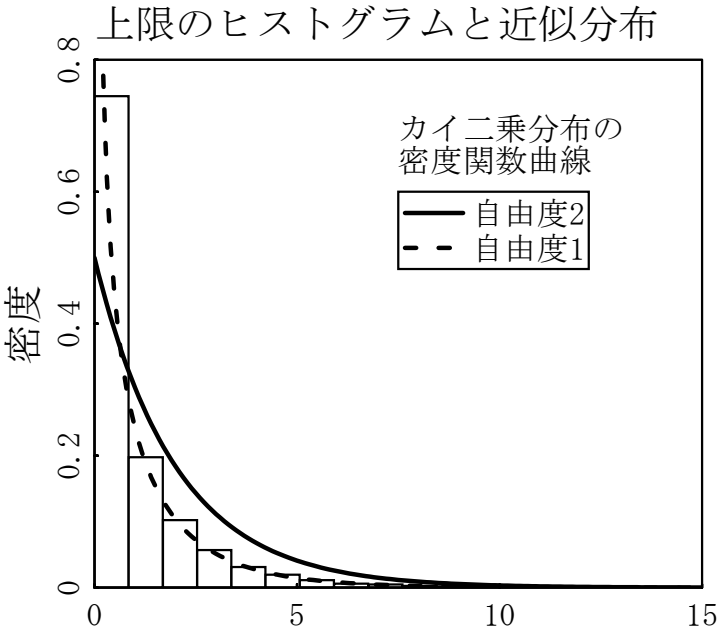
上限のヒストグラムと近似分布
密度
0
0.2
0.4
0.6
0.8
0
5
10
15
カイ二乗分布の
密度関数曲線
自由度2
自由度1

図2.