応用統計学　Vol. 42, No. 1 (2013), 1–10

研究ノート

混合効果モデルを用いたセミパラメトリック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

県立広島大学 経営情報学部　冨　田　哲　治

要　旨　経時測定データにおいて時間とともに変化する回帰係数は変化係数と呼ばれる．Satoh and Yanagihara (2010) は変化係数に線形性を仮定することで，関数としての同時信頼区間を提案した．基底関数として直線が適用された線形な変化係数は解釈が容易であるが，測定時点数が多くなると非線形曲線の近似として充分でないことがある．本稿では，直線を 1 次スプライン関数で補ったセミパラメトリックな変化係数を考え，Brumback et al. (1999) の提案した混合効果モデルを用いた推定方法を適用する．

1. はじめに

時刻 t における観測値を $y(t)$ とし，これを説明する p 個の共変量 $a_1(t),\cdots,a_p(t)$ の回帰係数を $\beta_1(t),\cdots,\beta_p(t)$ とおくと，観測値に対する回帰モデルは次式でかける．

$$y(t)=\sum_{j=1}^{p}\beta_j(t)a_j(t)+\varepsilon(t),\qquad t=t_1,\cdots,t_n. \tag{1}$$

ただし, $\varepsilon(t)\sim N(0,\sigma^2)$. ここで, $\beta(t)$ は時間とともに変化する共変量効果をあらわし, 変化係数と呼ばれ Hastie and Tibshirani (1993) によって提案された．変化係数の推定には従来, 固定された時間近傍ごとに重み付き重回帰を行うカーネル平滑化が利用されてきた(例えば, Hoover et al. (1998), Satoh and Ohtaki (2006), Tonda et al. (2011))．今, 観測値ベクトルを $\boldsymbol{y}=(y_1,\cdots,y_n)'$, 既知計画行列を $A=(\boldsymbol{a}_1,\cdots,\boldsymbol{a}_n)'$ とすれば, 変化係数ベクトル $\boldsymbol{\beta}(t)=(\beta_1(t),\cdots,\beta_p(t))'$ の時刻 t_0 における推定値は $\hat{\boldsymbol{\beta}}(t_0)=\{A'W(t_0)A\}^{-1}A'W(t_0)\boldsymbol{y}$ とかける．ただし，$y_i=y(t_i)$, $a_{ij}=a_j(t_i)$, $\boldsymbol{a}_i=(a_{i1},\cdots,a_{ip})'$, $W(t)=\mathrm{diag}(w_1(t),\cdots,w_n(t))$. ここで，時刻 t_i の観測値に対する重み関数 $w_i(t)$ は時刻 t_i で大きな値を取る釣鐘状の非負関数が使われることが多く，例えば，平均 0, 標準偏差 h の正規密度関数であれば $w_i(t)=(2\pi h^2)^{-0.5}\exp\{-(t-t_i)^2/2h^2\}$ とかける．しかしながら，この方法では変化係数の各点ごとの信頼区間(pointwise confidence interval)しか構成できず，時間の曲線としての同時信頼区間(simultaneous confidence region)の構成は困難であった．これに対して，Satoh and Yanagihara (2010) は成長曲線モデルの平均構造が線形性を持つ変化係数として解釈できることに注目し，線形な変化係数の推定方法とその同時信頼区間の構成法を

示した．また，佐藤・柳原・加茂 (2009) では一般化推定方程式 (Liang and Zeger, 1986) を利用して離散型の観測値に対する線形な変化係数の推定を実現し，冨田・佐藤・柳原 (2010) では空間データに対して線形な基底で記述できる変化係数曲面を適用している．

ここでは，線形な変化係数が線形重回帰モデルの枠組で推定できることを端的に示す．時刻 t に関する基底として長さ q の基底関数ベクトル $\boldsymbol{x}(t)$ を考えると，線形な変化係数はパラメータベクトル $\boldsymbol{b}$ を用いて，$\beta_j(t) = \boldsymbol{x}(t)'\boldsymbol{b}_j,\ j = 1, \cdots, p$ とかける．ここで，時間に関して直線を仮定すれば，$\boldsymbol{x}(t) = (1, t)'$ となる．このとき観測値 y_i の平均は $E[y_i] = \sum_{j=1}^{p} a_{ij}\boldsymbol{x}(t_i)'\boldsymbol{b}_j$ とかけるので，$n \times qp$ の既知行列 $X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)'$, $\boldsymbol{x}_i = \{a_{i1}\boldsymbol{x}(t_i)', \cdots, a_{ip}\boldsymbol{x}(t_i)'\}'$, 長さ qp の回帰係数ベクトル $\boldsymbol{b} = (\boldsymbol{b}_1', \cdots, \boldsymbol{b}_p')'$ を用いて $E[\boldsymbol{y}] = X\boldsymbol{b}$ とかけ，回帰係数ベクトル $\boldsymbol{b}$ は例えば，最小二乗推定量，$\hat{\boldsymbol{b}} = (X'X)^{-1}X'\boldsymbol{y}$ として求めることができる．このように線形な変化係数は説明変数と基底関数の交互作用項を新たな説明変数とすることで推定できるという利点を持つ．以下の 2 章では線形項に加えて非線形性をあらわす基底関数を導入し，3 章においてその最適化を混合効果モデルを利用して行う．そして，4 章では実データに提案手法を適用し，5 章で問題点などを補足する．

2. セミパラメトリックな変化係数

線形な変化係数は測定時点数が多い場合，あるいは観測期間が長い場合に非線形関数を十分に近似できなくなることがある．これに対して非線形関数を表現できる線形な基底を利用する方法が考えられる．ここでは，r 個の節点 $\kappa_1, \cdots, \kappa_r$ を持つ折れ線をあらわす 1 次スプライン基底 $\boldsymbol{z}(t) = \{(t-\kappa_1)_+, \cdots, (t-\kappa_r)_+\}'$ を用いる．ただし，δ_+ は δ が正であれば δ, そうでなければ 0 を取る関数とする．このような非線形曲線を表現する基底は自由度 q を増やせば散布図に対する適合は向上するが，その一方で適合した曲線の解釈は容易ではなくなる．そこで，本稿では解釈が容易な直線をあらわす線形基底 $\boldsymbol{x}(t)$ と非線形性をあらわす線形基底 $\boldsymbol{z}(t)$ を同時に用いるセミパラメトリックな変化係数を考える．節点の個数や配置，また基底関数の種類などのセミパラメトリック回帰モデルの一般的な話題は Ruppert et al. (2003) に詳しく紹介されているので参考にして頂きたい．

$$\beta_j(t) = \boldsymbol{x}(t)'\boldsymbol{b}_j + \boldsymbol{z}(t)'\boldsymbol{u}_j, \qquad j = 1, \cdots, p. \tag{2}$$

ここで，$\boldsymbol{x}(t)$ および $\boldsymbol{z}(t)$ は時刻 t に関する基底関数ベクトル，$\boldsymbol{b}_j$ および $\boldsymbol{u}_j$ は回帰係数ベクトルである．そこで観測値ベクトル $\boldsymbol{y} = (y(t_1), \cdots, y(t_n))'$ に対して，$n \times rp$ の既知計画行列 $Z = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)'$, $\boldsymbol{z}_i = \{a_{i1}\boldsymbol{z}(t_i)', \cdots, a_{ip}\boldsymbol{z}(t_i)'\}'$, 長さ rp の回帰ベクトル $\boldsymbol{u} = (\boldsymbol{u}_1', \cdots, \boldsymbol{u}_p')'$, 誤差分散行列 $R = \sigma^2 I_n$ を用いて次の回帰モデルを考える．

$$\boldsymbol{y} = X\boldsymbol{b} + Z\boldsymbol{u} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, R). \tag{3}$$

次に，散布図 $\{t, y(t)\}$, $t = t_1, \cdots, t_n$ に対する (3) 式の非線形部分の過剰適合を抑えるために回帰係数ベクトル $\boldsymbol{u}$ に対する罰則を加えた残差平方和を考える．

$$(\boldsymbol{y} - X\boldsymbol{b} - Z\boldsymbol{u})'R^{-1}(\boldsymbol{y} - X\boldsymbol{b} - Z\boldsymbol{u}) + \boldsymbol{u}'\Lambda\boldsymbol{u}, \qquad \Lambda = \mathrm{diag}(\lambda_1^2 I_r, \cdots, \lambda_p^2 I_r). \tag{4}$$

ここで，罰則パラメータ $\lambda_1^2, \cdots, \lambda_p^2$ はリッジパラメータと呼ばれ，$\boldsymbol{u}'\Lambda\boldsymbol{u} = \sum_{j=1}^{p} \lambda_j^2 ||\boldsymbol{u}_j||^2$ を満

たすことから，p 個の説明変数 $a_1, \cdots, a_p$ それぞれに対して異なる罰則を考慮している．罰則パラメータは正の値をとるが，大きな正の値をとれば非線形性が抑制され，逆に 0 に近い値をとれば自由度の高い 1 次スプラインによって非線形が表現されるようになる．仮に，(4) 式のリッジパラメータが得られたとすれば既知計画行列 $C = (X, Z)$ を用いて回帰係数ベクトル $\boldsymbol{\theta} = (\boldsymbol{b}', \boldsymbol{u}')'$ の一般化最小二乗推定量は

$$\hat{\boldsymbol{\theta}} = (C'R^{-1}C + B)^{-1}C'R^{-1}\boldsymbol{y}, \qquad B = \mathrm{diag}(0I_{qp}, \Lambda), \tag{5}$$

で与えられる．

3. 混合効果モデルによる推定法

リッジパラメータの選択方法としては，与えられたリッジパラメータの組ごとに (5) 式の推定量を求めて (4) 式の残差平方和を評価することを繰り返し，これを最小とするパラメータを探索することが基本的な考え方である．本稿では，Brumback et al. (1999) に提案された混合効果モデルを利用した方法を紹介する．今，非線形の既知計画行列 Z に対応する回帰ベクトル $\boldsymbol{u}$ を $\boldsymbol{\varepsilon}$ と独立で正規分布 $N(\boldsymbol{0}, G)$ にしたがうランダム効果ベクトルとする．ただし，$G = \mathrm{diag}(\sigma_1^2 I_r, \cdots, \sigma_p^2 I_r)$. このとき，$\mathrm{Var}(\boldsymbol{y}|\boldsymbol{u}) = R$ および $\mathrm{Var}(\boldsymbol{y}) = ZGZ' + R$ が成り立つ．また，ランダム効果ベクトル $\boldsymbol{u}$ と観測値ベクトル $\boldsymbol{y}$ の同時密度関数は次式を満たす．

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{u}) &= f(\boldsymbol{y}|\boldsymbol{u})f(\boldsymbol{u}) \\ &\propto \exp\left\{-(\boldsymbol{y} - C\boldsymbol{\theta})'(\sigma^2 I_n)^{-1}(\boldsymbol{y} - C\boldsymbol{\theta})/2\right\} \exp\left(-\boldsymbol{u}'G^{-1}\boldsymbol{u}/2\right) \\ &\propto \exp\left[-\left\{(\boldsymbol{y} - C\boldsymbol{\theta})'(\boldsymbol{y} - C\boldsymbol{\theta}) + \boldsymbol{u}'(\sigma^2 G^{-1})\boldsymbol{u}\right\}/(2\sigma^2)\right]. \end{aligned} \tag{6}$$

ここで，$\sigma^2 G^{-1} = \mathrm{diag}\{\sigma^2/\sigma_1^2 I_r, \cdots, \sigma^2/\sigma_p^2 I_r\}$ となるので，改めて $\lambda_j^2 = \sigma^2/\sigma_j^2$, $j = 1, \ldots, p$ とおけば，$\Lambda = \sigma^2 G^{-1}$ が成り立つ．したがって，(6) 式を最大化することと (4) 式の罰則付き残差平方和を最小化することが同値であることが分かる．こうして，(6) 式の最尤推定量 $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{b}}', \hat{\boldsymbol{u}}')'$ は (5) 式の一般化最小二乗推定量と一致し，$\boldsymbol{\theta}$ の最良線形不偏推定量であることが示される．また，分散パラメータ $\{\sigma^2, \sigma_1^2, \cdots, \sigma_p^2\}$ は最尤法によって推定可能であることから，結果としてリッジパラメータの最適化と回帰係数ベクトルの推定を同時に行うことができる．

推定された回帰係数ベクトル $\hat{\boldsymbol{\theta}}$ の分散共分散行列は，

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}|\boldsymbol{u}) = (C'R^{-1}C + B)^{-1}C'R^{-1}C(C'R^{-1}C + B)^{-1} = \Omega, \tag{7}$$

とかける（例えば，Lee et al. (2006) の 5 章）．実際に計算するときには，分散パラメータの推定値 $\{\hat{\sigma}^2, \hat{\sigma}_1^2, \cdots, \hat{\sigma}_p^2\}$ から $\hat{R} = R(\hat{\sigma}^2)$ および $\hat{B} = B(\hat{\sigma}_1^2, \cdots, \hat{\sigma}_p^2)$ を求め，R および B を置き換えればよい．それゆえ，変化係数 $\beta_j(t)$, $j \in \{1, \cdots, p\}$ の推定量は $\boldsymbol{c}(t) = \{\boldsymbol{x}(t)', \boldsymbol{z}(t)'\}'$ および $\hat{\boldsymbol{\theta}}_j = (\hat{\boldsymbol{b}}_j', \hat{\boldsymbol{u}}_j')'$ を用いて $\hat{\beta}_j(t) = \boldsymbol{c}(t)'\hat{\boldsymbol{\theta}}_j$ として得られ，その漸近分散は $v_j^2(t) = \boldsymbol{c}(t)'\Omega_j \boldsymbol{c}(t)$ で与えられる．ただし，Ω_j は $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{b}}_1', \cdots, \hat{\boldsymbol{b}}_p', \hat{\boldsymbol{u}}_1', \cdots, \hat{\boldsymbol{u}}_p')'$ に対応して $\Omega = (\Omega_{i,j})_{1 \le i,j \le 2p}$ と分割したとき，

$$\Omega_j = \begin{pmatrix} \Omega_{j,j} & \Omega_{j,p+j} \\ \Omega_{p+j,j} & \Omega_{p+j,p+j} \end{pmatrix}. \tag{8}$$

とかける．ここで，$\mathrm{Var}(\hat{\boldsymbol{b}}_j|\boldsymbol{u}) = \Omega_{j,j}$, $\mathrm{Var}(\hat{\boldsymbol{u}}_j|\boldsymbol{u}) = \Omega_{p+j,p+j}$ および $\mathrm{Cov}(\hat{\boldsymbol{b}}_j, \hat{\boldsymbol{u}}_j|\boldsymbol{u}) = \Omega_{j,p+j}$.

次に，変化係数 $\beta_j(t)$, $j \in \{1, \cdots, p\}$ に関する $100(1-\alpha)\%$ 信頼区間

$$\mathcal{I}_{j,1-\alpha}(t|u_\alpha) = \left[\hat{\beta}_j(t) - u_\alpha v_j(t),\ \hat{\beta}_j(t) + u_\alpha v_j(t)\right] \tag{9}$$

の構築法について考える．ここで，u_α は信頼区間 $\mathcal{I}_{j,1-\alpha}(t|u_\alpha)$ の被覆確率 $\Pr(\beta_j(t) \in \mathcal{I}_{j,1-\alpha}(t|u_\alpha))$ を決める閾値である．$\hat{\boldsymbol{\theta}}$ の漸近正規性から Satoh and Yanagihara (2010) の結果が適用でき，

$$\begin{aligned} \sup_{t\in\mathbb{R}} \left\{ \frac{\hat{\beta}_j(t) - \beta_j(t)}{v_j(t)} \right\}^2 &= \sup_{t\in\mathbb{R}} \frac{\left\{ \boldsymbol{c}(t)' \left(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j \right) \right\}^2}{\boldsymbol{c}(t)'\Omega_j \boldsymbol{c}(t)} \\ &\leq \sup_{\boldsymbol{c}\in\mathbb{R}^{q+r}} \frac{\left\{ \boldsymbol{c}' \left(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j \right) \right\}^2}{\boldsymbol{c}'\Omega_j \boldsymbol{c}} = \left(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j \right)' \Omega_j^{-1} \left(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j \right) \rightsquigarrow \chi^2_{q+r} \end{aligned} \tag{10}$$

が成り立つ．したがって，自由度 $q+r$ の χ^2 分布の上側 $100\alpha\%$ 点 $c_{q+r,\alpha}$ を用いて，$u_\alpha = \sqrt{c_{q,\alpha}}$ とすることで，$t \in \mathbb{R}$ における被覆確率が $\Pr\left(\beta_j(t) \in \mathcal{I}_{j,1-\alpha}(t|\sqrt{c_{q+r,1-\alpha}})\right) \geq 1-\alpha$ を満たす同時信頼領域が構築できる．

4. 適 用 例

ここでは 2 章で提案したセミパラメトリックな変化係数を 3 章で紹介した推定法を用いて 3 つの実データに適用する．1 つ目の骨塩量の相対変化データは非線形曲線の例であり性差を変化係数で記述する．2 つ目のテオフィリンの血中濃度データの経口投与量は連続型の共変量であり，その時間とともに変化する効果に関心がある．そして，最後に CD4 陽性リンパ球細胞数データの観測値は計数値なので一般化線形混合モデルの枠組で推定を行う．また，その結果得られる変化係数は直線で近似できる例となっている．なお，このデータは (1) 式で仮定したように同一個体であっても観測時点によって共変量である喫煙数が変化する．

共通の設定としては時間に関して直線をあらわす長さ $q = 2$ の基底関数ベクトル $\boldsymbol{x}(t_i) = (1, t_i)'$ を用いた．また，非線形曲線の節点として測定時点の $\{10, \cdots, 90\}\%$分位点を使い，長さ $r = 9$ の基底関数ベクトル $\boldsymbol{z}(t_i) = \{(t_i - \kappa_1)_+, \cdots, (t_i - \kappa_r)_+\}$ を構成した．ここで，節点で区切られた区間の中にほぼ同数の標本数が含まれることに注意する．また，分散パラメータの推定については最尤法ではなく，混合効果モデルでよく用いられる制限付最尤推定量(例えば，Harville (1977))の結果を示した．

4.1. 骨塩量の相対変化データ

北米に住む若者 261 人(男性 116 人，女性 145 人)の脊柱における骨塩量の相対変化を観測値として用いる．骨塩量とは一定量の骨の中に含まれるミネラル分の量を示す指標であり，骨粗鬆症の診断に用いられる．観測値は連続した 2 回の測定時の骨塩量の差をその平均で割った相対変化を示しており，測定時年齢の平均年齢を測定時点としている．測定時点数は 1 回のみが 107 人，2 回が 84 人そして 3 回が 70 人，したがって，観測値数は合計 485 であった．なお，このデータは Hastie et al. (2009) の 5 章において平滑化の適用例として挙げられている．

観測値を y_i, $i = 1, \cdots, 485$, これを説明する $p = 2$ 個の共変量として，a_{i1} はすべて 1, a_{i2} は

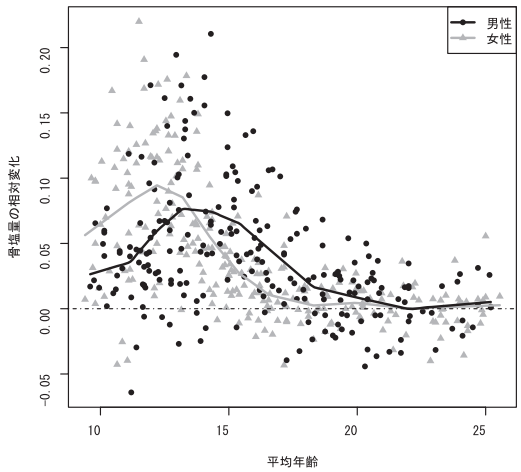

図 1. 骨塩量の相対変化の推定曲線

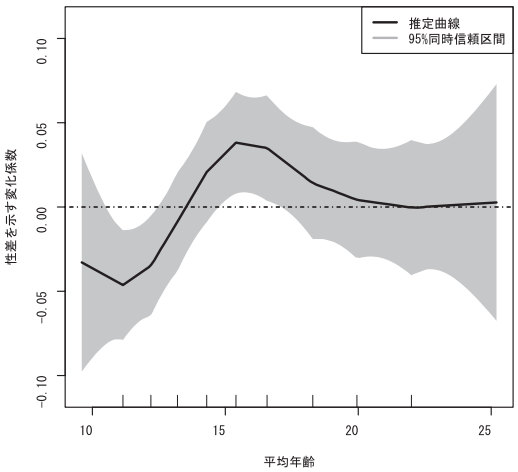

図 2. 女性に対する男性の効果を示す変化係数

男性なら 1, 女性なら 0 をとるダミー変数，したがって $\boldsymbol{a}_i = (a_{i1}, a_{i2})'$, 観測時点として平均年齢 t_i を用いた．

推定の結果，分散パラメータは $(\hat{\sigma}^2, \hat{\sigma}_1^2, \hat{\sigma}_2^2) = (1.45 \times 10^{-3}, 2.06 \times 10^{-4}, 2.68 \times 10^{-4})$, よってリッジパラメータとしては $\hat{\lambda}_1^2 = 7.07$, $\hat{\lambda}_2^2 = 5.43$ と推定された．図 1 に女性および男性に対する推定曲線 $\hat{\beta}_1(t)$ および $\hat{\beta}_1(t) + \hat{\beta}_2(t)$ を，図 2 に共変量 a_2 の時間とともに変化する効果 $\hat{\beta}_2(t)$ を示す．ここで，X 軸上の縦線は節点の位置を示す．以下に，統計解析ソフトウェア R (R Core Team, 2012) において実行可能な推定曲線を描くスクリプトを記述する．

```
library(ElemStatLearn); data(bone) # データの初期設定
n <- nrow(bone); t <- bone$age; y <- bone$spnbmd
a <- 1*(bone$gender=="male"); group <- rep(1,length(y))
X.1 <- cbind(1,t); X.2 <- a*X.1;
X <- cbind(X.1,X.2); colnames(X) <- 1:ncol(X)
knots <- quantile(t,(1:9)/10); q <- length(knots)
Z.1 <- matrix(0,nrow=n,ncol=q)
for(j in 1:q) Z.1[,j] <- (t-knots[j])*(t-knots[j]>0)
Z.2 <- a*Z.1; Z <- cbind(Z.1,Z.2)
library(nlme) # 混合効果モデルによる推定
d <- groupedData(y~-1+X|group, data=data.frame(y,X))
pdIdents <- list(pdIdent(~Z.1-1),pdIdent(~Z.2-1))
res <- lme(y~-1+X,data=d,random=list(group=pdBlocked(pdIdents)))
var.hat <- unique(as.numeric(VarCorr(res)[,"Variance"]))
names(var.hat) <- c("var.hat.1","var.hat.2","var.hat")
b.hat <- as.matrix(res$coef$fixed)
u.hat <- as.matrix(unlist(res$coef$random))
y.hat <- X %*% b.hat + Z %*% u.hat
plot(t,y.hat,col=2*(a+1)) # モデルによる理論値
beta.1.hat <- X.1 %*% b.hat[1:2,] + Z.1 %*% u.hat[1:q,]
plot(t, beta.1.hat) # 女性(定数)の効果を示す変化係数
beta.2.hat <- X.2 %*% b.hat[2+1:2,] + Z.2 %*% u.hat[q+1:q,]
plot(t[a==1], beta.2.hat[a==1,]) # 男性の効果を示す変化係数
```

4.2. テオフィリンの血中濃度データ

抗喘息薬テオフィリンを経口投与された 12 人の血中濃度を観測値として用いる．観測回数は

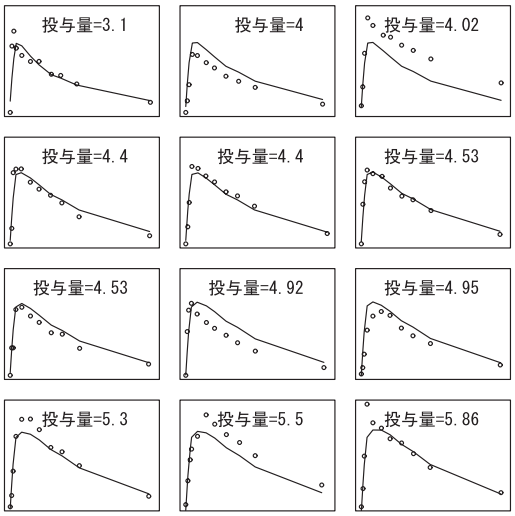

図 3. 個体別血中テオフィリン濃度の時間変化と推定曲線

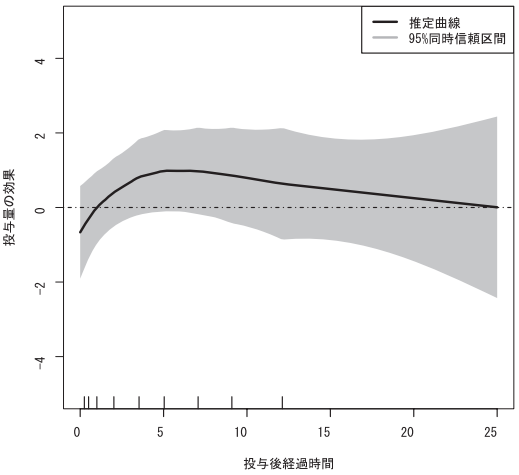

図 4. 経口投与量の効果を示す変化係数

すべての個体で 11 回と揃っているが，観測時点である投与後の経過時間 [h] は個体ごとに異なっており最小値 0.00, 最大値 24.65 であった．投与量 [mg/kg] は 320 [mg] を体重 [kg] で割った値でほぼ算出でき平均 4.63, 最小値 3.10, 最大値 5.86 であった．なお，データは統計ソフト R に Theoph として格納されている．

観測値を y_i, $i = 1, \cdots, 132$, これを説明する $p = 2$ 個の共変量として，a_{i1} はすべて 1, a_{i2} は投与量，したがって $\boldsymbol{a}_i = (a_{i1}, a_{i2})'$, 観測時点として投与後経過時間 t_i を用いた．

推定の結果，分散パラメータは $(\hat{\sigma}^2, \hat{\sigma}_1^2, \hat{\sigma}_2^2) = (1.75, 4.27, 4.11 \times 10^{-2})$, よってリッジパラメータとしては $\hat{\lambda}_1^2 = 0.409$, $\hat{\lambda}_2^2 = 48.3$ と推定されていた．図 3 に個体ごとの観測値と推定曲線を示す．横軸は経過時間を，縦軸は血中濃度をあらわす．なお，個体ごとの図の横軸および縦軸の範囲は共通とした．図 4 に共変量 a_2 の時間とともに変化する効果 $\hat{\beta}_2(t)$ を示す．ここで，X 軸上の縦線は節点の位置を示す．

4.3. CD4 陽性リンパ球細胞数データ

ヒト免疫不全ウイルス(HIV)に感染した 369 人の CD4 陽性リンパ球細胞数データを観測値として用いる．HIV 感染は後天性免疫不全症候群(AIDS)の原因であり，感染が進行すると CD4 陽性リンパ球細胞数が減少する．HIV 抗体を検出した時点を 0 とし，負の値を含む経過年数を観測時点として扱う．観測時点数が 1 回から 12 回までの人数はぞれぞれ $\{5, 24, 25, 47, 43, 52, 40, 41, 38, 21, 23, 10\}$ であった．背景要因として 1 日あたりの喫煙数(0 箱から 4 箱)が記録されていた．なお，このデータは Kaslow et al. (1987) に報告された調査の一部である．

観測値を y_i, $i = 1, \cdots, 2376$, これを説明する $p = 3$ 個の共変量として，a_{i1} はすべて 1, a_{i2} および a_{i3} はそれぞれ，1 日あたりの喫煙数が $\{1, 2\}$ 箱および $\{3, 4\}$ 箱に対応し，該当すれば 1, そうでなければ 0 をとるダミー変数，したがって $\boldsymbol{a}_i = (a_{i1}, a_{i2}, a_{i3})'$, 観測時点として HIV 抗体検出からの経過年数 t_i を用いた．

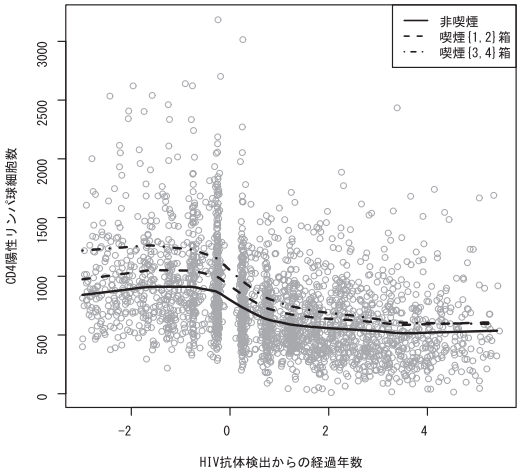

図 5. CD4 陽性リンパ球細胞数と推定曲線

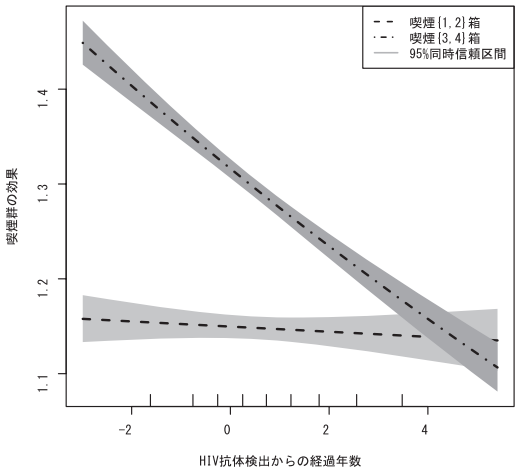

図 6. 非喫煙群に対する比を示す変化係数

観測値は計数値であるためポアソン分布にしたがうと考えられる．そこで，観測値と独立なランダム効果 $\boldsymbol{u} \sim N(\mathbf{0}, G)$ を用いて一般化線形混合モデル $\boldsymbol{y}|\boldsymbol{u} \sim \text{Poisson}\{\exp(X\boldsymbol{b}+Z\boldsymbol{u})\}$ を考える．Breslow and Clayton (1993) は条件付期待値 $\boldsymbol{\mu} = E(\boldsymbol{y}|\boldsymbol{u})$ および条件付分散 $W = \text{var}(\boldsymbol{y}|\boldsymbol{u})$ を持つ一般化線形混合モデルに対して，疑似観測値ベクトル $\boldsymbol{y}_{pseudo} = X\boldsymbol{b}+Z\boldsymbol{u}+W^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$ を考え，混合効果モデルに帰着することで近似的に $\boldsymbol{\theta}$ を推定する罰則付疑似尤度法を提案している．ここで，$\text{var}\{W^{-1}(\boldsymbol{y}-\boldsymbol{\mu})|\boldsymbol{u}\} = W^{-1}$ が成り立つので，(3) 式において $R = W^{-1}$ とおけば (7) 式から $\hat{\boldsymbol{\theta}}$ の条件付分散は

$$\text{Var}(\hat{\boldsymbol{\theta}}|\boldsymbol{u}) = (C'WC+B)^{-1}C'WC(C'WC+B)^{-1}, \tag{11}$$

として与えられる．罰則付疑似尤度法の詳細については Ruppert et al. (2003) の 10 章を参照して頂きたい．特に，観測値がポアソン分布に従うならば，$\boldsymbol{\mu} = \exp(X\boldsymbol{b}+Z\boldsymbol{u})$ および $W = \text{diag}(\boldsymbol{\mu})$ である．

推定の結果，分散パラメータは $(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\sigma}_3^2) = (1.87\times10^{-2}, 2.12\times10^{-11}, 2.19\times10^{-11})$ と推定された．$\hat{\sigma}_2^2$ および $\hat{\sigma}_3^2$ は $\hat{\sigma}_1^2$ と比較して非常に小さい値として推定されており，対応する変化係数がほぼ直線，$\hat{\beta}_2(t) = 0.139 - 2.36\times10^{-3}t$, $\hat{\beta}_3(t) = 0.275 - 3.21\times10^{-2}t$ で近似されることが分かる．これが検定によって検証されるのであれば，帰無仮説として「非線形部分の回帰係数が 0 である」を考えればよい．(8) および (10) 式より，帰無仮説 $\boldsymbol{u}_j = \mathbf{0}$ に対して Wald 型検定統計量 $\mathcal{T}_j = \hat{\boldsymbol{u}}_j\Omega^{-1}_{p+j,p+j}\hat{\boldsymbol{u}}'_j$, $j = 2, 3$ を考えると，帰無仮説の下で漸近的に自由度 $q = 9$ の χ^2 分布にしたがう．実際，$\mathcal{T}_2 = 0.068$ および $\mathcal{T}_3 = 0.077$ となり $c_{q,0.05} = 16.9$ より小さく，帰無仮説は棄却されない．図 5 に観測値と喫煙量ごとの推定曲線 $e^{\hat{\beta}_1(t)}$, $e^{\hat{\beta}_1(t)+\hat{\beta}_2(t)}$ および $e^{\hat{\beta}_1(t)+\hat{\beta}_3(t)}$ を示す．図 6 に共変量 a_2 の時間とともに変化する効果 $e^{\hat{\beta}_2(t)}$ および $e^{\hat{\beta}_3(t)}$ を示す．ここで，X 軸上の縦線は節点の位置を示す．なお，推定には統計解析ソフトウェア R の MASS ライブラリにある罰則付疑似尤度法を実装した glmmPQL 関数を利用した．

5. おわりに

線形性を仮定した変化係数は時間と共変量の交互作用項として様々な回帰モデルへ拡張できるため応用範囲が広い．冨田・佐藤・大谷ら (2010, 2012) および Tonda et al. (2012) ではコックスの比例ハザードモデルを用いて広島原爆被爆者の被爆時所在地によって変化する死亡リスクを推定し，従来爆心地からほぼ同心円上に減少すると考えられていた死亡リスクが方角によっては非対称であることを示した．また，冨田ら (2011) では時系列解析に応用し，加茂・冨田・佐藤 (2011) ではがん統計データに対してポアソン回帰モデルを用いて年齢–時代空間上のがん死亡リスクの視覚化を試みている．

本稿で提案したセミパラメトリックな変化係数は新たに非線形構造が記述できるだけでなく，線形性で十分であるという結果を導くこともできるため探索的な解析に有用である．補われた非線形部分に関する最適化は Brumback et al. (1999) の提案手法により混合効果モデルを利用することで推定と同時に実行される．さらに，同時信頼区間の構成についても線形な場合の変化係数における理論的な結果がほぼ同様に適用でき，非線形を考慮したにも関わらず線形と同様の使い易さを保つことができた．一方で，Brumback et al. (1999) の推定法は非線形性を抑制するためのリッジパラメータをランダム効果の分散に置き換えて最適化するため，利用できるモデルが混合効果モデルおよび一般化線形混合モデルなどに限定されてしまう．また，$\mathrm{Var}(\boldsymbol{y}) = ZGZ' + R$ に見られるようにすべての観測値に相関構造が仮定されてしまうため，通常の経時測定データのように個体ごとの独立性を保っていないという問題もある．変化係数は同一個体において繰り返し観測されるデータの解釈に有用なので，個体間および個体内の相関構造を自由に規定できる手法が望まれる．

謝　辞　本論文の修正にあたり，建設的なご意見と有益なコメントを頂いた 2 人の査読者および編集委員の方々に深く感謝いたします．この研究の一部は，文部科学省科学研究費の若手研究 (B) 課題番号 23790694 および 23700337，統計数理研究所共同研究プログラム (24-共研-4104)，広島大学原爆放射線医科学研究所共同研究重点プログラムの援助を受けています．

参考文献

Breslow, N. E. and Clayton, D. G. (1993): Approximate inference in generalized linear mixed models, *J. Amer. Statist. Assoc.* **88**, 9–25.

Brumback, B. A., Ruppert, D. and Wand, M. P. (1999): Variable selection and function estimation in additive nonparametric regression using a data-based prior: Comment, *J. Amer. Statist. Assoc.* **94**, 794–797.

David, A. and Harville, D. A. (1977): Maximum likelihood approaches to variance component estimation and to related problems, *J. Amer. Statist. Assoc.* **72**, 320–338.

Hastie, T., Tibshirani, R. and Friedman, J. (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), Springer-Verlag.

Hastie, T. and Tibshirani, R. (1993): Varying-coefficient models, *J. Roy. Statist. Soc. Ser.* 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.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F. and Rinaldo, C. R. Jr. (1987): The multicenter AIDS cohort study: rationale, organization, and selected characteristics of the participants, *American Journal of Epidemiology* **126**, 310–318.

加茂憲一，冨田哲治，佐藤健一 (2011): 年齢–時代平面上における癌死亡リスクの視覚化，統計数理 59(2), 217–237.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006): *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood* (2nd ed.), Chapman & Hall/CRC.

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

R Core Team (2012): *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

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

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

佐藤健一，柳原宏和，加茂憲一 (2009): 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，応用統計学 38, 1–11.

Satoh, K. and Yanagihara, H. (2010): Estimation of varying coefficients for a growth curve model, *Amer. J. Math. Management Sci.* **30**, 243–256.

Tonda, T., Satoh, K., Nakayama, T., Katanoda, K., Sobue, T. and Ohtaki, M. (2011): A nonparametric mixed-effects model for cancer mortality, *Aust. N. Z. J. Stat.* **53**, 247–256.

冨田哲治，佐藤健一，大谷敬子，佐藤裕哉，丸山博文，川上秀史，星正治，大瀧慈 (2010): 広島原爆被爆者コホートにおける被爆時所在地に基づく死亡危険度地図作成の試み，長崎医学会誌 85, 185–188.

冨田哲治，佐藤健一，大谷敬子，佐藤裕哉，丸山博文，川上秀史，田代聡，星正治，大瀧慈 (2012): 広島原爆被爆者コホート 1970～2010 年におけるリスク地図の推定，広島医学 65(4), 255–258.

T. Tonda, K. Satoh, K. Otani, Y. Sato, H. Maruyama, H. Kawakami, S. Tashiro, M. Hoshi and M. Ohtaki (2012): Investigation on circular asymmetry of geographical distribution in cancer mortality of Hiroshima atomic bomb survivors based on risk maps: analysis of spatial survival data, *Radiation and Environmental Biophysics* **51**(2), 133–141.

冨田哲治，佐藤健一，中山晃志，片野田耕太，祖父江友孝，大瀧 慈 (2011): 変化係数を用いたがん死亡危険度の年次変動要因の推測，統計数理 59(2), 205–215.

冨田哲治，佐藤健一，柳原宏和 (2010): 空間データに対する交互作用モデルを用いた変化係数曲面の推測について，応用統計学 39, 59–70.

（2012 年 10 月 29 日受付　2013 年 2 月 5 日最終修正　2 月 6 日採択）

著者連絡先：〒 734–8551 広島市南区霞 1–2–3
佐藤 健一（Tel. 082–257–5857）
E-mail: ksatoh@hiroshima-u.ac.jp

Japanese J. Appl. Statist. 42 (1) (2013), 1–10

Statistical Inference of Semiparametric Varying Coefficients Using Mixed Effects Model

Kenichi Satoh[1,*] **and Tetsuji Tonda**[2]

[1] Research Institute for Radiation Biology and Medicine, Hiroshima University
[2] Faculty of Management and Information Systems, Prefectural University of Hiroshima

Abstract

Varying coefficients can be used for visualizations or interpretations of the covariate effects which might be varying on time axis. Satoh and Yanagihara (2010) proposed linear varying coefficients and constructed a simultaneous confidence interval as a function of time. Linear curves are useful to summarize and interpret the scatter plot, but these might be not enough for approximating non-linear relationship especially when there are many observed time points. In this paper we consider semiparametric varying coefficients with splines and estimate them using a framework of linear mixed effect model which was proposed by Brumback et al. (1999), and then we construct a simultaneous confidence interval.

Key words: longitudinal data, varying coefficient, semiparametric, mixed effects model

*Corresponding author
E-mail address: ksatoh@hiroshima-u.ac.jp (Kenichi Satoh)

Received October 29, 2012; Received in final form February 5, 2013; Accepted February 6, 2013.