ノート

学習アルゴリズムによる整合的多重脱退モデル推定

尾上辰徳 *　　松山直樹 †

2017 年 1 月 31 日投稿
2017 年 2 月 14 日受理

概要

本研究は，保険数理で用いられる多重脱退モデル推定の基本問題を解消するための新たな手法（交互イタレーション）を提案するものである．交互イタレーションは公的統計による推移確率の不完全観測から出発して状態間制約と年齢間制約を充足する整合的多重状態モデルを推定する学習アルゴリズムであり，尤度の意味での収束性を示すことができる．また，粗い年齢群団数値しか得られない公的統計データを元データの再現性を確保しつつ各歳別に展開する手法（積分補間）も用いる．具体的に，がん疾病の多重脱退モデル推定に本手法を適用し，それらの有用性を示す．

キーワード：がん多重脱退，定常社会モデル，交互イタレーション，欠測値，公的統計

1. 多重脱退モデル推定問題

(1) 観測推移確率と多重脱退モデルの不整合

生命保険数理は，長期の保険期間に対応するために，状態推移確率が状態と年齢のみによって決まる定常社会モデルに基づき構成されているが，現実において定常社会の観測は不可能である．死亡のような一状態モデルでは現実の状態推移確率の観測値を定常社会モデルの状態推移確率とみなしたモデル構築が可能であるが，医療保険分野の多重脱退モデルにおいては根本的な不整合が存在する．すなわち，仮に状態推移確率の完全観測が可能だとしても，状態別の推移確率と全状態を統合した推移確率の関係は現実の非定常な状態別人口構成の影響を受けるため，定常社会モデルにおける状態別人口構成における関係と不整合となる．たとえば，ℓ_x, ℓ_x^j はそれぞれ定常社会モデルの年齢別人口，状態 j における年齢別人口を表し，$\widehat{\ell_x}, \widehat{q_x}, \widehat{\ell_x^j}, \widehat{q_x^j}$ は現実の人口，死亡率，状態 j における人口，死亡率の観測値とすると，現実の非定常観測では，状態別の死亡数合計は全死亡数に一致するため $\widehat{\ell_x}\widehat{q_x} = \sum_j \widehat{\ell_x^j}\widehat{q_x^j}$ が成り立つが，定常社会モデルの人口構成では $\ell_x\widehat{q_x} \neq \sum_j \ell_x^j \widehat{q_x^j}$ となり不整合を生ずる．多重脱退モデルは，年齢間制約条件（年齢による漸化式が成り立つ）と状態間制約条件（ある年齢の状態別人口合計は全人口に一致する加法性が成り立つ）という二重の制約条件を満たすものでなければならないが，観測値を直接用いる限り，いずれかの制約条件は無視せざるを得ない．例えば，先行研究 (友寄 [2015]) では年齢間制約条件は無視されている．

(2) 公的統計に依存する制約

医療保険分野の多重脱退モデル推定では状態数の増加に伴い状態別の標本数が少なくなるため，一般的には観測数が大きい国レベルの公的統計に依存せざるを得ない．公的統計の制約として，(ⅰ) 必

*明治大学大学院先端数理科学研究科
†明治大学総合数理学部　〒 164-8525 中野区中野 4-21-1　ma2yama(at)meiji.ac.jp
本稿作成に当たり，実務的アドバイスを頂戴した住友生命の金村慶二氏と有益なコメントを頂戴した匿名査読者に感謝する．

要な観測値が得られないこと（欠測値の存在），(ii) データ観測区分の粒度の荒さ，等が挙げられる．例えば，疾病からの回復の観測困難性や，伝統的な脱退残存モデルの有病率と回復の影響を受ける観測有病率の不整合や，5 歳区分ごとにしか観測値が得られないことなどである．

先行研究（友寄 [2015]）や標準的な実務 (山内, 金村, 富島 [2012]) においては，これらの問題は特に議論の対象とはされてこなかった．

一般に観測値 x からパラメータ θ を決定する問題は尤度関数 $p(x;\theta)$ の最大化問題として解かれるが，ここで扱う問題では尤度関数を書き下すことが困難であるために新たな手法が必要となる．

2. 交互イタレーションの導入

上記の問題点のうち (1) と (2)-(i) に対処するために，一般化した問題の定式化を行う．不完全観測をもとにした推定を行う必要があるために，変数を以下で定める．

X：観測値, Z：潜在変数,θ: モデルパラメータ (年齢 $\times$ 状態数のベクトル)

X は定常社会モデルに組み込み可能な観測データを用いることにし，Z は観測困難なものに加え，観測がなされていても定常社会モデルの前提と相容れないものも含めて考える．

ちなみに，後述のがん多重脱退の実装例では，

X(観測値) : $\{q_x\}_{x=0}^{\omega}$(死亡率), $\{\delta_x\}_{x=0}^{\omega}$(死亡割合), $\{\lambda_x^*\}_{x=0}^{\omega}$(絶対罹患率)

Z(潜在変数) : $\{h_x\}_{x=0}^{\omega}$(非がん患者数), $\{k_x\}_{x=0}^{\omega}$(がん患者数)

θ(パラメータ) : $\{q_x^N\}_{x=0}^{\omega}$(非がん集団の死亡率), $\{q_x^{c1}\}_{x=0}^{\omega}$(がん集団のがん以外の原因による死亡率), $\{q_x^{c2}\}_{x=0}^{\omega}$(がん集団のがんが直接的な原因による死亡率), $\{\lambda_x\}_{x=0}^{\omega}$(罹患率)

が対応する．

ここで，完全観測 (X, Z) はモデルパラメータ θ を決定し，モデルパラメータ θ が得られれば潜在変数 Z が決定できることから状態間，年齢間制約条件は次式のように記述できる．

$$\text{状態間制約条件：}\theta = g(X, Z) \tag{1}$$

$$\text{年齢間制約条件：}Z = h(\theta) \tag{2}$$

g, h はそれぞれ X と Z を変数に持つベクトル値関数であり問題ごとに異なる．また，この 2 つの制約条件は多次元の非線形連立方程式となるために代数的に解を得ることが出来ない．そこで，漸化式的構造を導入し繰り返し計算を用いて解を求めることを考える．イタレーション回数を t で表し，欠測値を $Z(1)$ とする．(1)(2) 式を，

$$\theta(t+1) = g(X, Z(t+1)) \tag{3}$$

$$Z(t+1) = h(\theta(t)) \tag{4}$$

と表すと，$\theta(t+1)$ は上記 2 式より，

$$\theta(t+1) = g\left(X,\, h(\theta(t))\right) \tag{5}$$

と漸化式構造にまとめられる．$\theta(t)$ の極限値 α が存在するとすると，極限値 α は (1)(2) 式を満たす解となる．極限値を持つ時には次の条件式

$$^{\forall}\varepsilon > 0\,,\, ^{\exists}T\,,\, \text{s.t. } ||\theta(t+1) - \theta(t)|| \leq \varepsilon\,,\, \text{for } T \leq t \tag{6}$$

とともにイタレーション・アルゴリズムにより (1)(2) を解くことができるが，構造が複雑なために漸化式 (5) の収束条件の確認は困難である．そのため，確率的構造を導入し学習アルゴリズムとして定式化し，尤度の上昇という観点から収束性を示す．
観測変数の分布を X，潜在変数の分布を Z とする．
状態間制約条件 (1)，年齢間制約条件 (2) は推定値の最尤性として以下のように書き換えられる．

$$\text{状態間制約条件} \quad : \quad p(x,\, z\,;\, g(x,\, z)) \geq p(x,\, z\,;\, \theta) \tag{7}$$

$$\text{年齢間制約条件} \quad : \quad p(h(\theta)\,;\, \theta) \geq p(z\,;\, \theta) \tag{8}$$

このとき，尤度 $p(x;\theta)$ を具体的に書き下すことは困難であるが次の定理が成り立つ．

定理

制約条件 (7)(8) のもとで，漸化式：$\theta(t+1) = g(x, h(\theta(t)))$ で定まる $\{\theta(t)\}$ は，

$$p(x;\theta(t+1)) \geq p(x;\theta(t))\,(t = 1, 2, \cdots)$$

を満たす．

[証明]
観測データ x の対数尤度関数 $\ell(\theta,\, x)$ は，任意の確率分布 $z \sim Z$ で周辺化すると，

$$\ell(\theta,\, x) = \log p(x\,;\, \theta) = \log p(x,\, z\,;\, \theta) - \log p(z|x\,;\, \theta)$$

と表せる．$\ell(\theta, x)$ と $\ell(\theta(t), x)$ の尤度差は，

$$\log p(x;\theta) - \log p(x;\theta(t)) = \log p(x,\, z;\theta) - \log p(z|x;\theta) - \log p(x,\, z;\theta(t)) + \log p(z|x;\theta(t)) \tag{9}$$

となる．ここで，Z は任意なので $z \sim Z|x;\theta(t)$ として，$Z|x;\theta(t)$ は年齢間制約条件を満たす次の一点分布とする．

$$\begin{cases} p(z = h(\theta(t))) = 1 \\ p(z \neq h(\theta(t))) = 0 \end{cases}$$

上式 (9) の第四項は，$\log p(z|x;\theta(t)) = 0$ となり，上式 (9) は，

$$\begin{aligned} \text{上式 (9)} &= \log p(x,z;\theta) - \log p(x,z;\theta(t)) - \log p(z|x,\theta) \\ &\geq \log p(x,h(\theta(t));\theta) - \log p(x,h(\theta(t));\theta(t)) \end{aligned} \tag{10}$$

となる．(10) の第一項で，$\theta = \theta(t)$ ととると 0 となるので，

$$\theta = \text{argmax}_\theta \log p(x,h(\theta(t));\theta) \tag{11}$$

とおくと，尤度差 (9) は非負となる．
ここで，(7) より (11) を満たす θ は $\theta = g(x,\, h(\theta(t)))$ となる．
よって，$\theta(t+1) = g(x,\, h(\theta(t)))$ とおくと，

$$\log p(x;\theta(t+1)) - \log p(x;\theta(t)) \geq 0 \tag{12}$$

が得られる． □

この定理から尤度は t に関して単調非減少で上に有界であるので極限が存在し尤度の意味で $\theta(t)$ は収束することがわかる．

この手法は，2 つの制約アルゴリズムを交互に用いて推定値を更新することから以下では交互イタレーションと呼称することにする．

交互イタレーションのアルゴリズムをフローチャートで書くと図 1 の通りである．

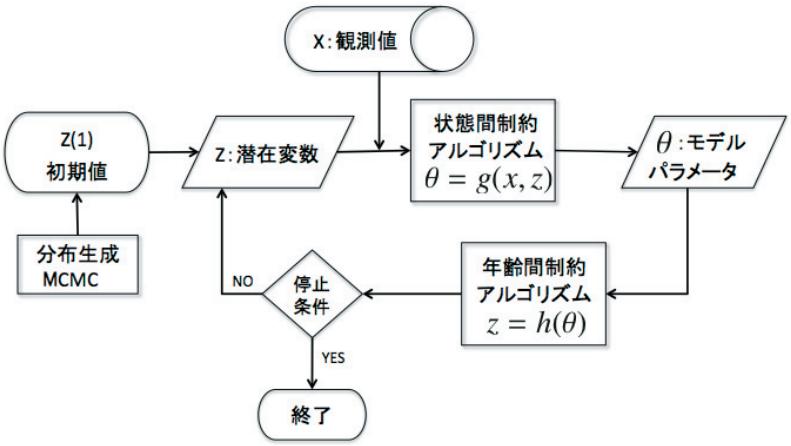

図 1: 交互イタレーションのアルゴリズム

一般には収束点が複数存在する可能性がある．そのような場合には潜在変数の初期値をランダムに生成して収束値の選択を行う必要がある．ただし，多次元における「次元の呪い」を回避するために MCMC 法等による分布生成を行う必要がある．

3. 積分補間の導入

上記の問題点のうち (2)-(ii) に関しては，交互イタレーションは各歳化された観測値に依存するため，各歳化にあたってのデータの滑らかさと観測値の再現性には特別な配慮が必要となる．一般的に公的統計は 5 歳年齢群での開示にとどまっており各歳別の多重脱退を表現するのに必要な観測値が得られない．仮にこれらのデータに対してある年齢区分で与えられた観測値を区分の中央年齢に当てはめて直線補間するとデータの再現性と滑らかさは失われる．各年齢での数値を求める上で，

1. 各年齢における数値の年齢区分における合計は元データと一致すること．
2. 年齢区分による境界点において連続であること．

という二つの制約を満たす手法として以下の計算ステップを導入し，以下では積分補間と呼ぶことにする．

前提として公的統計から区分 $i = 1, 2, \cdots, n$ に対するデータ $\{(x_i, y_i)\}_{i=1}^{n}$ が得られているとする．x_i は年齢区分を表す．

(ステップ 1) 観測値から累積度数 $Y_i = \sum_{j=1}^{i} y_j$ を求め，データ $\{(x_i, Y_i)\}_{i=1}^{n}$ を得る．

(ステップ 2)Y_i の傾きを見ながら連続した年齢区分を選択し滑らかな累積度数関数 $F(x)$ を多項式近似により推定する．区分を m 個選択した場合 $F(x)$ は

$$F(x) = \sum_{i=1}^{m} a_i x^{m-i}$$

となり，選択した (x_i, Y_i) の点を全て通過する条件から $(m-1)$ 次連立方程式により係数 $a_i (i = 1, \cdots, m)$ は求まる．

(ステップ 3) 年齢区分の選択により複数個の $F(x)$ が求まるが重複部分に関しては和半等により調整を行い，0 歳とデータにおける最終年齢に関しては定数捕外による調整を行う．

(ステップ 4) 各歳の値 $f(x)$ は

$$f(x) = F(x+1) - F(x) \quad (x = 0, 1, \cdots \omega)$$

により求める．

ただし，$f(x)$ が負値とならないようにステップ 3 において調整を行う．

4．がん多重脱退モデル推定への実装

この節では，多重脱退モデル推定の問題点 (1)(2) に対処する方法として提案した交互イタレーションと積分補間を具体的にがん疾病を対象とした回復を見込まない多重脱退モデルの推定に適用する．回復は一般には定義・観測が困難であるということに加え，再発の可能性もあるため，がん集団をがん罹患経験者集団と位置付ける立場をとるためである．また，この立場は後述する地域がん登録の考え方と整合的である．

4.1 使用する公的統計データ

がんの多重脱退モデルの推定においては厚生労働省の患者調査の有病率を用いることが最も簡単

であるが，患者調査の有病率の定義は回復を含めないモデルの前提と整合しないため，有病率を未知の潜在変数としてがん罹患率とがん死亡割合を観測値とする推定を行う．がんの罹患率の推計については患者調査における患者数の年齢間差分を使う方法 (友寄 [2015]) と，地域がん登録の罹患率を直接使う方法 (山内, 金村, 富島 [2012]) 等が知られている．前者は定常社会モデルと明らかに不整合である．後者は特定地域の観測に基づくものであるが全国規模に換算するための統計処理がなされていることから，ここでは地域がん登録の罹患率を観測値として採用する．

・厚生労働省　平成 23 年度　人口動態調査　上巻　死亡　第 5.15 表人口動態
「性・年齢別にみた死因年次推移分類別死亡数及び率（人口 10 万対)」
・国立がん研究センターがん対策情報センター
「地域がん登録全国推計によるがん罹患データ（1975 年 ~2012 年)」
・厚生労働省　平成 23 年度　簡易生命表（男）第 1 表
各基礎率を求めるための人口として総務省の総人口を用いた．
これらのデータは 3 . で述べた積分補間を用いて年齢別に展開したうえで使用する．

4.2 モデルの構造

図 2 は脱退残存モデル（回復なし）のイメージ図である．状態は非がん集団とがん集団，死亡の 3 状態とし，回復を見込まないので，がん集団の脱退は死亡のみである．

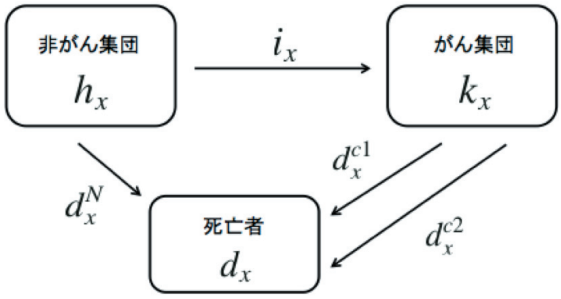

図 2: 脱退残存モデルのイメージ図

図 2 中で用いた記号をまとめる．
h_x：非がん集団の生存者数
k_x：がん集団の生存者数
$d_x^N (= h_x q_x^N)$：非がん集団における死亡者数
$d_x^{c1} (= k_x q_x^{c1})$：がん集団のがん以外の原因による死亡者数 (ただし, $q_x^{c1} = (1+m) q_x^N$ の仮定を置く)
$d_x^{c2} (= k_x q_x^{c2})$ ：がん集団のがんが直接的な原因による死亡者数
$i_x (= h_x \lambda_x)$：罹患者数（λ_x は罹患率を表す）

実装における X，Z，θ は，

$$\begin{aligned} X(\text{観測値}) &: \{q_x\}_{x=0}^{\omega}(\text{死亡率}), \{\delta_x\}_{x=0}^{\omega}(\text{死亡割合}), \{\lambda_x^*\}_{x=0}^{\omega}(\text{絶対罹患率}) \\ Z(\text{潜在変数}) &: \{h_x\}_{x=0}^{\omega}, \{k_x\}_{x=0}^{\omega} \\ \theta(\text{パラメータ}) &: \{q_x^N\}_{x=0}^{\omega}, \{q_x^{c1}\}_{x=0}^{\omega}, \{q_x^{c2}\}_{x=0}^{\omega}, \{\lambda_x\}_{x=0}^{\omega} \end{aligned}$$

である．地域がん登録より得られる罹患率は近似的に絶対罹患率 λ_x^* が得られているものとしている．患者調査有病率は入院・退院ベースによる患者数なのに対して，今回の多重脱退モデル及び地域がん登録による（絶対）罹患率 λ_x^* は診断ベースによる患者数となっているため，患者調査有病率は定常社会モデルと明らかに不整合である．従って，実装において患者調査有病率を観測値と見なした h_x, k_x の導出は行っていない．また，非がん患者のがん以外の死亡指数 m に関する仮定 $q_x^{c1} = (1+m)q_x^N$ は先行研究に準じて $m = 0$ とした．

次に，ベクトル値関数 g, h のアルゴリズムを述べる．

X，Z，θ を，

$X_x = (q_x, \delta_x, \lambda_x^*)$，$X = \{X_x\}_{x=0}^{\omega}$

$Z_x = (h_x, k_x)$，$Z = \{Z_x\}_{x=0}^{\omega}$

θ_x=(q_x^N, q_x^{c1}, q_x^{c2},λ_x)，$\theta = \{\theta_x\}_{x=0}^{\omega}$

とおくと，

(1) 年齢間制約条件：$Z = h(\theta)$ を与えるベクトル値関数 h は，年齢間の漸化式

$$h_{x+1} = h_x - h_x\lambda_x - h_x q_x^N$$

$$k_{x+1} = k_x + h_x\lambda_x - k_x q_x^{c1} - k_z q_x^{c2}$$

をもとに $\{\theta_y\}_{y=0}^{x-1}$ $(x = 1, 2, \cdots, \omega)$ を入力として $Z_x = (h_x, k_x)$ を生成するアルゴリズムである．この漸化式を解くと次式が得られる．

$$\begin{aligned} h_x &= h_0 \prod_{y=0}^{x-1}(1 - \lambda_y - q_y^N) \\ k_x &= \sum_{z=0}^{x-1} h_z\lambda_z \prod_{y=z+1}^{x-1}(1 - q_y^{c1} - q_y^{c2}) + k_0 \prod_{y=0}^{x-1}(1 - q_y^{c1} - q_y^{c2}) \end{aligned}$$

ただし，h_0, k_0 は ℓ_0 より定まる初期値である．

(2) 状態間制約条件：$\theta = g(X, Z)$ を与える関数 g は，$(X_x, Z_x) = (q_x, \delta_x, \lambda_x^*, h_x, k_x)$ を入力として $\theta_x = (q_x^N, q_x^{c1}, q_x^{c2}, \lambda_x)$ を定めるアルゴリズムである．

ただし，モデルパラメータ θ_x=(q_x^N, q_x^{c1}, q_x^{c2},λ_x) をそれぞれ絶対脱退率 (肩に $*$ を記して表す) を用

いて次のように表す.

$$
\begin{aligned}
q_x^N &= q_x^{N*}\left(1-\frac{1}{2}\lambda_x^*\right)\\
q_x^{c1} &= q_x^{c1*}\left(1-\frac{1}{2}q_x^{c2*}\right)\\
q_x^{c2} &= q_x^{c2*}\left(1-\frac{1}{2}q_x^{c1*}\right)\\
\lambda_x &= \lambda_x^*\left(1-\frac{1}{2}q_x^{N*}\right)
\end{aligned}
$$

$(q_x^{N*}, q_x^{c1*}, q_x^{c2*}, \lambda_x)$ は以下のように定まる.

$$
\begin{aligned}
q_x^{N*} &= \frac{(1-\delta_x)d_x}{(h_x+(1+m)k_x)\left(1-\frac{1}{2}\lambda_x^*\right)}\\
q_x^{c1*} &= \frac{(1+m)(1-\delta_x)d_x}{h_x+(1+m)k_x}\left(1-\frac{1}{2}\frac{k_x(1-A)+\frac{1}{2}\delta_x d_x-\sqrt{(k_x(1-A)+\frac{1}{2}\delta_x d_x)^2-2k_x\delta_x d_x}}{k_x}\right)^{-1}\\
q_x^{c2*} &= \frac{k_x(1-A)+\frac{1}{2}\delta_x d_x-\sqrt{(k_x(1-A)+\frac{1}{2}\delta_x d_x)^2-2k_x\delta_x d_x}}{k_x}\\
\lambda_x &= \lambda_x^*\left(1-\frac{1}{2}\frac{(1-\delta_x)d_x}{(h_x+(1+m)k_x)(1-\frac{1}{2}\lambda_x^*)}\right)
\end{aligned}
$$

ただし，A, d_x はそれぞれ次の式である.

$$
\begin{aligned}
A &= \frac{(1+m)(1-\delta_x)d_x}{2\left(h_x+k_x(1+m)\right)}\\
d_x &= \ell_0 q_x\prod_{y=0}^{x-1}(1-q_y)
\end{aligned}
$$

式の導出の詳細に関しては付録 1 参照.

4.3 推定結果

この問題では交互イタレーションの収束に初期値依存性は認められなかった．停止条件の誤差を $\varepsilon = 10^{-4}$ とし，イタレーション回数は概ね 30 回前後で収束した．推定結果は付録 2 のとおりであるが，各年齢の総人口は簡易生命表と一致しており（年齢間制約条件），全死亡数が状態別死亡数の和に一致（状態間制約条件）していることから，目的が達成されていることがわかる．ここで推定有病率のグラフ (図 3) を見ると，有病率の観測値（患者調査）と推定値に大きな差異があることを確認できる．これは観測値が回復の影響を受けているためと考えられ，患者調査有病率が回復を見込まないがん多重脱退モデル推定に適さないことがわかる.

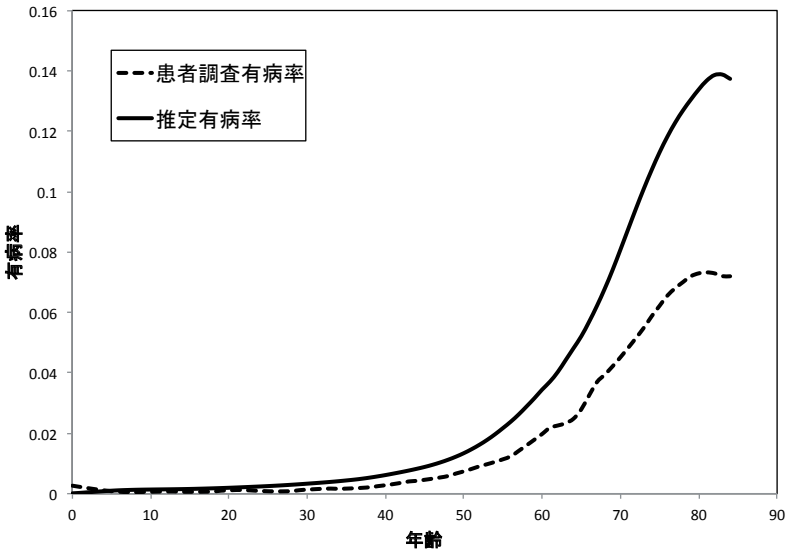

図 3: 有病率の推定結果

5．今後の課題点

実装では死亡指数 m=0 といたが，通常死亡率とがん患者のがん以外による死亡率との間には免疫力の低下等による差異があると考えられるので年齢要素を加味した推定を行うことが必要である．また，今回はがん疾病の多重脱退を対象としたが，交互イタレーション手法は就業不能のように復帰のある多重状態モデル推定などにも幅広く応用が可能であると考えられる．

ただし，年齢要素を加味した死亡指数 m や回復率を含む多重状態モデル推定では初期値依存性への対処が必要となるため，今後の課題とする．

参考文献

[1] 二見隆 [1999], 『生命保険数学（上巻・下巻)』財団法人生命保険文化研究所

[2] 山内恒人 [2014] 『生命保険数理の基礎』東京大学出版社

[3] 友寄一郎 [2015] ”がんの数理” 日本アクチュアリー会 会報 68 号,

[4] 山内恒人，金村慶二，富島直紀 [2012] ”新人アクチュアリー奮闘記〜第 3 分野基礎率作成を命じられて〜” 日本アクチュアリー会 会報第 65 号（第 2 冊分）

[5] 伊藤ゆり，杉本知之 [2011]『地域がん登録資料に基づくがん患者の治癒確率の推定』数計統計；59(2)：287-300

[6] 小西貞則，越智義道，大森祐浩 [2015]『計算統計学の方法―ブートストラップ・EM アルゴリズム・MCMC―』 朝倉書店

[7] David C.M.Dickson,Mary R.Hardy,Howard R. Waters(2013)
”Actuarial Mathematics for Life Contingent Risks”,Cambridge University Press

[8] 厚生労働省大臣官房統計情報部，平成 23 年度患者調査，一般社団法人厚生労働統計協会

[9] 厚生労働省大臣官房統計情報部，平成 23 年度人口動態調査，一般社団法人厚生労働統計協会

[10] 厚生労働省 人口動態・保険社会統計課，平成 23 年 簡易生命表

[11] 国立がん研究センターがん情報サービス「がん登録・統計」地域がん登録全国推計によるがん罹患データ (1975 年〜2011 年)

付録 1：g,h のアルゴリズム

(1) 関数 h のアルゴリズム

初期値：h_0, k_0 は $h_0 + k_0 = \ell_0$を満たすように 絶対脱退率を用いて，

$$h_0 = \frac{\ell_0}{1+\frac{1}{2}\lambda_0} \tag{13}$$

$$k_0 = \frac{1}{2}\frac{\ell_0\lambda_0}{1+\frac{1}{2}\lambda_0} \tag{14}$$

とおく．非がん患者に関する漸化式から h_x は，

$$\begin{aligned}
h_1 &= h_0 - h_0\lambda_0 - h_0 q_0^N = h_0(1-\lambda_0-q_0^N)\\
h_2 &= h_1 - h_1\lambda_1 - h_1 q_1^N = h_0(1-\lambda_0-q_0^N)(1-\lambda_1-q_1^N)\\
&\vdots\\
h_x &= h_{x-1} - h_{x-1}\lambda_{x-1} - h_{x-1}q_{x-1}^N = h_0\prod_{y=0}^{x-1}(1-\lambda_y-q_y^N)
\end{aligned} \tag{15}$$

と定まる．次に，がん患者に関する漸化式から k_x は，

$$\begin{aligned}
k_1 &= k_0 + h_0\lambda_0 - k_0 q_0^{c1} - k_0 q_0^{c2} = h_0\lambda_0 + k_0(1-q_0^{c1}-q_0^{c2})\\
k_2 &= k_1 + h_1\lambda_1 - k_1 q_1^{c1} - k_1 q_1^{c2} = h_1\lambda_1 + k_1(1-q_1^{c1}-q_1^{c2})\\
&= h_1\lambda_1 + h_0\lambda_0(1-q_1^{c1}-q_1^{c2}) + k_0(1-q_0^{c1}-q_0^{c2})(1-q_1^{c1}-q_1^{c2})\\
k_3 &= k_2 + h_2\lambda_2 - k_2 q_2^{c1} - k_2 q_2^{c2} = h_2\lambda_2 + k_2(1-q_2^{c1}-q_2^{c2})\\
&= h_2\lambda_2 + h_1\lambda_1(1-q_2^{c1}-q_2^{c2}) + h_0\lambda_0(1-q_1^{c1}-q_1^{c2})(1-q_2^{c1}-q_2^{c2})\\
&\quad + k_0(1-q_0^{c1}-q_0^{c2})(1-q_1^{c1}-q_1^{c2})(1-q_2^{c1}-q_2^{c2})\\
&\vdots\\
k_x &= \sum_{z=0}^{x-1} h_z\lambda_z \prod_{y=z+1}^{x-1}(1-q_y^{c1}-q_y^{c2}) + k_0\prod_{y=0}^{x-1}(1-q_y^{c1}-q_y^{c2})
\end{aligned} \tag{16}$$

と定まる．

(2) 関数 g のアルゴリズム

モデルパラメータ θ_x=(q_x^N, q_x^{c1}, q_x^{c2},λ_x) をそれぞれ絶対脱退率を用いて次のように表す．

$$q_x^N = q_x^{N*}\left(1-\frac{1}{2}\lambda_x^*\right) \tag{17}$$

$$q_x^{c1} = q_x^{c1*}\left(1-\frac{1}{2}q_x^{c2*}\right) \tag{18}$$

$$q_x^{c2} = q_x^{c2*}\left(1-\frac{1}{2}q_x^{c1*}\right) \tag{19}$$

$$\lambda_x = \lambda_x^*\left(1-\frac{1}{2}q_x^{N*}\right) \tag{20}$$

(17)～(20) 中の $q_x^{N*}, q_x^{c1*}, q_x^{c2*}, \lambda_x$ を以下で求める．

各年齢 $x(x=0,1,\cdots,\omega)$ において状態間で成り立つ関係式が (21)(22) である．また，パラメータ θ のうち q_x^N と q_x^{c1} の間には死亡指数 $m(\in\mathbb{R})$ を用いた仮定 (23) を置く．

$$d_x = h_x q_x^N + k_x q_x^{c1} + k_x q_x^{c2} \tag{21}$$

$$\delta_x : 1-\delta_x = k_x q_x^{c2} : h_x q_x^N + k_x q_x^{c1} \tag{22}$$

$$q_x^{c1} = (1+m) q_x^N \tag{23}$$

ただし，(21) は死亡者数合計の関係式であり，(22) は死亡者割合の関係式である．(23) は絶対脱退率 (17)(18) を用いて表すと，

$$q_x^{c1*}\left(1-\frac{1}{2}q_x^{c2*}\right) = (1+m) q_x^{N*}\left(1-\frac{1}{2}\lambda_x^*\right) \tag{24}$$

と表せる．(21) に (17)∼(19) を代入する．

$$\begin{aligned} d_x &= h_x q_x^N + k_x q_x^{c1} + k_x q_x^{c2} \\ &= h_x q_x^{N*}\left(1-\frac{1}{2}\lambda_x^*\right) + k_x q_x^{c1*}\left(1-\frac{1}{2}q_x^{c2*}\right) + k_x q_x^{c2*}\left(1-\frac{1}{2}q_x^{c1*}\right) \\ &= (h_x + k_x(1+m)) q_x^{N*}\left(1-\frac{1}{2}\lambda_x^*\right) + k_x q_x^{c2*}\left(1-\frac{1}{2}q_x^{c1*}\right) \end{aligned} \tag{25}$$

また，(22) の右辺は (17)∼(19) を用いて，

$$k_x q_x^{c2} : h_x q_x^N + k_x q_x^{c1} = k_x q_x^{c2*}\left(1-\frac{1}{2}q_x^{c1*}\right) : h_x q_x^{N*}\left(1-\frac{1}{2}\lambda_x^*\right) + k_x q_x^{c1*}\left(1-\frac{1}{2}q_x^{c2*}\right) \tag{26}$$

となるので，(22) は (24)(25)(26) を用いて整理すると，

$$\begin{aligned} \delta_x q_x^{N*}\left(1-\frac{1}{2}\lambda_x^*\right)(h_x+(1+m)k_x) &= (1-\delta_x) k_x q_x^{c2*}\left(1-\frac{1}{2}q_x^{c1*}\right) \\ &= (1-\delta_x)\left\{ d_x - (h_x+k_x(1+m)) q_x^{N*}\left(1-\frac{1}{2}\lambda_x^*\right)\right\} \\ &= (1-\delta_x) d_x - (h_x+k_x(1+m)) q_x^{N*}\left(1-\frac{1}{2}\lambda_x^*\right) \\ &\quad + \delta_x q_x^{N*}\left(1-\frac{1}{2}\lambda_x^*\right)(h_x+(1+m)k_x) \end{aligned} \tag{27}$$

となる．(27) を q_x^{N*} に関して解くと，

$$q_x^{N*} = \frac{(1-\delta_x) d_x}{(h_x+(1+m)k_x)\left(1-\frac{1}{2}\lambda_x^*\right)} \tag{28}$$

を得る．q_x^{N*} が定まることで，(17) から λ_x は，

$$\lambda_x = \lambda_x^*\left(1-\frac{1}{2}\frac{(1-\delta_x) d_x}{(h_x+(1+m)k_x)(1-\frac{1}{2}\lambda_x^*)}\right) \tag{29}$$

と定まる．次に，q_x^{c2*} を求める．(24)(25) から q_x^{c1*} を消去すると，

$$\begin{aligned} d_x &= (h_x + k_x(1+m))q_x^{N*}\left(1-\frac{1}{2}\lambda_x^*\right) + k_x q_x^{c2*} - \frac{1}{2}k_x q_x^{c2*} q_x^{c1*} \\ &= (h_x + k_x(1+m))q_x^{N*}\left(1-\frac{1}{2}\lambda_x^*\right) + k_x q_x^{c2*} - \frac{k_x}{2}q_x^{c2*}\left(\frac{1+m}{1-\frac{1}{2}q_x^{c2*}}\left(1-\frac{1}{2}\lambda_x^*\right)q_x^{N*}\right) \end{aligned} \tag{30}$$

となる．ここで，

$$A = \frac{(1+m)(1-\delta_x)d_x}{2\left(h_x + k_x(1+m)\right)} \tag{31}$$

とおく．(30) に (31) を用いて整理すると

$$\begin{aligned} & k_x q_x^{c2*} - d_x\delta_x = \frac{k_x q_x^{c2*}}{1-\frac{1}{2}q_x^{c2*}}A \\ \longleftrightarrow\quad & \frac{k_x}{2}\left(q_x^{c2*}\right)^2 + \left(Ak_x - \frac{d_x\delta_x}{2} - k_x\right)q_x^{c2*} + d_x\delta_x = 0 \\ \longleftrightarrow\quad & q_x^{c2*} = \frac{k_x(1-A) + \frac{1}{2}\delta_x d_x - \sqrt{(k_x(1-A)+\frac{1}{2}\delta_x d_x)^2 - 2k_x\delta_x d_x}}{k_x} \end{aligned} \tag{32}$$

となり，q_x^{c2*} を得る．(24) に求めた q_x^{c2*} を代入することで q_x^{c1*} は，

$$\begin{aligned} q_x^{c1*} &= \frac{1}{1-\frac{1}{2}q_x^{c2*}}(1+m)q_x^{N*}\left(1-\frac{1}{2}\lambda_x^*\right) \\ &= \frac{(1+m)(1-\delta_x)d_x}{h_x + (1+m)k_x}\left(1-\frac{1}{2}\frac{k_x(1-A)+\frac{1}{2}\delta_x d_x - \sqrt{(k_x(1-A)+\frac{1}{2}\delta_x d_x)^2 - 2k_x\delta_x d_x}}{k_x}\right)^{-1} \end{aligned} \tag{33}$$

と定まる． □

付録 2:脱退残存表 (回復なしモデル)

表 1: 脱退残存表 (回復なしモデル)

年齢	ℓ_x	d_x	h_x	k_x	d_x^N	d_x^{c1}	d_x^{c2}	i_x
0	100000	234	99989	11	230	0	4	22
1	99766	41	99737	29	40	0	1	21
2	99725	31	99676	49	30	0	1	20
3	99694	23	99625	69	22	0	1	19
4	99671	19	99584	87	18	0	1	18
5	99652	17	99548	104	16	0	1	15
6	99635	16	99517	118	14	0	2	13
7	99619	14	99490	129	12	0	2	11
8	99605	13	99467	138	11	0	2	9
9	99592	12	99447	145	10	0	2	8
10	99580	11	99430	151	9	0	2	7
11	99569	12	99414	156	10	0	2	7
12	99557	13	99397	161	11	0	2	7
13	99544	16	99379	166	14	0	2	8
14	99528	19	99357	172	17	0	2	9
15	99509	24	99331	179	21	0	2	11
16	99485	30	99299	187	28	0	3	13
17	99455	37	99259	197	33	0	3	14
18	99418	43	99211	207	39	1	4	16
19	99375	48	99157	218	43	1	4	17
20	99327	52	99097	231	48	1	4	18
21	99274	57	99031	244	53	1	4	19
22	99217	63	98960	258	57	1	4	20
23	99154	67	98883	272	62	1	4	20
24	99087	68	98801	287	63	1	4	21
25	99019	67	98717	302	62	1	4	22
26	98952	65	98634	318	60	1	4	23
27	98887	65	98551	336	59	1	4	25
28	98822	66	98468	355	60	1	5	27
29	98756	69	98381	375	62	2	6	29
30	98687	71	98290	397	63	2	6	31
31	98616	73	98196	420	64	2	7	33
32	98543	75	98099	444	65	2	8	36
33	98468	78	97998	470	68	2	9	40
34	98390	83	97890	499	71	2	10	44
35	98307	88	97776	531	74	3	11	47
36	98220	93	97655	564	78	3	12	53
37	98127	99	97524	603	83	3	13	61
38	98028	106	97380	647	88	4	14	68
39	97922	116	97225	697	95	4	16	74
40	97806	128	97055	751	103	5	19	85
41	97679	140	96868	811	112	6	23	93
42	97538	153	96663	876	120	7	26	102
43	97385	165	96441	944	127	8	29	112
44	97220	178	96202	1019	135	9	33	124
45	97043	192	95942	1101	144	11	37	140
46	96851	208	95658	1193	152	12	43	160
47	96643	228	95346	1297	163	15	51	185
48	96415	251	94999	1417	175	17	60	214
49	96164	278	94611	1553	188	20	70	247
50	95886	309	94176	1709	200	24	85	290
51	95577	341	93685	1891	215	29	98	329
52	95235	372	93141	2093	227	34	110	368
53	94863	398	92546	2318	237	39	123	410
54	94465	427	91899	2566	244	45	138	455
55	94038	461	91200	2838	251	51	158	501
56	93577	505	90449	3129	261	60	183	570
57	93072	557	89618	3455	273	69	215	650
58	92515	613	88695	3821	277	79	257	736
59	91902	671	87682	4221	284	90	296	823
60	91231	733	86575	4657	298	106	329	855
61	90499	802	85422	5077	317	124	361	958
62	89697	867	84147	5551	333	145	390	1066
63	88830	928	82748	6082	341	165	421	1116
64	87902	993	81291	6611	347	186	459	1181
65	86909	1068	79764	7146	360	213	496	1304
66	85841	1155	78100	7741	374	245	536	1396
67	84686	1246	76330	8358	389	281	577	1488
68	83440	1332	74453	8988	400	318	613	1572
69	82108	1410	72482	9628	408	358	644	1641
70	80699	1498	70432	10268	423	407	668	1666
71	79200	1609	68343	10859	439	460	710	1706
72	77591	1741	66199	11395	460	523	758	1738
73	75850	1890	64000	11853	487	595	808	1763
74	73961	2049	61751	12212	517	675	857	1782
75	71912	2222	59453	12462	550	761	909	1796
76	69690	2411	57107	12587	591	861	959	1805
77	67279	2625	54711	12572	642	975	1007	1810
78	64654	2847	52260	12400	700	1099	1047	1808
79	61807	3067	49752	12062	765	1227	1073	1797
80	58740	3272	47190	11558	831	1347	1093	1771
81	55468	3457	44587	10890	896	1448	1112	1724
82	52011	3624	41967	10054	968	1533	1122	1643
83	48387	3771	39357	9042	1052	1599	1120	1567
84	44616	3891	36738	7890	1151	1635	1104	1467

A Learning Algorithm for Coherent Multiple State Modeling

Tatsunori Onoe　Naoki Matsuyama

Meiji University, 4-21-1 Nakano Nakano-ku Tokyo 164-8525, Japan
ma2yama(at)meiji.ac.jp

Abstract

Multiple state modeling from official statistics has many difficulties caused by the data deficiency, data inconsistency with stationary model assumptions and rough granularity of the data. The purpose of this study is to propose a learning algorithm and an interpolation algorithm, to be called as alternate iteration and integral interpolation, respectively, to cope with these difficulties. As an implementation of these algorithms, we estimate a coherent multiple state model of cancer disease from Japanese official statistics.