
研究論文

一般化加法モデルを用いた脱退率推定の改善

車谷優樹^{*†} 岩沢宏和[‡]

2025 年 8 月 27 日投稿

2025 年 11 月 17 日受理

概要

本研究は、確定給付企業年金制度の脱退率推定における現行手法の問題点を明らかにし、この問題点を改善する脱退率推定を提案する。具体的には、ベルヌーイ分布を仮定した一般化加法モデル (GAM) を導入し、個々の加入者の年齢と脱退の有無を記録した観測データに対し、各年齢の脱退実績数に応じて脱退率を滑らかに推定する方法を示す。数値実験の結果、提案手法は現行手法に比べ安定的かつ精度の高い推定を可能にし、また信頼区間を用いることで財政悪化リスク相当額の算定に応用可能であることが確認された。

キーワード：企業年金、確定給付企業年金、脱退率推定、一般化加法モデル (GAM)

1 はじめに

確定給付企業年金制度 (Defined Benefit Pension Plan, 以下、「DB」) における脱退率とは、加入者が自己都合や転職などの理由により生存状態で脱退する確率を年齢ごとに示したものであり、制度運営において掛金額の算定や数理債務の評価に用いられる重要な計算基礎の 1 つである。脱退率は DB ごとに過去数年間の脱退実績を用いて算定され、通常 5 年ごとに見直しが行われる。しかし現行の手法による脱退率推定には、いくつかの問題点が指摘されており、これは特に加入者数が小規模な DB において顕著である。

本研究の目的は、現行の手法の問題点を改善し、より安定的かつ客観的な脱退率推定を可能にするため、一般化加法モデル (Generalized Additive Model, GAM) を用いた脱退率推定手法を提案し、数

値実験によりその有効性を検証することにある。GAM は Wood[2025]による R パッケージ “mgcv” (以下、「mgcv パッケージ」) を用いた実装が容易であることから、アクチュアリー実務にも導入しやすいと考えられる。更に、この提案手法は脱退率算定に関する財政悪化リスク相当額の算定に応用可能であることを述べる。

2 現行の推定方法とその問題点

2.1 現行の脱退率推定

現行の脱退率推定は、直近 3 年以上の加入者の脱退実績等をもとに、推定に使用する年度を決定 (以下では 3 年分の場合を想定) した上で、次の 2 段階を経て行われる。

1. 粗製脱退率の作成
2. 粗製脱退率の補整

^{*}大阪大学大学院基礎工学研究科 〒560-8531 大阪府豊中市待兼山町 1-3
email: kurumadani@sigmath.es.osaka-u.ac.jp

[†]株式会社りそな銀行 〒540-8610 大阪府大阪市中央区備後町 2-2-1 email: Yuki.Kurumadani@resonabank.co.jp

[‡]早稲田大学大学院会計研究科 〒169-8050 東京都新宿区西早稲田 1-6-1 email: iwahiro@bb.mbn.or.jp

2.1.1 粗製脱退率の作成

まず「粗製脱退率」と呼ばれる、各年齢の脱退者を加入者数で除した値を算定する(図2-1参照)。

$$x \text{ 歳の粗製脱退率} = \frac{x \text{ 歳の脱退者数}}{x \text{ 歳の加入者数}}$$

分子分母は過去3年間の実績を合計して算定することが一般的である。ただし分母が0の場合はこの式を用いて計算することはできない。

X-2年度			X-1年度			X年度			3年合計			脱退率		
年齢	加入者	脱退者	加入者	脱退者	加入者	脱退者	加入者	脱退者	年齢	加入者	脱退者	粗製脱退率	年齢	補正後
19	8	1	1	0	0	0	0	0	19	9	1	11.11%	19	7.17%
20	23	0	7	1	1	0	0	0	20	31	1	3.23%	20	6.74%
21	37	4	23	0	8	0	0	0	21	68	4	5.88%	21	6.55%
22	41	1	35	1	24	4	0	0	22	100	6	6.00%	22	5.04%

図2-1 粗製脱退率の作成

粗製脱退率は年齢ごとに大きなばらつきが生じる傾向にある。これは特に小規模なDBにおいて顕著である。このため日本年金数理人会[2025](以下、「DB実務ガイダンス」)では、「連続した複数の年齢による集団を作成し、粗製脱退率を求めることもできる」(例えば5歳幅など)と規定されている。このように算定年齢幅を拡大することで、粗製脱退率の年齢ごとのばらつきを軽減でき、ある年齢の加入者数が0人である場合でも、粗製脱退率が適切に計算できるようになる(図2-2参照)。

年齢	各歳幅			5歳幅		
	加入者	脱退者	粗製脱退率	加入者	脱退者	粗製脱退率
19	9	1	11.11%	250	15	6.00%
20	31	1	3.23%	250	15	6.00%
21	68	4	5.88%	250	15	6.00%
22	100	6	6.00%	250	15	6.00%
23	42	3	7.14%	250	15	6.00%
24	25	1	4.00%	80	5	6.25%
25	10	0	0.00%	80	5	6.25%
26	0	0	#DIV/0!	80	5	6.25%
27	12	2	16.67%	80	5	6.25%
28	33	2	6.06%	80	5	6.25%

図2-2 粗製脱退率の作成(左:各歳幅, 右:5歳幅)

「5歳幅」以外の算定年齢幅の拡大の方法として、「その年齢の前後に同年齢だけ、粗製脱退率の分母となる人数が30人以上となるまで拡大する」といった方法もある。これは日本年金数理人会[2014]に記載されている方法であり、「年齢毎の粗製脱退率の分母となる人数が算定年度幅を通算して30人以

上となる時その脱退率は有意と判定し実績値を用いる」との考えに由来するものである。

なお、これは代行保険料率の計算に使用される方法であり、代行保険料率の計算では予定脱退率の平滑化(補整)は行われなかったとされていることに注意が必要である。

2.1.2 粗製脱退率の補整

次に、粗製脱退率に対して移動平均法やGreville[1979]による補整式などを用いて、年齢ごとの脱退率を滑らかな曲線に補整する(図2-3)。これは「年齢ごとの脱退率に大きなばらつきが無いはず」との考え方(経験則)から、統計資料の偶発的な要素や変動を排除し、凹凸の少ない滑らかな曲線となるように粗製脱退率を補整するものであり、実務上広く行われている。

年齢 x における粗製脱退率 c_x は、有限の加入者から得られた経験比率であり、真の脱退率を中心としたノイズ付き観測値とみなせる。補整の目的は、ノイズを抑えつつ真の滑らかな関数構造を保存することである。



図2-3 粗製脱退率の補整

粗製脱退率 $\{c_x\}$ に対し、補整後の脱退率 $\{q_x\}$ は次式で与えられる。

・ $2n+1$ 点移動平均法:

$$q_x = \sum_{t=-n}^n \frac{1}{2n+1} c_{x+t}$$

・ m 次 $2n+1$ 項式Greville法:¹

$$q_x = \sum_{t=-n}^n L_t c_{x+t}$$

¹ $L_{-t} = L_t$ を満たす。厳密には外挿値を計算するための係数 $\{a_t\}$ も必要になるが、ここでは触れない。

t \ n	2	3	4	5	6
0	0.559440	0.412588	0.331140	0.277944	0.240058
1	0.293706	0.293706	0.266557	0.238693	0.214337
2	-0.073426	0.058741	0.118470	0.141268	0.147356
3	—	-0.058741	-0.009873	0.035723	0.065492
4	—	—	-0.040724	-0.026792	0.000000
5	—	—	—	-0.027864	-0.027864
6	—	—	—	—	-0.019350

図 2-4 Greville 法 ($m = 3$) における補整係数 $\{L_t\}$

m 次 $2n + 1$ 項式 Greville 法における補整係数 $\{L_t\}$ (図 2-4) は、任意の m 次以下の多項式 $P(x)$ を補整前後で保つという制約条件 (m 次多項式再現条件)

$$P(x) = \sum_{t=-n}^n L_t P(x + t)$$

のもとで、平滑後の系列の 3 回差分

$$\Delta^3 q_x = q_{x+3} - 3q_{x+2} + 3q_{x+1} - q_x$$

の 2 乗和 $\sum_x (\Delta^3 q_x)^2$ を (ガウスノイズに対して) 平均的に最小化するように $\{L_t\}$ を選択する。「 m 次多項式再現条件」は、真の脱退率が m 次多項式で表せる場合に、それが補整後も保持されるべきとの考えに基づくものであり、本来の関数系を壊さずに雑音だけを除去するという補整操作の本質を支える基礎条件であると言える。

上記の補整方法は、離散的かつ決定論的な枠組みで平滑化を行うものであり、経験率のノイズをある程度抑える実務的な有効性を備えている。

2.1.3 外部データ脱退率の使用

小規模な DB における脱退率として、「DB 実務ガイドダンス」には、「同種の企業で使用されている脱退率」、「業界毎の統計資料から推定した脱退率」(以下、「外部データ脱退率」) または「外部データ脱退率に対して合理的な補正を行った脱退率」を使用できるとされている。

2.2 現行の脱退率推定の問題点

現行の推定方法には、以下の問題点が存在する。
問題点 1: 移動平均法や Greville などの補整方法は、粗製脱退率の信頼性を、その算出分母の大きさに関わらず一様に扱ってしまう。

問題点 2: Greville 法による補整結果はマイナスになる場合があり、補整結果をそのまま使用できない場合がある。

問題点 3: 粗製脱退率作成の際、「算定年齢幅の拡大」の方法によって結果が大きく変わり得る。

問題点 4: 外部データ脱退率をどのように脱退率推定に組み込むべきかが明らかでない。

問題点 5: 現行の脱退率は決定論的な手法により推定されており、リスク定量化が難しい。

2.2.1 問題点 1: 補整に粗製脱退率の各年齢の情報量の価値を反映できない

移動平均法や Greville などの補整手法は、各年齢の脱退実績数の大小を考慮しないため、情報量の少ないデータと多いデータを同一に扱ってしまう問題を含んでいる。この問題はとりわけ小規模 DB において顕著である。

問題点 1 を簡単な例で確かめる。表 2-1 の通り、31 歳の実績数のみ極端に少ないケース 1 と、どの年齢でも実績数が等しいケース 2 を考える。両者の粗製脱退率は同じであるため、現行の手法であれば両者の脱退率推定を機械的に区別することはできない。

表 2-1 年齢規模が異なるケース (脱退者数/加入者数)

	ケース 1	ケース 2	粗製脱退率
30 歳	50/100 人	1/2 人	50%
31 歳	0/2 人	0/2 人	0%
32 歳	50/100 人	1/2 人	50%
33 歳	50/100 人	1/2 人	50%

表 2-1 の粗製脱退率に対して移動平均法による補整結果は図 2-5 となる。これは全ての年齢の実績を同列に扱った結果であるから、ケース 2 の推定結果としては適切であるが、31 歳の実績の信頼度が他の年齢よりも極端に低いケース 1 の場合は適切とは言えず、31 歳の実績をほぼ無視したような結果 (50% 一定率) に修正することが望ましい。

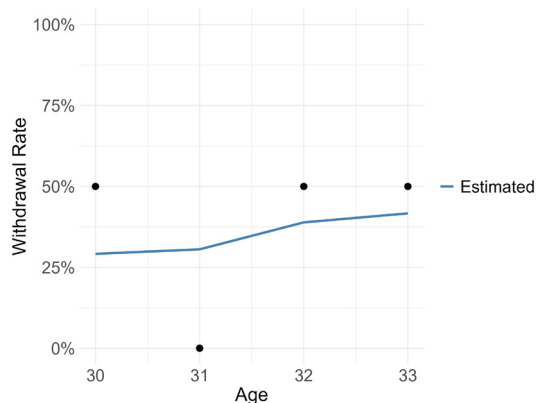


図 2-5 移動平均法による推定結果

2.2.2 問題点 2：補整結果が負値を取り得る

Greville 法における補整係数にはマイナスのものを含むため、補整結果がマイナスになる場合がある。この場合、脱退率（0 以上 1 以下の値を取る）の推定結果としては不適切であり、実務上「マイナスのものは 0 とする」といった補整が必要になる。この場合、 m 次多項式再現条件が満たされないこととなり、補整操作の意図を結果に十分に反映できない恐れがある。

2.2.3 問題点 3：算定年齢幅の拡大方法

小規模集団における算定年齢幅の拡大には恣意性が介入する恐れが高く、適切な幅の設定が困難である。

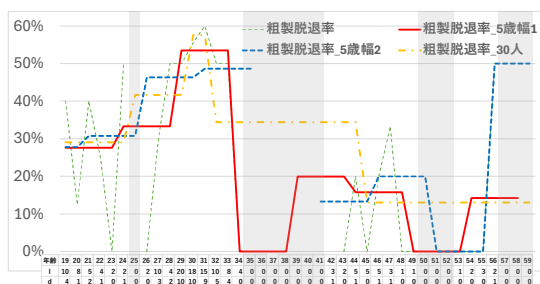


図 2-6 様々な算定年齢幅の拡大方法による粗製脱退率の違い

例えば図 2-6 において、19 歳から 5 歳ごとに集団を分けて粗製脱退率を作成した場合（「粗製脱退率_5 歳幅_1」）、最も高年齢の集団が 54 歳～58 歳、59 歳となり、均等に年齢を割り当てることができない。

例えば最後の集団のみ 54 歳～59 歳の集団とすることも考えられるが、1 つだけ 6 歳の集団が作成されることとなり、恣意性が介入した結果となる可能性がある。

図 2-6 の「粗製脱退率_5 歳幅_2」は 21 歳から 5 歳ごとに集団を分ける場合であるが、この場合、36 歳～41 歳の分母が 0 となってしまう、そのままでは粗製脱退率を作成することができない。

図 2-6 の「粗製脱退率_30 人」は分母が 30 人となるまで年齢幅を拡大して粗製脱退率を作成した場合であるが、33 歳以降 2 つの値しか取っておらず、年齢ごとの脱退実績の違いを十分に反映できていない。

以上 3 通りの拡大方法を示したが、得られる粗製脱退率の形状は全く異なっており、その拡大方法の選び方に結果が大きく依存していることが分かる。

2.2.4 問題点 4：外部データ脱退率の取り入れ方

外部データ脱退率をどのように適切に取得するかという課題は存在するが、仮に外部データ脱退率を適切に取得できたとしても、その情報を推定にどのように組み込むべきかについては明確な基準が存在せず、実際には外部データ脱退率はほとんど活用されていないのが現状である。

2.2.5 問題点 5：リスク定量化が困難

これについては第 5 章で述べる。

2.3 先行研究

2.3.1 補整に関する一般論

本節では年齢を x 、粗製脱退率を $\{c_x\}$ 、補整後の脱退率を $\{q_x\}$ と表す。脱退率に限らず、補整方法に関する先行研究は多くある。基本的な考え方は、

- (1) $\{c_x\}, \{q_x\}$ の乖離度合（乖離度、逸脱度）
- (2) $\{q_x\}$ の平滑度合（平滑度）

をそれぞれ定量化し、適当な関数空間内で

乖離度 $+$ $\lambda \times$ 平滑度； λ ：正の定数 (2.1)
の最小化問題を解くことによって、(1) (2) をともに小さくする $\{q_x\}$ を得る、というものである。

(1) の乖離度は、例えば 2 乗残差、つまり

$$\text{乖離度} = \sum_x (q_x - c_x)^2$$

のように定式化される．更に年齢 x ごとの価値を反映した乖離度は，年齢ごとに異なる重み w_x を用いて，

$$\text{乖離度} = \sum_x w_x (q_x - c_x)^2 \quad (2.2)$$

などと定式化される．一方で(2)の平滑度は，

$$m \text{ 次差分平滑度} = \sum_x (\Delta^m q_x)^2 \quad (2.3)$$

$$m \text{ 次微分平滑度} = \int_x \left(\frac{d^m q_x}{dx^m} \right)^2 dx \quad (2.4)$$

(ここで $\Delta q_x := q_{x+1} - q_x$) などが一般的である．既に述べた通り，Greville の補整法は 3 次差分平滑度を使用している．(2.4)について，次に注意する．

$$(2.4) = 0 \Leftrightarrow q_x : m-1 \text{ 次式} \quad (2.5)$$

つまり(2.1)において平滑度が重視される場合の推定結果は $m-1$ 次式に近づく．これは例えば真の脱退率の形状が $m-1$ 次式と想定される場合などに合理的である．

(2.1)における λ は 0 に近づけると乖離度を小さくすることが重視され， $+\infty$ に近づけると平滑度を小さくすることが重視される．この意味で λ は乖離度と平滑度をどの程度重視するかを司るパラメータであり，平滑化パラメータ (smoothing parameter) と呼ばれる．

Whittaker[1923]は最小化問題(2.1)の解を考察することで，等間隔な観測データ点を平滑化する方法を提案した．

Schoenberg[1964]は，任意間隔のデータに対し，乖離度を(2.2)，平滑度を(2.4)とした最小化問題の解が自然 $2m-1$ 次スプライン関数で一意に与えられることを示した(下記命題 2-2 参照)．特にこれは大域的に滑らかな関数である．これらにより問題点 1 の解決が図られるが，年齢ごとの重み w_x や正の定数 λ をどのように決定すれば良いかに関する問題点が生じる．

2.3.2 自然 $2m-1$ 次スプラインの一般論

以下，1 次元データの場合に限定した議論を行う．

ここでは自然 $2m-1$ 次スプライン ($m = 1, 2, \dots$) の定義と，良く知られた一般論について述べる (詳細は桜井[1981]，Wood[2017]を参照)．後で命題 2-2 として述べる通り，(2.2)に限定しない一般的な乖離度の場合に，平滑度を(2.4)とした場合の最小化問題(2.1)の解は自然 $2m-1$ 次スプライン関数で与えられることが知られている．

定義 2-1.

$2m-2$ 回微分まで連続な区分 $2m-1$ 次多項式で両端では高々 $m-1$ 次多項式である関数 (ないし曲線) を自然 $2m-1$ 次スプラインという．複数の区分 $2m-1$ 次式が連続する点を結び目 (knot) という．

例えば knot の数が 11 個の自然 3 次スプラインの一例を図 2-7 に示した．それぞれの区間では 3 次多項式で，両端では 1 次式 (直線) となっている．

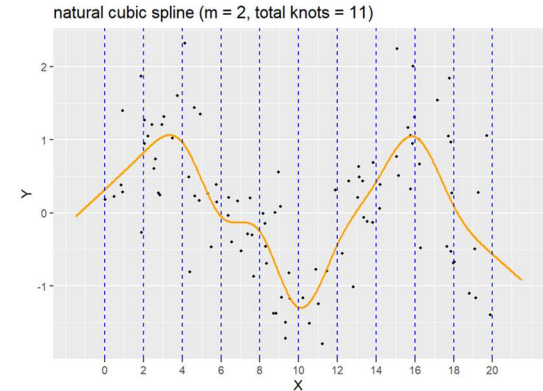


図 2-7 knot の数が 11 個の自然 3 次スプライン

以下，本節において， $a \leq x_1 < \dots < x_K \leq b$ とし， $\{x_1, \dots, x_K\}$ を knot とする自然 $2m-1$ 次スプラインの (関数空間としての) 基底について調べる．

まず $m = 2$ の場合，自然 3 次スプラインの基底関数の取り方の 1 つとして，3 次回帰スプラインと呼ばれるものの場合を述べる． K 個の knot $x_1, \dots, x_K$ を持つ 3 次スプライン関数 s について， $\beta_j = s(x_j)$ ， $\delta_j = s''(x_j)$ とおいたとき， s は $x_j \leq x \leq x_{j+1}$ の範囲で以下のように表される．

$$s(x) = \beta_j a_j^-(x) + \beta_{j+1} a_j^+(x) + \delta_j c_j^-(x) + \delta_{j+1} c_j^+(x)$$

$$a_j^-(x) := \frac{x_{j+1} - x}{h_j}, \quad a_j^+(x) := \frac{x - x_j}{h_j}$$

$$c_j^-(x) = \frac{\frac{(x_{j+1} - x)^3}{h_j} - h_j(x_{j+1} - x)}{6}$$

$$c_j^+(x) = \frac{\frac{(x - x_j)^3}{h_j} - h_j(x - x_j)}{6}$$

$$h_j = x_{j+1} - x_j$$

これは既に各 x_j において s が連続かつ 2 回微分が連続であることを満たしている. 更に s が「1 回微分が連続」, 「2 回微分が端点 x_1, x_K においては 0」との条件から $(\delta_1, \dots, \delta_K)$ は $(\beta_1, \dots, \beta_K)$ を用いて表すことができ,

$$s(x) = \sum_{i=1}^K \beta_i b_i(x) \quad (2.6)$$

と書き直すことができる. つまり K 個の knot を持つ自然 3 次スプラインは, knot における y 座標 β_j を指定すると一意に存在し (命題 C-1 参照), (2.6) のように表される. 特に自然 3 次スプラインのパラメータ数 (自由度) は knot の個数 K に等しいことが分かる. この $\{b_i(x)\}_i$ を基底関数とする自然 3 次スプラインを 3 次回帰スプライン (cubic regression splines) と呼び, mgcv パッケージでは, bs= “cr” として実装されている.

この場合, knot を指定する必要がある. なお, mgcv パッケージにおいて knot を指示しない場合, 説明変数を上から 9 等分した各分位点 (0 分位点, 100/9 分位点, 200/9 分位点, ..., 100 分位点; 計 10 個) が既定値として用いられる.

次に, m 次微分平滑度に対する薄板スプライン (thin plate splines) について述べる. まず一般に自然 $2m-1$ 次スプラインは次の基底を持つことを確認する.

命題 2-1.

$m < K$ のとき, $\{x_1, \dots, x_K\}$ を knot とする自然 $2m-1$ 次スプラインは,

$$p_{m-1}(x) + \sum_{i=1}^K c_i |x - x_i|^{2m-1} \quad (2.7)$$

と一意的に表される. ここで $p_{m-1}(x)$ は $m-1$ 次多項式であり, 係数 $\{c_i\}$ は

$$\sum_{i=1}^K c_i x_i^r = 0 \quad (r = 0, \dots, m-1) \quad (2.8)$$

を満たす.

(証明は Appendix A を参照.)

命題 2-1 において, $(c_1, \dots, c_K)$ に関する制約条件 (2.8) を考慮した (2.7) の表現をもとに, knot を定めずに最適化問題を解くことによって求まる自然 $2m-1$ 次スプラインを薄板スプライン (thin plate splines : Duchon[1977]) という. 2 次元以上のデータについても同じような命題が成立するが, 詳細は Wood[2017] を参照.

mgcv パッケージでは bs= “tp” (既定値) として実装されている. mgcv パッケージでは knot を指定する必要はなく, 後述の通り, 観測データの全ての観測値をベースに自動で決定される. また微分平滑度の次数 m を指示できるが, 既定値は $m=2$ と設定されている.

mgcv パッケージにおいて, cr と tp ($m=2$ の場合) の基底関数の違いを図 2-8 に図示した. 主な違いとして, cr の場合の基底関数は, 各 knot において, 1 つを除いて $(x, y) = (\text{knot}, 0)$ を通り, 除いた 1 個の knot においては $(x, y) = (\text{knot}, 1)$ を通る (いわゆる cardinal スプライン基底関数) が, tp ($m=2$) の場合はこのような挙動を取らず, $\{1, x\}$ が基底関数に含まれていることが挙げられる.

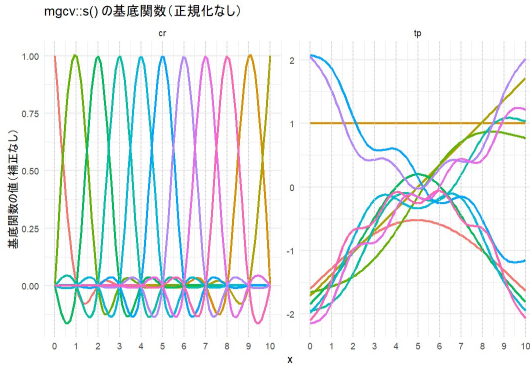


図 2-8 cr(左)と tp(右)の基底関数の違い(knot = {0, 1, ..., 10} の場合)

以下、 $a \leq x_1 \leq x_2 \leq \dots \leq x_n \leq b$ とする．与えられた観測データ

$$(x_1, y_1), \dots, (x_n, y_n) \quad (x_i \in \mathbb{R}, y_i \in \mathbb{R})$$

に対し、任意の損失関数 L を用いた汎関数

$$J(f) := L(f(x_1), y_1, \dots, f(x_n), y_n) + \lambda \int_a^b \{f^{(m)}(x)\}^2 dx$$

を最小化する関数 $f: \mathbb{R} \rightarrow \mathbb{R}$ を考える．ここで $\lambda > 0$ 、 m は 1 以上の整数とし、関数 f は

$$W_m[a, b] := \{f: \mathbb{R} \rightarrow \mathbb{R} \mid m \text{ 回以下の微分は連続, } \int_a^b \{f^{(m)}(x)\}^2 dx < +\infty\}$$

の元であるとする．なお、 $(x_1, \dots, x_n)$ は同じものを含んでいてもよく、相異なる年齢の個数を K とする．

命題 2-2.

$m < K$ を満たすとする．任意の $f \in W_m[a, b]$ に対して、 $J(f) \geq J(s)$ を満たすような、 $\{x_1, \dots, x_n\}$ を knot とする自然 $2m - 1$ 次スプライン s が存在する．

(証明は Appendix B を参照)

注意 2-1.

命題 2-2 から、 $J(f)$ が最小となる関数を見つける問題は、 $\{x_1, \dots, x_n\}$ を knot とする自然 $2m - 1$ 次スプラインの中から見つければ良いことが分かる．

この事実は、 $(x_1, \dots, x_n)$ に同じものを含まず、かつ損失関数が重み付き 2 乗損失

$$L(f(x_1), y_1, \dots, f(x_n), y_n) = \sum_{i=1}^n w_i \{y_i - f(x_i)\}^2$$

の場合、Schoenberg[1964]により証明されているが、

ここでは $(x_1, \dots, x_n)$ が同じものを含み、かつ一般の損失関数においても同様の命題が成り立つことを主張している．

2.3.3 脱退率に関する先行研究

脱退率に関する先行研究として、岩沢[1991]による「乖離度期待値一致法」がある．これは最小化問題 (2.1) において、乖離度を尤もらしい値(確率変数としての期待値)に固定し、3 次差分平滑度の最小化を図るものであり、年齢ごとの重み w_x は、年齢 x での人数を l_x として、

$$w_x := \frac{l_x}{q_x(1 - q_x)}$$

を使用する．この手法により問題点 1 は解決されるが、「解が一意に定まらない場合がある」、「満足のいくような計算アルゴリズムが出来ていない」、といった新たな問題が生じる．

また同じ問題に対する別のアプローチとして、小川[1992]によるベイズ統計を用いた脱退率推定の方法もある．これは粗製脱退率を作成する際に経験則を加味することで、問題点 1 の解決を図ろうとするものであるが、事前情報として「少数集団の脱退率を多数集めた集団」が必要となり、導入の負荷が大きいものと言える．

3 提案手法

本研究では、2.2 節で明らかとなった問題点を解決する、一般化加法モデル (Generalized Additive Model : GAM. Hastie and Tibshirani [1986], Wood[2017]) に基づく脱退率の推定手法を提案する．

3.1 GAM の概要

GAM は、当てはまりの良さと平滑度合いの良さを両立した推定が可能となる統計的手法であり、統計モデリングにおいて広く使用されている．現行手法との関連では、Greville 法の平滑化の理念を統計モデルの文脈で自然に拡張したものと言える．

加入者の脱退に関して、「 x : 年齢, y : 脱退なら 1, 非脱退なら 0」なる観測データ $\{(x_1, y_1), \dots, (x_n, y_n)\}$

が与えられているとする。ここで同じ年齢 x に対する観測データ y が複数存在してもよい。年齢 x の観測データ y が真の脱退率 q_x を成功確率とするベルヌーイ分布に従って発生したと仮定する。

脱退率 q_x のロジット関数 $\text{logit}(x) := \log x / (1 - x)$ による変換を滑らかな関数 $f(x)$ によって $\text{logit}(q_x) = f(x)$ とモデル化し、次のような平滑化付き尤度最大化により f を決定する。

$$\min_f \left\{ L(f(x_1), y_1, \dots, f(x_n), y_n) + \lambda \int \{f^{(m)}(x)\}^2 dx \right\}$$

ここで L は観測データのベルヌーイ分布に基づく負の対数尤度

$$\begin{aligned} & L(f(x_1), y_1, \dots, f(x_n), y_n) \\ &= - \sum_i \{ y_i \log q_{x_i} + (1 - y_i) \log(1 - q_{x_i}) \} \\ &= - \sum_i \{ y_i \log \text{logit}^{-1} f(x_i) + (1 - y_i) \log(1 - \text{logit}^{-1} f(x_i)) \} \end{aligned}$$

である。ここで、 $\text{logit}^{-1}(x) := 1/(1 + e^{-x})$ である。この最小化問題は、「乖離度」として「負の対数尤度」（つまり逸脱度）、「平滑度」として「 m 次微分平滑度」を使用した最小化問題 (2.1) と等価である。特に $m = 3$ のとき、Greville 法における 3 次差分に対応する概念であり、差分による滑らかさ制御を連続関数区間に一般化したものである。

この最適化問題の解は、命題 2-2 より、 f を自然 $2m - 1$ 次スプライン (knot は $\{x_1, \dots, x_n\}$) に限定してよく、この関数空間としての基底を 1 つ取ること、その 1 次結合による係数を求める問題に帰着され、解を数値的に求めることが可能となる。つまり（ロジット関数を用いた）GAM は、 $\text{logit}(q_x)$ を目的変数とし、基底関数 $\{B_j(x)\}$ の 1 次結合で表される自然 $2m - 1$ 次スプライン関数 $s(x)$ を説明変数に設定した回帰モデルであると言える。

注意 3-1.

ロジット関数で変換するのは、推定される脱退率が 0 と 1 の間の値を取ることを保証するためである。このようにロジット関数で変換されたものを推定対象とすることは現行手法や先行研究と異なるアプローチであるが、ベルヌーイ分布を扱う場合にロ

ジット関数を使用することは統計モデリングにおいては標準的な取扱いである。

以上をまとめて、本研究で使用する GAM は以下のモデル式で表現される：

$$\text{logit}(q_x) = s(x) \quad (3.1)$$

$$s(x) = \sum_j \beta_j B_j(x) \quad (3.2)$$

$B_j(x)$: 自然 $2m - 1$ 次スプラインの基底関数 (mgcv パッケージにおけるパラメータ bs : “cr” , “tp” など)

$\lambda > 0$ を固定したとき、 $\beta = \{\beta_j\}$ を次の最小化問題の解として定める。

$$\min_{\beta} \left\{ L(s(x_1), y_1, \dots, s(x_n), y_n) + \lambda \int \{s^{(m)}(x)\}^2 dx \right\} \quad (3.3)$$

平滑化パラメータ λ は統計的指標 (mgcv パッケージにおけるパラメータ method, 後述) を用いて自動的に選択される。

後述の通り、mgcv パッケージには、上記の計算が 1 つの関数 gam により実装されており、これらの計算を個別にコーディングする必要はないため、実務への応用可能性も高いと言える。

したがって、GAM による推定手法は、Greville 法が目指した「統計的ノイズを抑えた滑らかな脱退率推定」を、確率モデルと関数空間の枠組みにおいて体系的に実現するものであり、理論的にも実務的にもより洗練されたアプローチであると位置づけることができる。また平滑化パラメータ λ を統計的規準により自動選択されるため、恣意性が排除され、推定結果の再現性と客観性が担保される。

3.1.1 他の説明変数の導入

上記では年齢のみを説明変数としているが、性別や加入年数、外部データ脱退率といった他の変数の導入も容易である。例えば外部データ脱退率 q_x^{ex} を使用するモデルとして、(3.1) の代わりに

$$\text{logit}(q_x) = s(x) + \text{logit}(q_x^{\text{ex}}) \quad (3.4)$$

$$\text{logit}(q_x) = s(x) + \alpha \cdot q_x^{\text{ex}} \quad (3.5)$$

$$\text{logit}(q_x) = s(x) + \alpha \cdot \text{logit}(q_x^{\text{ex}}) \quad (3.6)$$

などが想定される。(3.4)は $\text{logit}(q_x^{\text{ex}})$ をオフセット項として追加するものである。一般に $\text{logit}(q_x)$ は脱退リスクの対数オッズであるから、(3.4)はオッズ比の対数を推定対象とするモデルとして解釈できる。(3.5)は q_x^{ex} を、(3.6)は $\text{logit}(q_x^{\text{ex}})$ を説明変数として追加するモデルである。なお、 $q_x^{\text{ex}} = 0, 1$ となる年齢 x が存在するとき、その年齢において $\text{logit}(q_x^{\text{ex}})$ が定義できないため、モデル(3.4)、(3.6)は使用できないことに注意が必要である。

3.1.2 ハイパーパラメータの選び方：情報量規準

GAMを用いて推定を行う場合、選択可能なパラメータ（ハイパーパラメータ）として、

- ・微分平滑度の次数 m
- ・スプラインの基底関数（例：cr, tp など）
- ・crを使用する場合の knot
- ・外部データを説明変数に加えるかの有無
- ・加える場合、使用するモデル（例：(3.4)–(3.6)）

などがあるが、統計的理論に基づいた選択を行いたい場合、各モデルに対してGAMを用いた推定を行った後、それぞれの情報量規準を比較し、最も小さいモデルを選択する、という方法がある。例えば、最も予測精度が高くなるように選びたい場合、情報量規準として、赤池情報量規準(Akaike's Information criterion: AIC)を使用し、真のモデルに近くなるように選びたい場合、ベイズ情報量規準(Bayesian information criterion: BIC)を使用することが考えられる。いずれも詳細はAppendix Cを参照。後述の通り、mgcvパッケージではAIC、BICの計算はそれぞれ関数AIC、BICにより実装されているため、容易に計算可能である。

このように統計的理論の裏付けがある数値的基準を用いてモデル選択を行うことで、ハイパーパラメータの客観的な選択が可能となる。

注意 3-2.

平滑化パラメータの選択方法(REML, GCV.Cp)もハイパーパラメータの一種であり、これをAICで判断する場合、GCV.Cpは実質的にAICが最小になるように選ぶ(Appendix Cを参照)ため、この方法では

常にGCV.Cpが選ばれることとなる。

3.1.3 平滑化パラメータ λ の選択方法

平滑化パラメータ λ の選択方法についてmgcvパッケージには、method="GCV.Cp"（既定値）、"REML"などが実装されている。前者は最も予測精度が高いもの（予測誤差基準）として、後者はベイズ的な観点から周辺尤度を最大化する平滑度を選ぶもの（周辺尤度基準）として選ぶものである。詳細はAppendix Cを参照。

本論文では、推定の安定性の観点から、method="REML"を第一に扱い、推定結果に違和感がある場合はGCV.Cpでの結果を計算し、より良いと考えられる結果を採用する、という方法を提案する。

この理論的背景としてWood[2025]には、「ML/REMLと異なり、GCV/UBRE/AICなどの予測誤差基準はときどき深刻な過剰適合を生み、信頼区間やp値の性能も悪化する傾向にある」、「ML/REMLは他の基準と比べて局所最小値に陥る可能性が低いため、より好ましい選択肢となることがある。」、「ML/REMLを用いると、シミュレーションにおいて最良の結果が得られる。これは、これらの方法が時折発生する深刻な過小推定を回避する傾向があるためである。」などと述べられている。実際、これらは第4章の数値実験の結果と整合しており、GCV.Cpの推定結果に深刻な過剰適合を生んでいるものが確認され、REMLではそのような結果が無いことが確認された。なお、これはスプラインの基底関数としてcr(knotは全年齢)を用いる場合のみであり、tpを用いる場合にはGCV.Cpであっても確認されなかった。

また、GCV.Cpと比べてREMLは大きめの λ を選ぶ（つまり有効自由度EDF（後述）が低めに推定される）傾向にあるとされている。実際、これは第4章の数値実験によって裏付けられており、REMLで推定された脱退率曲線は年齢ごとの細かな実績に追従しにくく安定して真の脱退率に近い形状が得られた。一方で、追従しないがゆえ、GCV.Cpと比べて推定結果は局所的な起伏がほとんど見られない単純な形状になる傾向がある。これは現行の実務では、何らかの加工を加えないとほとんど実現されない

形状である。

これらのことから、GCV.Cp と比べて REML の方が理論的にも数値実験の結果からも良いと判断される。

3.1.4 スプラインの基底関数、 m の選び方

本論文では、GAM で使用するスプラインの基底関数として、tp(微分平滑度の次数 $m = 2, 3$) を第一に扱い、更に良いモデルを探したい場合は、tp($m \neq 2, 3$) や全実現値を knot とする cr でも当てはめて、AIC や BIC が最も小さくなるモデルを選択する、といった方法を提案する。このように提案するのは、第 4 章で述べる数値実験において、tp($m = 2, 3$ の場合) が良い結果を示したためである。

m の選び方として、mgcv パッケージにおける既定値である $m = 2$ とすることは標準的であり、また 3.5.1 節で確認する通り、人数規模が小さく無情報に近い場合の推定結果は (2.5) から $m - 1$ 次式 (の逆ロジット変換) に近づくが、この場合の推定結果を 1 次式(直線)とすることは自然な考え方である。一方、Greville の方法との連続性から $m = 3$ とすることも想定される。

tp と cr の比較に関する理論的背景として、Wood[2025] には、「一般的に、tp は平均二乗誤差(MSE) が最良になる傾向があるが、他の基底関数より時間がかかる。cr は MSE で次点となり、p-スプライン(ps) は更にわずかに劣る。」との説明があり、それゆえ gam 関数において、引数 bs の既定値は tp となっていることから、cr より tp の方が望ましいと考えられる。なお、「時間がかかる」とあるが、本論文での状況では実行時間は問題にならない。実際、スプラインに使用する説明変数は年齢の 1 個のみであり、これは 40 個程度の値しか取らない離散変数である。実際、gam 関数の実行時間は 1 秒未満である場合がほとんどである。

更に tp は knot を指定する必要がない点も cr と比べると実務に導入しやすいと言える。cr (knot は x 座標の全観測値) の場合は、全観測値による基底関数をそのまま使用するため、 λ を固定したときの最小化問題 (3.3) の厳密解を得ることはできるが、

関数空間としての自由度が高く、 λ を選択する際に GCV.Cp がうまく機能せずに、結果的に過剰適合となる恐れがあり、実際、第 4 章の数値実験において過剰適合した推定結果が確認されている。これに対して、tp は $(1, x, \dots, x^{m-1})$ を除いた基底関数に対して低ランク近似 (low rank approximation) を行い、主要な変動成分のみを上位 $10 - m$ 個自動で選択することから、適度に関数空間を制限しており、GCV.Cp であってもうまく機能し、安定的かつ頑健な推定結果を得ることができる。ここで、mgcv パッケージにおいて基底関数を何個選択するかを表すパラメータ k (既定値は 10) が関数 s の引数に用意されているが、通常は指定せず既定値をそのまま使用する。

なお、tp($m = 2$ の場合) と cr (knot は x 座標の全実現値) について、両者が表現できる関数空間は理論上は同一のものであるが、基底関数の取り方が異なるため、mgcv パッケージにおける両者の出力結果は一般に異なる。

3.2 R を用いた実装

統計ソフトウェア R における mgcv パッケージを用いることで、GAM を容易に実装できる。以下、本章において GAM の結果を示す場合、特に言及しない場合は `bs = "tp" ( $m = 2$ )` かつ `method = "REML"` を使用し、外部データ脱退率のモデルの結果を示す場合は (3.4)、つまり `logit( $q_x^{\text{ex}}$ )` をオフセット項として追加するモデルを前提とする。

3.2.1 R コード

脱退率の推定結果は以下のように得られる。

```
library(mgcv)
# data_original:加工前データ (脱退実績データ+ex)
# df:加工後データ (1レコードずつに分解)

# モデル(3.1)
model_base <- gam(
  y ~ s(x), data = df, family = binomial, method = "REML"
)
# モデル(3.4)
```

```

model_logit_ex_offset <- gam(
  y ~ offset(qlogis(ex)) + s(x),
  data = df, family = binomial, method = "REML"
)
# 予測用データの作成
all_ages <- seq(min(data_original$x, na.rm = TRUE),
  max(data_original$x, na.rm = TRUE))
newdata <- data.frame(
  x = all_ages,
  ex = data_original$qex[match(all_ages, data_original$x)]
)
# 予測値の計算
pred_base <- predict(
  model_base, newdata = newdata, type = "response"
)
pred_logit_ex_offset <- predict(
  model_logit_ex_offset,
  newdata = newdata, type = "response"
)

```

3.2.2 推定に使用するデータ

まず過去3年間の脱退実績データをデータフレーム `data_original` に格納する (図 3-1 でいう `x`: 年齢, `l`: 加入者数, `d`: 脱退者数). これは現行の手法で粗製脱退率の作成の元となっているものであるが, もし外部データ脱退率を推定に用いる場合, このデータも含めておく (図 3-1 でいう `qex`). なお, 加入者数が 0 人の年齢があっても特段の対応は不要である (図 3-1 でいう 23 歳).

> data_original				
	x	l	d	qex
1	19	5	1	0.26690124
2	20	2	1	0.22112131
3	21	3	1	0.18546692
4	22	4	1	0.15993808
5	23	0	0	0.14453478
6	24	1	1	0.13925702
7	25	1	1	0.14410481
8	26	2	0	0.14572831
9	27	10	3	0.14412751
10	28	30	2	0.13930242
11	29	20	10	0.13125304
12	30	18	10	0.11997936
13	31	40	21	0.10977181
14	32	20	10	0.10063038
15	33	8	4	0.09255509
16	34	20	0	0.08554592
17	35	2	1	0.07960289
18	36	18	0	0.07447700
19	37	17	0	0.07016825
20	38	16	1	0.06667666
21	39	17	1	0.06400221
22	40	1	1	0.06214491
23	41	2	1	0.05989284
24	42	18	2	0.05724601
25	43	16	3	0.05420442
26	44	15	1	0.05076807
27	45	14	0	0.04693695
28	46	13	1	0.04360614
29	47	12	1	0.04077562
30	48	11	0	0.03844541
31	49	12	0	0.03661549
32	50	13	0	0.03528588
33	51	12	1	0.03444621
34	52	13	0	0.03409650
35	53	14	0	0.03423675
36	54	15	0	0.03486695
37	55	12	0	0.03598710
38	56	14	1	0.03734942
39	57	6	0	0.03895392
40	58	8	0	0.04080060
41	59	4	1	0.04288944

図 3-1 脱退実績データ (過去3年間, `data_original`)

次にこのデータを加入者ごとに脱退なら 1, 非脱退なら 0 となるレコードに変換する. つまり加入者の数分だけレコードが作成される (図 3-2).

```

# df に変換
df <- do.call(
  rbind, lapply(
    seq_len(nrow(data_original)),
    function(i) {
      age <- data_original$x[i]
      count <- data_original$l[i]
      churn <- data_original$d[i]
      qex <- data_original$qex[i]
      if (count == 0) return(NULL)
      data.frame(
        x = rep(age, count),
        y = c(rep(1, churn), rep(0, count - churn)),
        ex = rep(qex, count)
      )
    }
  )
)

```

```
> head(df, 20)
  x y ex
1 19 1 0.2669012
2 19 0 0.2669012
3 19 0 0.2669012
4 19 0 0.2669012
5 19 0 0.2669012
6 20 1 0.2211213
7 20 0 0.2211213
8 21 1 0.1854669
9 21 0 0.1854669
10 21 0 0.1854669
11 22 1 0.1599381
12 22 0 0.1599381
13 22 0 0.1599381
14 22 0 0.1599381
15 24 1 0.1392570
16 25 1 0.1441048
17 26 0 0.1457283
18 26 0 0.1457283
19 27 1 0.1441275
20 27 1 0.1441275

> tail(df, 20)
  x y ex
460 56 0 0.03734942
461 56 0 0.03734942
462 57 0 0.03895392
463 57 0 0.03895392
464 57 0 0.03895392
465 57 0 0.03895392
466 57 0 0.03895392
467 57 0 0.03895392
468 58 0 0.04080060
469 58 0 0.04080060
470 58 0 0.04080060
471 58 0 0.04080060
472 58 0 0.04080060
473 58 0 0.04080060
474 58 0 0.04080060
475 58 0 0.04080060
476 59 1 0.04288944
477 59 0 0.04288944
478 59 0 0.04288944
479 59 0 0.04288944
```

図 3-2 各加入者の脱退実績データ (df)

3.2.3 実行例

この脱退データに対する実行結果を図 3-3 に示す。

```
> pred_base
  1      2      3      4      5      6
0.27790218 0.30846742 0.32471876 0.31304181 0.27589761 0.23220001
  7      8      9     10     11     12
0.19976263 0.19135149 0.21560168 0.27896396 0.37424367 0.46395363
 13     14     15     16     17     18
0.49846272 0.44853964 0.32118958 0.17956948 0.08903722 0.04947107
 19     20     21     22     23     24
0.03717332 0.04010188 0.05748380 0.09184837 0.13455065 0.15840630
 25     26     27     28     29     30
0.14340949 0.10222043 0.06280655 0.03827989 0.02626722 0.02146370
 31     32     33     34     35     36
0.02033320 0.02057560 0.02037114 0.01883637 0.01659785 0.01510923
 37     38     39     40     41
0.01580000 0.02060332 0.03406248 0.06700238 0.13885323
```

```
> pred_logit_ex_offset
  1      2      3      4      5      6
0.29466155 0.30880926 0.31098938 0.28946058 0.24997693 0.21304038
  7      8      9     10     11     12
0.19483436 0.19235293 0.21867086 0.28238306 0.37636469 0.46155275
 13     14     15     16     17     18
0.49611220 0.44861387 0.32333024 0.18144535 0.08965482 0.04921516
 19     20     21     22     23     24
0.03645897 0.03899293 0.05606600 0.09116980 0.13524031 0.16031225
 25     26     27     28     29     30
0.14526035 0.10282908 0.06207188 0.03741758 0.02559993 0.02100238
 31     32     33     34     35     36
0.02005342 0.02045292 0.02033888 0.01879968 0.01650018 0.01495684
 37     38     39     40     41
0.01563733 0.02040808 0.03390190 0.06731588 0.14118798
```

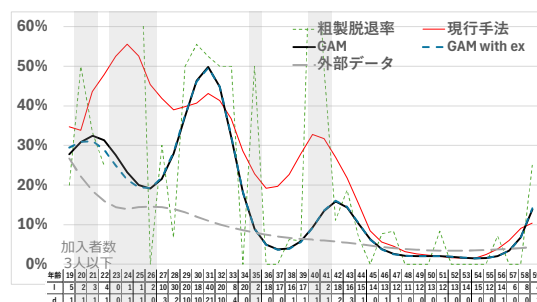


図 3-3 GAM の実行例と現行手法との比較

この脱退データは 20 歳台後半～30 歳台の加入者数が極端に多いなど、年齢集団ごとに人数がバラバラであり、各年齢での加入者数に応じた脱退率推定を行うのが望ましいデータである。

現行手法（5 点移動平均法を 2 回適用）を機械的に当てはめた結果を細い実線で表した。例えば 23 歳～26 歳や 40 歳～41 歳について、加入者の少ない年齢の極端に多い脱退実績に引きずられて高い値となっており、また 30 歳～32 歳は、加入者の少ない周りの年齢での極端な実績の影響を受け、加入者が多い年齢での脱退実績が十分に反映できておらず低めの水準となっている等、不十分な結果となっている。

これに対して GAM(3.1)（太い実線）、GAM(3.4)（破線、外部データ脱退率をオフセット項として取り入れる）の結果を見ると、現行手法の欠点についても妥当な結果となっており、また 28 歳～32 歳や 37 歳～39 歳について、加入者の多い年齢での脱退実績を反映できているなど、現行手法よりも望ましい結果となっていることが分かる。

3.2.4 複数のモデルの比較

3.1.2 節で述べたパラメータの選択方法として、AIC を用いたモデル選択例を示す。

```
library(mgcv)

# data_original:加工前データ（脱退実績データ+ex）
# df:加工後データ（1レコードずつに分解）
# all_ages:年齢の範囲
# newdata:予測用データ
```

```

# モデルの構築
knots <- unique(data_original$x[data_original$l > 0])
models <- list(
  base = gam(y ~ s(x),
    data = df, family = binomial, method="REML"),
  m3 = gam(y ~ s(x, m = 3),
    data = df, family = binomial, method="REML"),
  GCV = gam(y ~ s(x), data = df, family = binomial),
  cr = gam(y ~ s(x, bs = "cr", k = length(knots)),
    data = df, family = binomial, method="REML"),
  logit_ex_offset = gam(y ~ offset(qlogis(ex)) + s(x),
    data = df, family = binomial, method="REML"),
  additive_ex = gam(y ~ ex + s(x),
    data = df, family = binomial, method="REML"),
  logit_ex = gam(y ~ I(qlogis(ex)) + s(x), data = df,
    family = binomial, method="REML")
)
# 予測値の計算
predictions <- lapply(models, predict, newdata =
  newdata, type = "response")
# AIC の表示
aic_values <- sapply(models, AIC)
print(aic_values)

```

```

> print(aic_values)
      base      m3      GCV      cr
logit_ex_offset 348.0569 350.4529 345.1002 346.2696
      additive_ex      logit_ex
logit_ex_offset 348.7206 347.2070 346.7306

```

図 3-4 AIC の実行例

図 3-4 において、外部データ脱退率を組み込むべきか、またその組み込み方について、モデル base (3.1), logit_ex_offset (3.4), additive_ex (3.5), logit_ex (3.6) に対する AIC を比較すると、logit_ex、つまり (3.6) が最小となった。よってこの例の場合、予測精度の観点からは $\text{logit}(q_x^{\text{ex}})$ を説明変数として追加するモデル (3.6) が最良であることが分かる。

また、基底関数として $\text{tp}(m=2,3)$, cr (knot は全年齢) のいずれを使用すべきかについて、モデル base, m3, cr の AIC を比較すると、cr の AIC が最小となった、よってこの例では、予測精度の観点からは cr を選択することが最良であることが分かる。

なお、平滑化パラメータの選択方法 REML, GCV について、モデル base, GCV の AIC を比較すると GCVの方が小さくなっており、これは注意 2-1 で確認した事実と整合している。

3.3 現行の問題点の解決策

GAM は現行手法の持つ 5 つの問題点 (2.2 節) を全て克服している。

まず問題点 2 に対して、ロジット関数を用いた GAM による推定では、推定結果は必ず 0 より大きく 1 より小さいため、問題点 2 は生じない。

次に、GAM による提案手法は粗製脱退率を作成しないため、問題点 3 のような粗製脱退率の作成に関わる問題は生じない。

更に、適切な外部データ脱退率を取得できた場合、GAM ではその外部データ脱退率を推定に組み込むことが可能である。また、外部データ脱退率を組み込むべきかを判断する際の一つの目安として、情報量規準といった統計的根拠に基づいた外部データの取入れが可能となっており、問題点 4 を克服していると言える。

問題点 1 について、GAM では乖離度として負の対数尤度を使用しているため、加入者数が多い年齢ではその脱退実績分だけ尤度に反映されることとなる。つまり各年齢における実績数を自動的に考慮した尤度最大化に基づく推定を行っており、問題点 1 も克服している。

問題点 1 に対して、2.2.1 節の表 2-1 の例について、それぞれ GAM に当てはめると、図 3-5、3-6 のような結果となった。

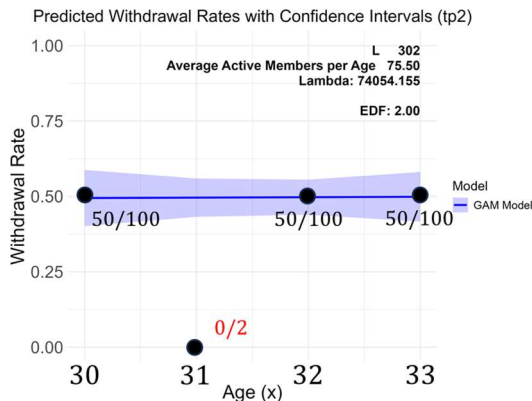


図 3-5 ケース 1 の推定結果

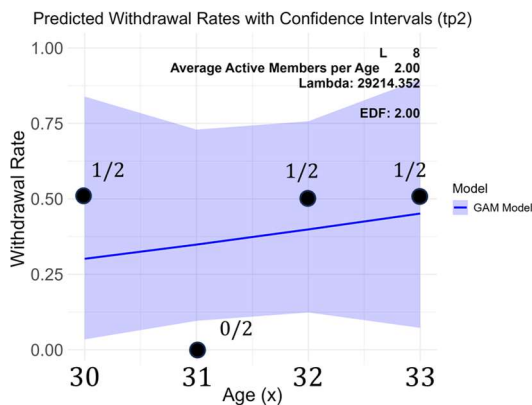


図 3-6 ケース 2 の推定結果

ケース 1 の推定結果（実線）では 31 歳の実績はほぼ無視されているが、ケース 2 では全ての年齢の実績を同列に扱っている。更に図中の 95%信頼区間（後述）の帯の幅に注目すると、ケース 1 の方が小さくなっている。この例からも本提案手法は実績数の違いを考慮できる手法であることが分かる。

3.4 有効自由度

GAM の推定結果を見る上で、重要な指標である有効自由度（Effective Freedom of Degree ; EDF）について、一般的に知られている事実を簡単に述べる。（詳細はHastie, Tibshirani and Friedman[2009], Wood[2017]を参照）

EDF は推定結果の実質的なパラメータの数（モデルの複雑さの程度）を表すものとして、構築されたモデルに基づいて計算される平滑化行列とよばれ

るもののトレースとして定義される。一般に 1 次元データの場合には、平滑化パラメータ λ 、平滑度の次数 m 、使用する基底関数の個数 k 、観測データの（重複を除いた）年齢 x の個数 J 、観測データや基底関数に依存する正の定数 δ_j を用いて、

$$\text{EDF} = m + \sum_{j=1}^{\min(k,J)-m} \frac{1}{1 + \lambda \delta_j}$$

と表せることが知られている。ここから、

$$m \leq \text{EDF} \leq \min(k, J) \quad (3.7)$$

であり、 $\lambda \rightarrow +0$ 、つまり乖離度が重視された推定結果（視覚的には粗製脱退率に近い）の場合には EDF は上限値である $\min(k, J)$ に近づき、 $\lambda \rightarrow +\infty$ 、つまり平滑度が重視された推定結果（視覚的にはリンクスケールで $m-1$ 次多項式）の場合には EDF は下限値である m に近づく。

特に全年齢を knot とする cr を採用した場合、 $k = J$ であることから、

$$2 \leq \text{EDF} \leq J \quad (3.8)$$

となり、tp を採用した場合、 $k = 10$ （既定値）、 J は通常 10 を上回ることから、

$$m \leq \text{EDF} \leq 10 \quad (3.9)$$

となる。

mgcv パッケージにおいて、EDF の計算結果は以下のコードで得られる。注意として、mgcv パッケージで得られる EDF は定数関数分の 1 次元が切り出されるため、ここでは上記の理論に沿うよう 1 を加えた値に補整している（図 3-7）。

```
# model_base:関数 gam の実行結果
edf <- summary(model_base)$s.table[1, "edf"] + 1
> edf
[1] 8.471427
```

図 3-7 EDF の実行例

3.5 提案手法のその他の特徴

3.5.1 人数規模を踏まえた推定

脱退率推定において人数規模を踏まえた推定が望ましい。つまり、人数規模が大きい場合は粗製脱退率に近づき、人数規模が小さい場合は情報が乏しく、例えば真の脱退率として想定される関数形（例

えばリンクスケールで $m - 1$ 次式), 外部データ脱退率を考慮する場合は外部データ脱退率に近い形状に近づくことが期待されるが, 本手法はこれを満たす.

例えば同じ粗製脱退率を持つ人員データに対し, 人数規模を小 (各年齢 3 人程度), 中 (各年齢 15 人程度), 大 (各年齢 50 人程度) の 3 パターンで GAM (モデル (3.1), (3.4)) による当てはめを行った. この結果を図 3-8~3-10 に示す. 図において粗製脱退率を大きめの点で表現した.

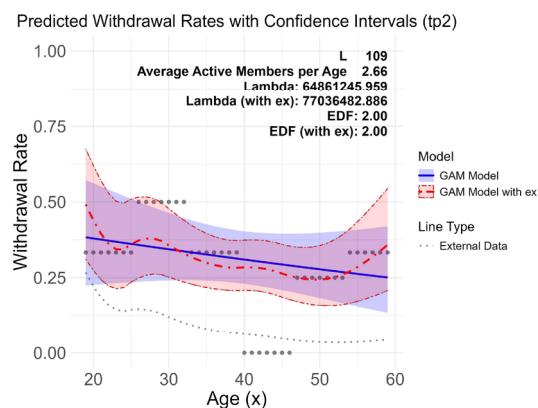


図 3-8 小規模集団の場合の推定結果 (tp, $m = 2$)

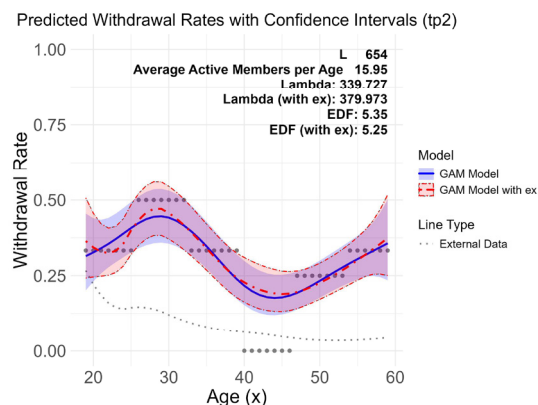


図 3-9 中規模集団の場合の推定結果 (tp, $m = 2$)

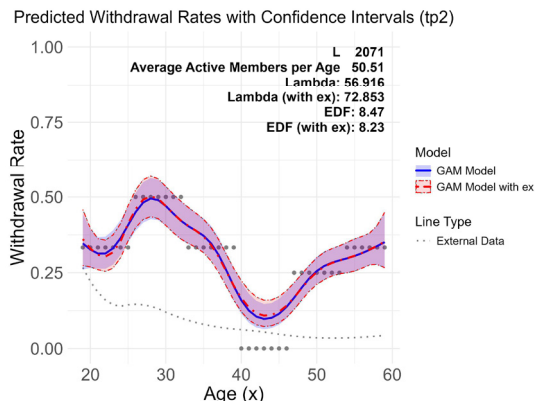


図 3-10 大規模集団の場合の推定結果 (tp, $m = 2$)

確かに小規模集団 (図 3-8) においては, 推定結果は直線, または外部データ脱退率に近い形となり, 規模が増加するにつれて粗製脱退率に近づいていく (EDF が増大していく) 様子が確認される.

これは最小化問題 (3.3) において, λ を固定すると, データの数を増やすと乖離度 (逸脱度) が増加していく一方で, 平滑度が一定値以上は大きくなることから, 最小化問題 (3.3) において相対的に乖離度 (逸脱度) の方が重視されるためである. 逆に人数規模が小さいと, 相対的に平滑度の方が重視されるため, (2.5) より推定されるスプライン関数は $m - 1$ 次式 ($m = 2$ の場合は直線) に近づく傾向にある.

なお, (knot を全年齢とする) cr の場合はより粗製脱退率に近づく. これは EDF が 10 (tp の場合の上限値) を超えていることから分かる (図 3-11, 縦の点線は使用した knot を表す).

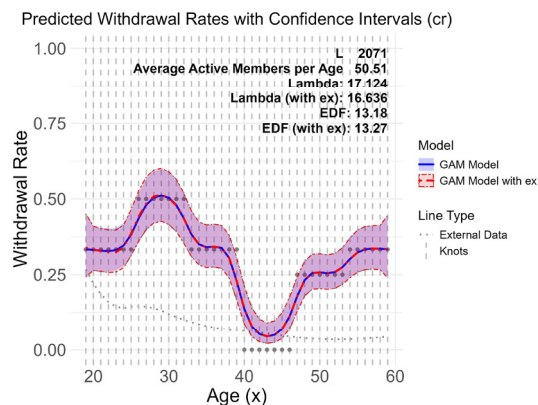


図 3-11 大規模集団の場合の推定結果 (cr)

3.5.2 補外

ある年齢以上の加入実績が無い場合、GAM による推定結果をそのまま加入実績が無い年齢まで伸ばして使用した場合、自然 $2m - 1$ 次スプラインの定義から、補外部分のスプライン関数は $m - 1$ 次式となる。例えば $m = 1$ の場合は 0 次式、つまり直前の年齢での脱退率をそのまま使用することになり、 $m = 2$ の場合は 1 次式、つまり直前の年齢での傾向をそのまま直線的に伸ばした推定結果となる。 $m = 3$ の場合は 2 次式となる。

図 3-12～3-14 は $tp(m = 1 \sim 3)$ による補外部分 (40 歳～59 歳) の違いを表したものである。

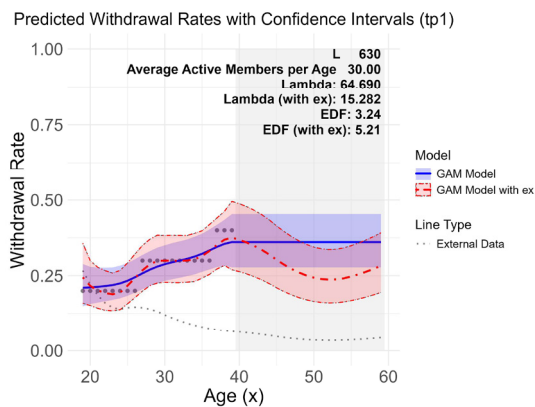


図 3-12 $tp(m = 1)$ の推定結果

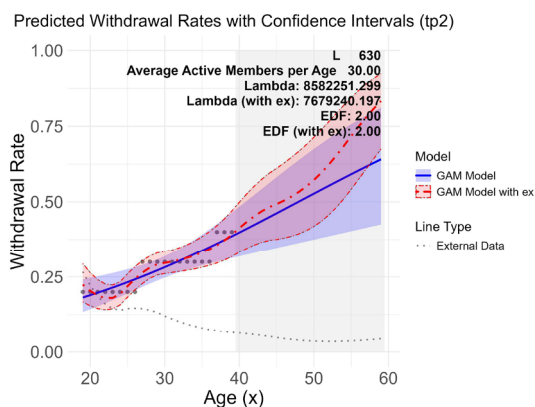


図 3-13 $tp(m = 2)$ の推定結果

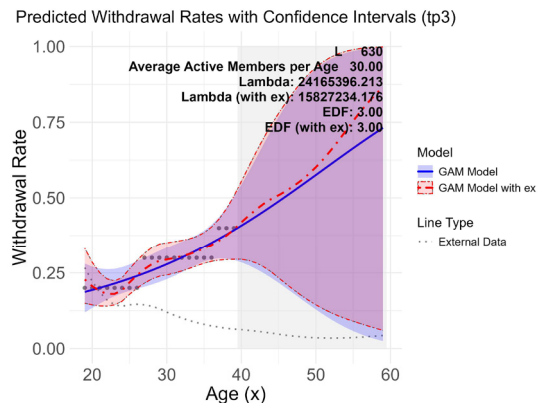


図 3-14 $tp(m = 3)$ の推定結果

3.5.3 極端な値を取らない

$\text{logit}(q_x)$ を推定対象としているため、 q_x は関数 logit^{-1} の値域、つまり $(0, 1)$ の範囲で値を取る。したがって、提案手法による脱退率の推定結果は値 0 や 1 を取らない。

3.5.4 特異な年齢における脱退率の補整

例えば定年延長を行った DB について、旧定年年齢における脱退が他の年齢よりも特異的に多く見込まれる場合がある。このとき旧定年年齢での脱退率のみ平滑化せずに、粗製脱退率を使用する（いわゆるコブを作る）形で脱退率推定することがある。この取扱いは「DB 実務ガイドンス」に「特異な年齢における補整方法」として規定されている。このような異常脱退の判定は恣意性を伴うものであり、客観性を持たせることが困難な意思決定と言える。

本提案手法をそのまま適用した場合、（特異な年齢における異常な粗製脱退率が周辺年齢と比べて統計的に有意な差として識別できるだけの加入者および脱退者データが存在しない限り、）平滑化の過程により、その年齢での推定値は周りの年齢との間でならされた推定結果となってしまう。つまり、本提案手法は真の脱退率曲線が年齢に関して滑らかな曲線である場合に有効であり、このように特異な年齢がある場合には、一般にそのまま使用することは適切でない。

なお、本提案手法を用いて「DB 実務ガイドンス」に準拠する形で特異な年齢における補整を行う場

合、例えば特異な年齢を除いて提案手法を適用して脱退率を推定し、特異な年齢においては粗製脱退率をそのまま使用する、といった方法が考えられる。

また、信頼区間（後述）に含まれない粗製脱退率を異常脱退データと判断することで、客観的な意思決定が可能になる可能性がある。

3.6 現行手法との比較

現行手法と提案手法の比較を表にまとめると表 3-1 の通りである。上 5 つは 2.2.1 節で明らかにした問題点である。

表 3-1 現行手法と提案手法の比較

	現行 GAM
各年齢の粗製脱退率の値の違いを考慮している	× ○
補正結果が必ず 0 から 1 までの実数になる	× ○
粗製脱退率作成に関わる問題点を抱えていない	× ○
外部データ脱退率 q_x^{ex} を推定に取り入れられる	× ○
脱退率に関する信頼区間が得られる	× ○
決定脱退率は粗製脱退率からかけ離れない	○ ○
決定脱退率は全体に滑らかな形態をしている	○ ○
集団の規模が大きくなるにつれ、	
決定脱退率は粗製脱退率に近くなる	× ○
大集団だけでなく小集団でも有効な推定方法である	× ○
計算結果がシステムを用いて容易に算出される	○ ○
計算時間は実務上問題にならない	○ ○
計算に乱数を用いない	○ ○
推定結果はノイズに引きずられず、その予測精度は高い	× ○
高度な理論を使用している（その分、説明が難しい）	※ ○

※ 移動平均法: ×, Greville: △

4 数値実験

前章までで提案手法の優位性は理論的に確かめられたが、実際に理論通りになっているかを検証するために数値実験を行う。また、GAM に使用するスプライン関数(bs)として、cr や tp（微分平滑度の次数 m ）、平滑化パラメータ λ の選択方法(method)として GCV, Cp や REML などの選択肢があるが、それぞれの手法がどのような推定結果となるかなど、各

手法の特徴を明らかにする。

具体的には、真の脱退率を仮定し、それに従う脱退データを乱数を用いて多数発生させ、その脱退データを用いて脱退率推定を数多く行い、現行手法と比較して真の脱退率に近い推定結果が得られるかを検証する。

4.1 数値実験の概要

数値実験に使用する加入者集団 $\{l_x\}$ として、小規模（340 人程度）、大規模（3,400 人程度）の 2 種類を想定し、各年齢の人数割合は同一とする。（図 4-1）

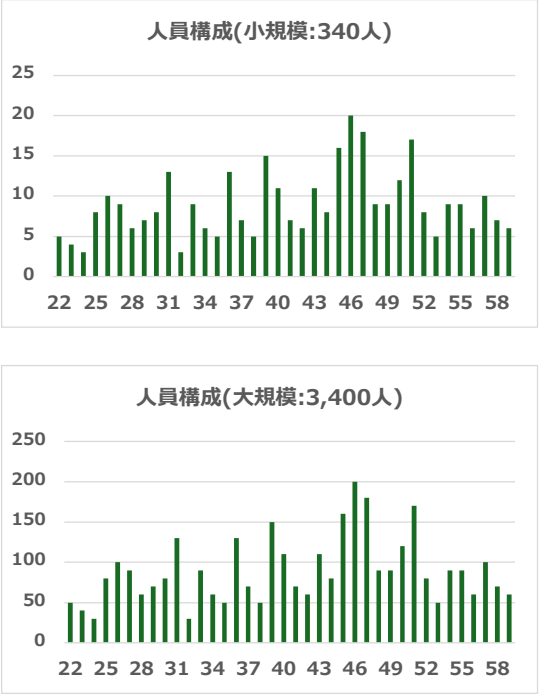


図 4-1 数値実験に使用する人員構成

真の脱退率として、次の 2 種類 q_{01} , q_{02} を想定する。どちらも年齢ごとに滑らかなものであるが、 q_{02} の方が q_{01} と比べてより凹凸が大きいものを想定している。（図 4-2）

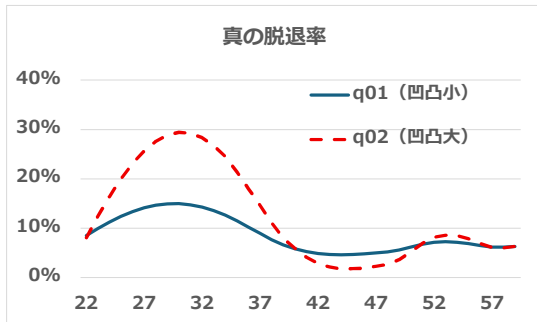


図 4-2 数値実験に使用する「真の脱退率」

x 歳の加入者数 l_x と真の脱退率 q_x に対して、試行回数 l_x 、成功確率 q_x の二項分布に従う確率変数の実績値（つまり年齢 x での実績脱退者数） d_x を二項乱数をもとに 300 個作成し、3 個の脱退実績を用いて 1 個の脱退率を推定する。つまり 100 個の脱退率が得られることとなる。「3 個の脱退実績をまとめて 1 個の脱退率を推定する」という考え方は、3 年実績で脱退率を推定する実務の取扱いを踏襲したものであり、加入者集団 $\{l_x\}$ から次の 1 年間に発生する脱退シナリオを 300 通り想定していることとなる。なお、実務との取扱いに完全に揃えるなら、3 年間の脱退実績を年度ごとに追跡すべきだが、ここでは簡単のため、全て同一の加入者集団 $\{l_x\}$ から 3 年分の脱退が発生するものとみなしている。

脱退率推定には次のモデルを使用する。

- ① Greville(3 次 13 項)【現行手法】
- ② GAM : bs= "tp" ($m = 2, 3$)【提案手法】
- ③ GAM : bs= "cr" (knot : 全年齢), bs= "tp" ($m = 1, 4 \sim 8$)【比較用】

ここで tp の微分平滑度の次数 m として幅広く想定するため、1~8 とする。②③については、method= "GCV. Cp", "REML" の両方を想定する。①について、算定年齢幅の拡大は行わない。

これら 100 個の推定脱退率から次を検証する。

- ・検証 1 : モデル①~③のうち推定脱退率が真の脱退率に近いものはどれか。
- ・検証 2 : 視覚的にみて、モデル②, ③(cr のみ)の推定結果は違和感のあるものになっていないか。

ここでは真の脱退率 $\{q_x\}$ と推定脱退率 $\{\hat{q}_x\}$ との近さを測る指標としてクロスエントロピー (Cross Entropy : CE, Appendix D を参照) を用いる。これは 2 値分類における損失関数として一般的なものである。

$$CE := -\frac{1}{x_r - x_e} \sum_{x=x_e}^{x_r-1} \{q_x \log \hat{q}_x + (1 - q_x) \log(1 - \hat{q}_x)\}$$

ここで $(x_e, x_r) = (\text{最低年齢}, \text{最高年齢}) = (22, 60)$ とする。クロスエントロピー CE が小さいほど、推定脱退率 $\{\hat{q}_x\}$ は真の脱退率 $\{q_x\}$ に近いと言える。

4.2 検証 1 (数値的検証)

推定脱退率 $\{\hat{q}_x\}$ と真の脱退率 $\{q_x\}$ との CE が 100 個得られるが、この分布について調べる。

4.2.1 CE の 100 回平均、標準偏差

次の図 4-3~4-6 は左の縦軸に CE の平均値(実線), 右の縦軸に標準偏差 (一点鎖線) を取ったグラフである。

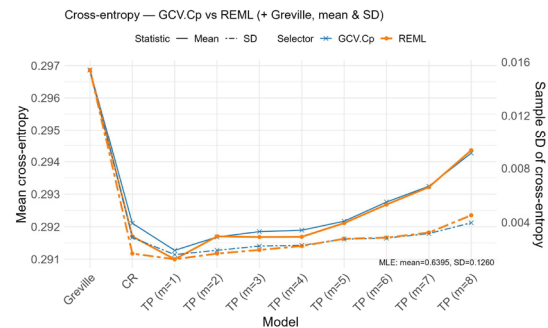


図 4-3 CE の平均・標準偏差 (q01, 小規模)

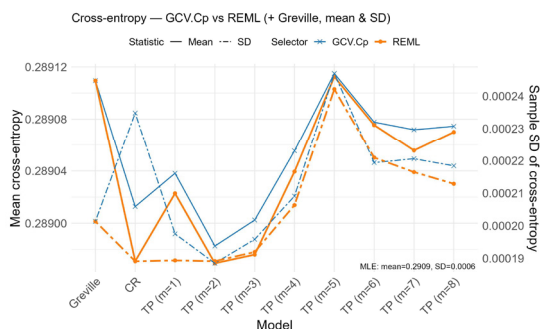


図 4-4 CE の平均・標準偏差 (q01, 大規模)

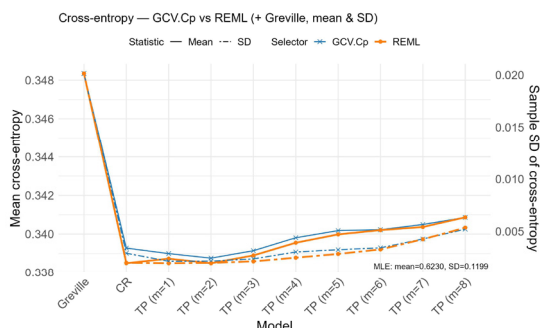


図 4-5 CE の平均・標準偏差 (q02, 小規模)

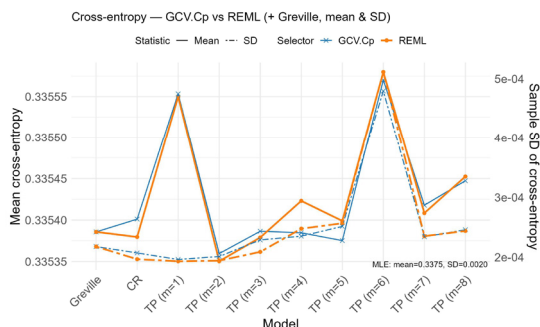


図 4-6 CE の平均・標準偏差 (q02, 大規模)

他のモデルと比較して、 $tp(m=2,3)$ は平均的に CE が小さく、そのブレも小さいことから、安定して真の脱退率に近い結果となることが確認される。また現行手法である Greville と比べると、特に小規模集団において精度を大きく改善できていることが分かる。他の $tp(m \neq 2,3)$ は CE の平均もブレも相対的に大きく、 $tp(m=2,3)$ と比べて優れてい

るとは言えない。

また GCV.Cp と REML の結果を比較すると、 $cr, tp(m=2,3)$ は全ての場合で REML の方が良い結果を示しており、特に cr の場合は改善の幅が非常に大きいことが分かる。

4.2.2 CE の violin plot

次に、各モデルでの CE の分布の形状を確認し、大きく外れている推定結果がどの程度存在するかを確認する。図 4-7～図 4-10 にその結果を violin plot として示した。なお、小規模集団において Greville の CE の最大値は他のモデルより著しく大きいいため、グラフから除外した。

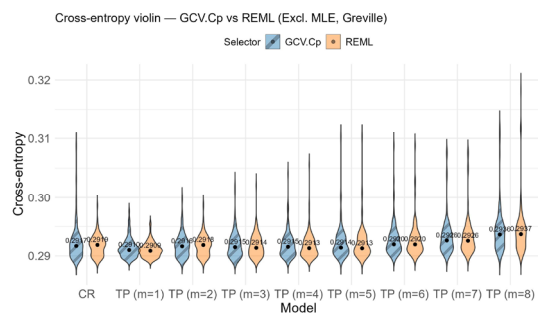


図 4-7 CE の violin plot (q01, 小規模)

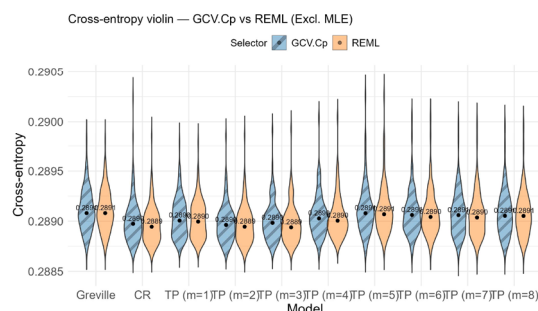


図 4-8 CE の violin plot (q01, 大規模)

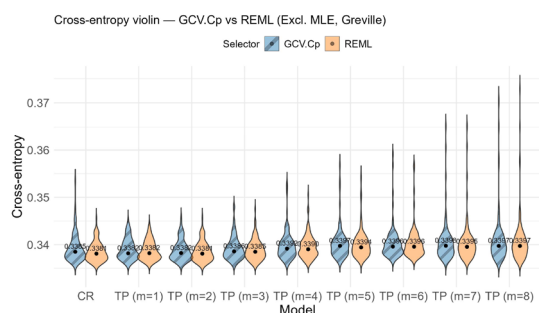


図 4-9 CE の violin plot (q02, 小規模)

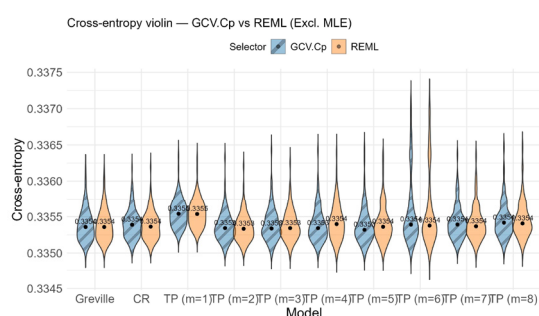


図 4-10 CE の violin plot (q02, 大規模)

他のモデルと比較して、 $tp(m=2,3)$ は CE の最大値も大きくなく、他のモデルよりも安定した結果となることが分かった。また GCV.Cp と REML を比較すると、わずかに REML が良い結果を示していることが分かる。一方、cr かつ GCV.CP の場合、特に小規模集団において CE の最大値は大きく、これが REML の場合に大きく改善されていることが分かる。これについて次節で詳しく見る。

4.3 検証 2 (視覚的検証)

検証 1 では、提案手法 (モデル②) が真の脱退率に近い推定結果を与えることが明らかとなった。次の検証として、この推定結果が視覚的に違和感のあるものでないかを調査する。ここでは全シミュレーションパスを重ねた軌跡図を用いる。

真の脱退率 q_{01} , 小規模集団の場合の軌跡図は以下ようになった。

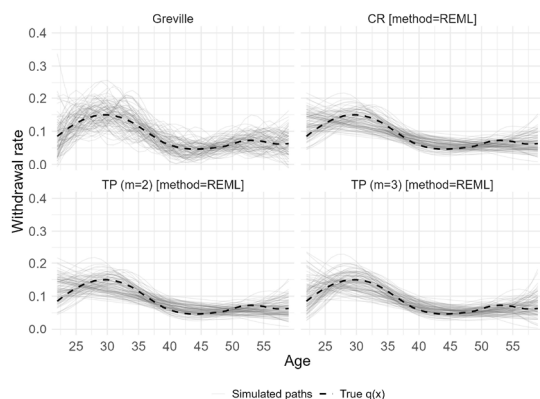


図 4-11 REML の場合の軌跡図 (q_{01} , 小規模)

次に, 大規模集団の場合の軌跡図を図示する。

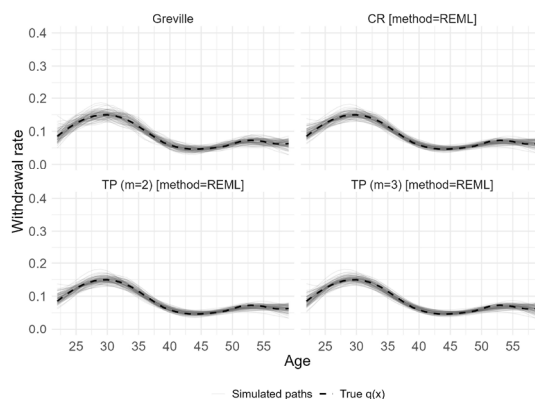


図 4-13 REML の場合の軌跡図 (q_{01} , 大規模)

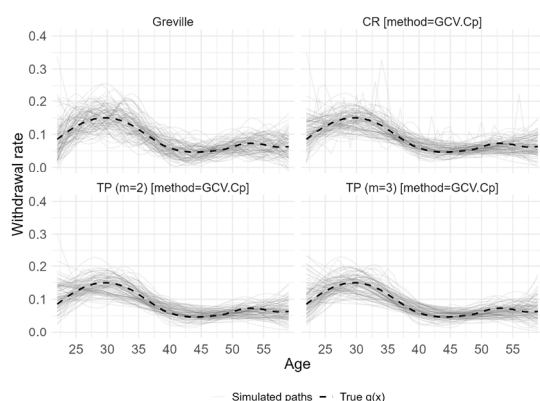


図 4-12 GCV.Cp の場合の軌跡図 (q_{01} , 小規模)

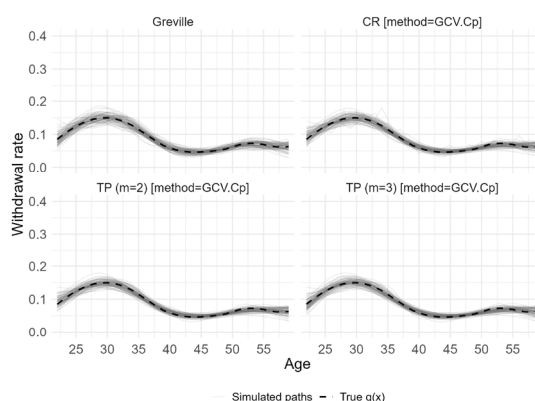


図 4-14 GCV.Cp の場合の軌跡図 (q_{01} , 大規模)

図 4-11 を見ると, 現行手法(Greville ; 左上)はややバラツキが大きく, 提案手法の方が平均的に真の脱退率に近い形状となっていることが分かる。

一方で, cr , $tp(m=2,3)$ の一部の推定結果が大域的に凹凸に変化が無い(例えば 1 次式, 2 次式の逆ロジット変換)ことが確認され, REML の方がこのような推定結果をやや多く含むように見て取れる。

また, 図 4-12 の cr の結果には, 滑らかでない推定結果が一部含まれているが, 図 4-11 にはそのような結果が見られないことが分かる。

大規模集団の場合は, 現行手法でも真の脱退率に近い形状となっているが, 提案手法の方がよりブレが抑えられていることが分かる。

図 4-13 を見ると, 小規模の場合に確認されたような, 凹凸に変化の無い推定結果は確認されず, ほぼ全ての推定結果で真の脱退率の形状をうまく捉えられている。一方, 図 4-14 の cr の結果を見ると, 小規模の場合と同様, 滑らかでない推定結果が依然確認され, 図 4-13 にはそのような結果が見られないことが分かる。 $tp(m=2,3)$ については両者に大きな違いは無い。

次は真の脱退率 q_{02} ，小規模集団の場合の軌跡図である。

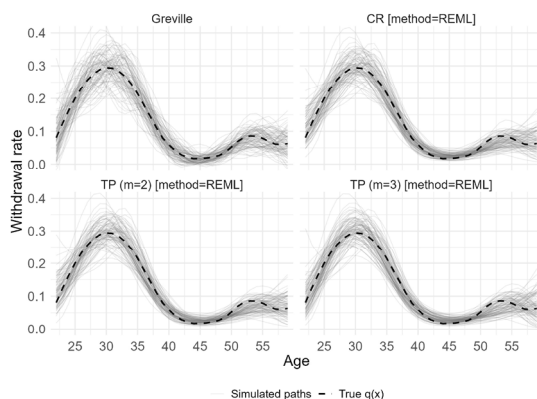


図 4-15 REML の場合の軌跡図 (q_{02} ，小規模)

最後に，大規模集団の場合の軌跡図を図示する。

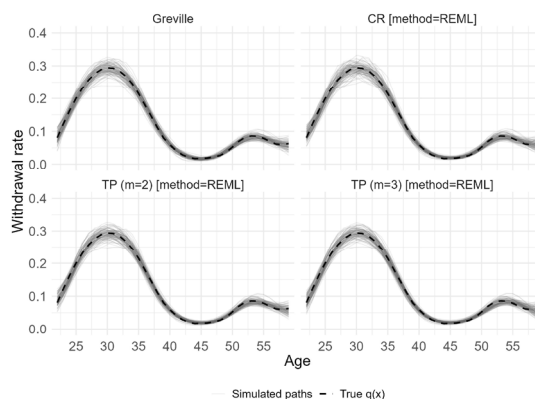


図 4-17 REML の場合の軌跡図 (q_{02} ，大規模)

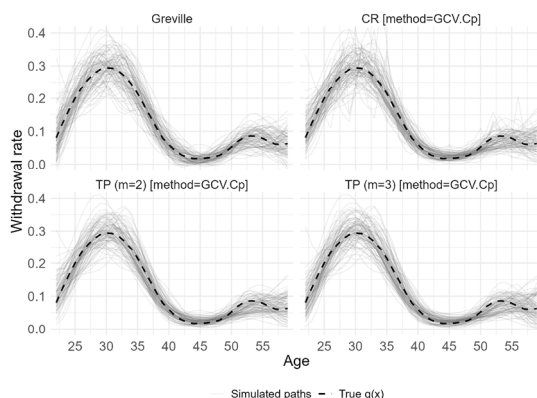


図 4-16 GCV.Cp の場合の軌跡図 (q_{02} ，小規模)

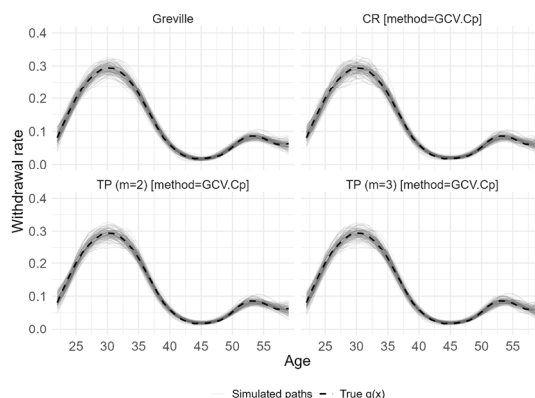


図 4-18 GCV.Cp の場合の軌跡図 (q_{02} ，大規模)

図 4-15 において，現行手法と比べると， cr ， $tp(m=2,3)$ の方がブレがやや少なく，平均的に真の脱退率に近い形状が得られていることが分かる。 q_{01} の場合と比べ，凹凸の変化がほとんど見られないような推定結果は確認されない。

一方，これまでと同様，図 4-16 の cr の結果には滑らかでない推定結果があり，図 4-15 ではこのような結果は見られないことが確認される。 $tp(m=2,3)$ については両者に大きな違いは無い。

図 4-17，4-18 ともにいずれの結果も真の脱退率に近い形状となっている。図 4-16 の cr の結果に確認された滑らかでない推定結果は図 4-18 には見られず，どの手法も視覚的にはほとんど同じであることが分かる。

以上より、次の3つが確認された。

- (1) cr かつ GCV. Cp の場合、一部の推定結果が滑らかな曲線にならない。これは真の脱退率の凹凸や人員規模に関わらず発生している。一方、REML の場合はこのような結果は見られない。
- (2) cr, tp($m = 2,3$) の場合、一部の推定結果が曲線として凹凸の変化がない。これは GCV. Cp よりも REML の方が多い。また、真の脱退率の凹凸が小さい場合に起こりやすく、人数規模が大きくなると生じなくなる。
- (3) REML と比べて GCV. Cp の推定結果は年齢ごとの細かな変化に追随する傾向にある。

4.3.1 (1) 推定結果が滑らかな曲線にならない

いくつかの推定結果において、推定脱退率がギザギザしたものとなっている。これは GAM の推定において乖離度（逸脱度）が優先された結果であり、有効自由度 EDF が異常に大きな値として現れる。このように GCV. Cp を用いる場合に過剰適合した結果が見られることは、3.1.3 節に引用した Wood[2025]の説明に整合している。

このような推定結果は、滑らかな曲線を得るという目的に反しており、望ましくないと言える。なお、年齢ごとの予測値をプロットしているためギザギザした結果となっているが、実際にはスプライン関数で当てはめているため、年齢を連続変数とみなすと滑らかな曲線にはなっている。

視覚的調査から 100 個のシミュレーション結果のうち、EDF ≥ 12 （ただし、真の脱退率が q02 で大規模集団の場合に限っては 12 でなく 18）を満たすものの推定結果はギザギザなものとなっており、そのような個数をまとめると表 4-1 の通りである。

表 4-1 (1)の個数(合計 100 個)

	cr, GCV. Cp	cr, REML
q01 小	7	0
q01 大	5	0
q02 小	9	0
q02 大	2	0

図 4-19 に EDF のヒストグラムを示した。確かに

異常に大きな値のものがいくつか確認できる。

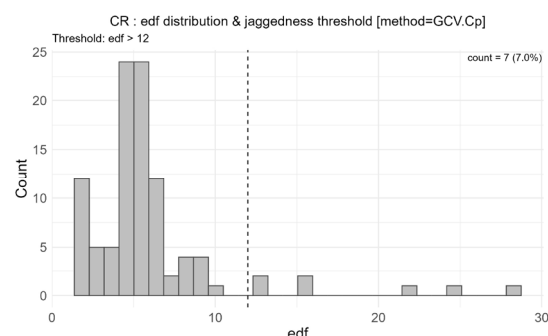


図 4-19 EDF のヒストグラム (cr, GCV. Cp, q01, 小規模)

cr の場合に REML と GCV. Cp の場合の EDF を比較すると図 4-20 のようになる。

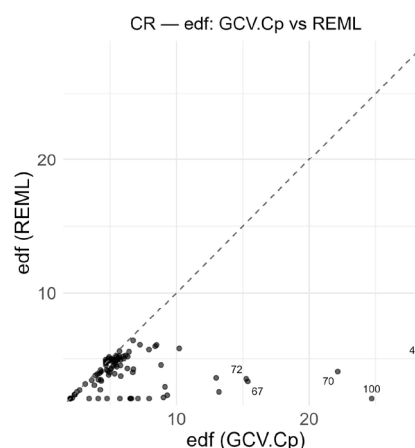


図 4-20 REML と GCV. Cp の EDF 比較 (cr, q01, 小規模)

例えばシミュレーション番号 4 の場合、GCV. Cp の EDF は 28 程度と非常に大きいですが、REML の場合は 6 程度と非常に小さい。REML の場合の EDF の最大値は 6 程度であり、確かに異常に大きな値は存在しない。

図 4-21 はシミュレーション番号 4 の結果である。左側は、GCV. Cp を用いた場合の推定結果であり、cr の推定結果に激しい凹凸が見られるが、REML を用いた場合 (図 4-21 右) は滑らかな曲線となっている。

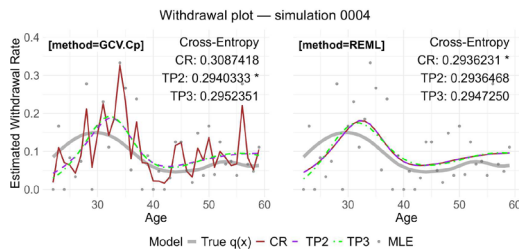


図 4-21 (1) の例 (q01, 小規模)

4.3.2 (2) 推定結果が凹凸の変化が無いものとなる

この推定結果は GAM の推定において平滑度が優先された結果であり、有効自由度 EDF が異常に小さな値として現れる。これは(1)と異なり、推定結果として特に問題はないと考えられるが、現行手法（5点移動平均法, Greville の方法）ではほとんど発生しない事象である。

視覚的調査から、EDF の下限値 (3.9) よりもわずかに大きな値として、 $EDF < m + 0.1$ を満たすものの推定結果に凹凸の変化が見られないことが分かった。このような個数をまとめると表 4-2 の通りである。

表 4-2 (2) の個数(合計 100 個)

	cr		tp($m = 2$)		tp($m = 3$)	
	GCV. Cp	REML	GCV. Cp	REML	GCV. Cp	REML
q01 小	11	23	11	23	10	15
q01 大	0	0	0	0	0	0
q02 小	0	0	0	0	0	0
q02 大	0	0	0	0	0	0

真の脱退率を q01, 小規模集団の場合の EDF のヒストグラムを図 4-22~4-25 に示す。ここで見やすさの観点から対数スケールに変換している。確かに EDF が異常に小さいものがいくつか存在することが確認される。

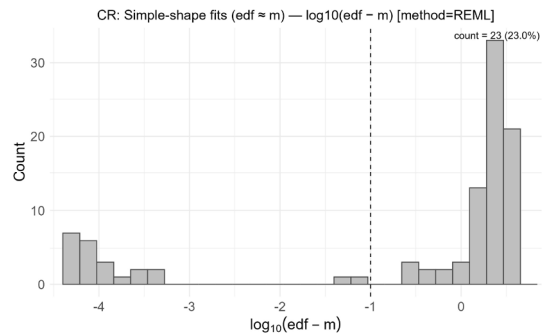


図 4-22 EDF のヒストグラム (cr, REML)

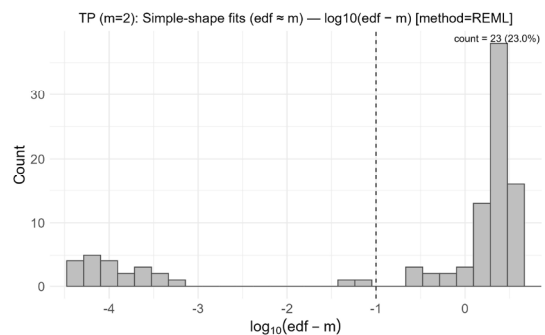


図 4-23 EDF のヒストグラム (tp($m = 2$), REML)

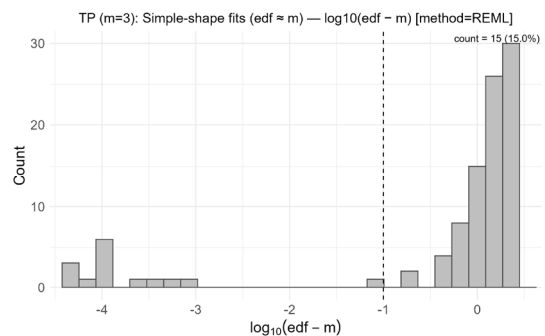


図 4-24 EDF のヒストグラム (tp($m = 3$), REML)

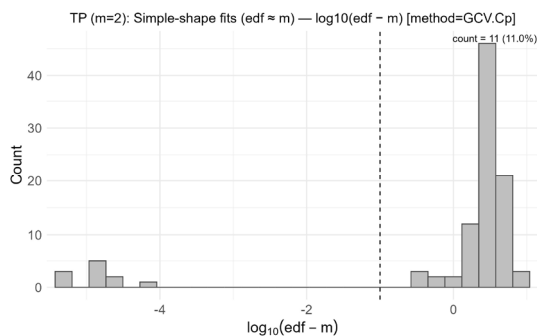


図 4-25 EDF のヒストグラム ($tp(m=2)$, GCV.Cp)

図 4-26 に一例を示した。GCV.Cp, REML のいずれの結果でも, cr と $tp(m=2)$ の推定結果がほぼ直線となっている (お互いのグラフは重なっている)。

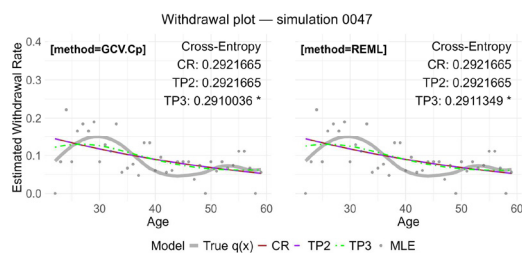


図 4-26 (2) の例 (q01, 小規模集団)

REML と GCV.Cp の EDF を比較すると図 4-27 のようになった。

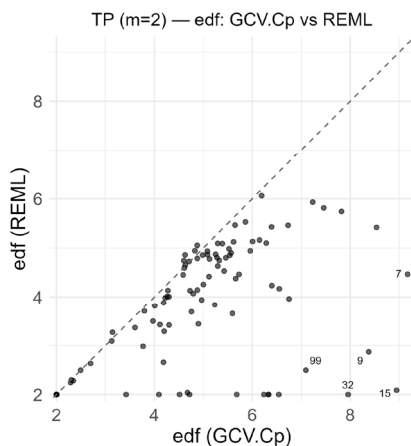


図 4-27 GCV.Cp と REML の EDF 比較 ($tp(m=2)$, q01, 小規模)

ほとんどの場合で REML の方が小さい, つまり GCV.Cp よりも REML の方がより平滑度最小化が重視される傾向にあることが分かる。これは(1)の結果とも整合的である。

図 4-27 において, 例えばシミュレーション番号 15 の結果は, GCV.Cp では 8.5 程度で, REML は 2 付近と大きく異なっている。この結果を図示すると図 4-28 の通りである。

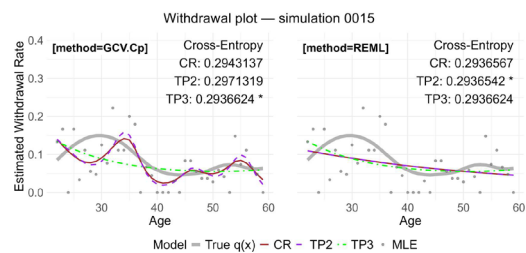


図 4-28 シミュレーション番号 15 の結果

$tp(m=2)$ の結果を見ると, 確かに GCV.Cp の結果 (図 4-28 左) は凹凸の変化が見られ, REML の結果 (図 4-28 右) に凹凸の変化は見られない。なお, GCV.Cp の推定結果の形状は真の脱退率からやや離れており, この場合は REML の結果の方が適切であると考えられる。

4.3.3 (3) GCV.Cp は細かな変化に追従しがち

既にいくつかの結果に見られた傾向であるが, REML と比べて GCV.Cp の結果は年齢ごとの細かな変化に追従する傾向にある。例えば, 図 4-29~図 4-31 もその一例である。これは真の脱退率, 人員規模に関わらず見られるものである。

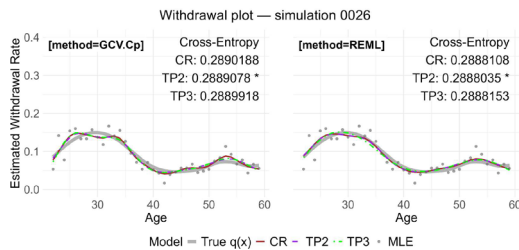


図 4-29 (3)の例 (q01, 大規模集団)

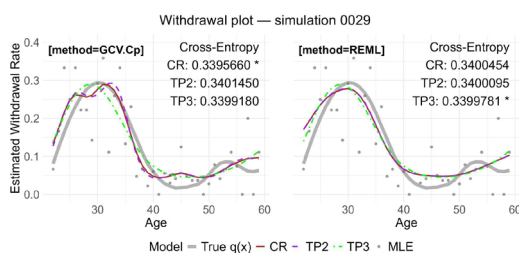


図 4-30 (3)の例 (q02, 小規模集団)

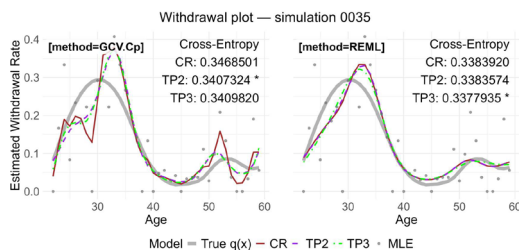


図 4-31 (3)の例 (q02, 小規模集団)

このことから、安定的な推定という観点からは、REMLの方が適切であると言える。

4.4 数値実験のまとめ

CE の分布を見ることにより、 $tp(m=2,3)$ が安定して良い推定結果を示した。また、これらの結果は視覚的にも違和感の無い結果となっていることが分かった。

パラメータの選択方法として GCV.Cp と REML を比較すると、後者の CE が平均的に小さく、より真の脱退率に近い推定結果を示していることが分かる。また cr かつ GCV.Cp の場合に滑らかな曲線にならない推定結果があったが、REML の場合はそのような結

果は無かった。一方で、GCV.Cp と比べて REML の場合の EDF は低めに出る傾向にあり、推定結果がより凹凸の変化が無い滑らかなものになる傾向は強いが、細かな変化に追従しない安定的な結果が得られることが明らかになった。

5 リスク評価への応用

本章ではリスク分担型企業年金以外の DB を想定する。ほとんどの DB において、財政悪化リスク相当額の算定が義務付けられている。ここで財政悪化リスク相当額とは、20 年に 1 回の頻度で生じると想定されるリスクに耐えうる額とされており、例えば、年金資産の運用結果により積立金の額が変動することで積立不足が生じるリスク（いわゆる価格変動リスク）、予定利率等の基礎率の変動に伴い債務が変動することで積立不足が生じるリスク（いわゆる負債変動リスク）などがリスクの例である。財政悪化リスク相当額の算定において、負債変動リスクを考慮するかどうかは任意とされている。

本論文では、負債変動リスクのうち、脱退率算定にかかわるリスクとして、推定脱退率と真の脱退率が異なるリスクを定量化する手法を提案し、その計算例を示す。

定義 5-1.

確率変数 X , 分布関数 $F(x)$, 信頼水準 α に対して、

$$\text{VaR}(\alpha) := \min\{x \mid F(x) \geq \alpha\}$$

$$\text{TVaR}(\alpha) := \frac{1}{1-\alpha} \int_{\alpha}^1 \text{VaR}(t) dt = E[X \mid X > \text{VaR}(\alpha)]$$

と定義する。

財政悪化リスク相当額の算定に使用するリスク尺度は TVaR(95%) を用いる。これは「DB 実務ガイド」に記載された取扱いである。VaR(95%) と比較すると、TVaR(95%) を使用することで分布の裾の厚さを反映したリスク定量化が可能となる。

5.1 定量化したいリスク

脱退率は脱退実績に関する観測データをもとに

推定され、それをもとに掛金額や DB 制度の債務である数理債務が計算されることとなる。すなわち脱退実績が変化することで、脱退率や掛金率、数理債務は変動することとなる。

ここではこのような数理債務の変動のうち、20 年に 1 回の頻度で生じると想定されるものを算定することを考える。つまり不確実性の源泉は、脱退率算定時の脱退実績のブレである。このようなリスクは「DB 実務ガイダンス」における「財政再計算時に発生する差損」（基礎率が変動することによる数理債務の変動）に分類されている。

例えば真の脱退率が与えられている場合、このリスクは次のように定量化できる。（図 5-1）

- ・ 真の脱退率に従う 30,000 個の脱退データを乱数により発生させ、10,000 個の脱退率を推定（3 回分で 1 個の脱退率を推定）
- ・ この脱退率を用いて 10,000 個の数理債務（V と表記する）を計算し、ヒストグラムにまとめる
- ・ リスク量＝「V の経験 TVaR (95%)－真の脱退率での V」

または、

- ・ リスク量＝「（V の標本平均/不偏標準偏差を平均/標準偏差とする正規分布の）TVaR (95%)－真の脱退率での V」

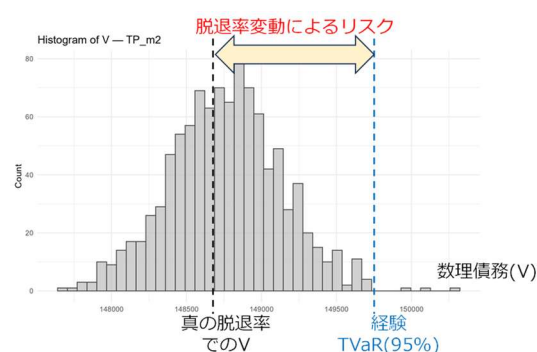


図 5-1 定量化したいリスクのイメージ

注意 5-1.

この手法は真の脱退率が既知であることを前提としているため、通常の実務では使用できない。また既知であったとしても、10,000 回数理債務を計算

することは実務上非常に負荷の大きい手法であり、極力数理債務の計算回数を減らしたリスク評価手法が実務上求められる。

5.2 リスク評価に関する現行の取扱い

5.2.1 「DB 実務ガイダンス」に記載の方法

財政再計算時に発生する負債変動リスクの計算方法として、「DB 実務ガイダンス」にはいくつかの方法が記載されているが、ここでは「基礎率を確率変数とする方法」について考える。

これは、「財政再計算時に設定する基礎率が、20 年に 1 回の頻度で発生すると予想される程度変動したと仮定した場合に、基礎率変動前後の数理債務の変動額を負債変動リスクとして考えるもの」であり、

- ・ 変動後の基礎率を作成する
- ・ リスク量＝「変動後の脱退率での V－通常の脱退率での V」

としてリスク量を計算するものである。

「変動後の脱退率」の作成方法として、「TVaR (95%) に対応する割掛率」を推定脱退率に乗じる方法が「DB 実務ガイダンス」に記載されている。ここで「TVaR (95%) に対応する割掛率」は、過去に適用した脱退率などを用いて、「脱退率の全年齢平均値」に関する分布を作成し、その TVaR (95%) を計算し、TVaR (95%) に対応する割掛率を「TVaR (95%)」÷「推定された脱退率の全年齢平均値」と計算される。

なお、「TVaR (95%) に対応する割掛率」として分布の上側と下側の 2 通り想定されるが、数理債務が脱退率に関して単調増加である場合は上側を、単調減少の場合は下側を用いることとされている。つまり変動後脱退率による数理債務は推定脱退率そのものによる数理債務を上回ることが保証される。

5.2.2 現行の手法の問題点（2.2 節の問題点 5）

「TVaR (95%) に対応する割掛率」の計算方法として、脱退実績の全年齢平均した値のみを用いるため、分布全体の情報を十分に反映できていない。また全年齢共通の割掛率を使用するため、年齢ごとのリス

クの違いを十分に捉えられない。また脱退率の算定に必要なデータだけでは算定できず、過去に採用された脱退率のデータ等が必要となり、新興企業が実施する DB など、十分なデータが得られない可能性がある。

上記の問題点は、現行手法（移動平均法, Greville 法）がいずれも決定論的な手法であることに起因する問題であると言える。

5.3 信頼区間を用いたリスク評価

GAM という統計モデリング手法を使用すれば、その信頼区間を用いることで、5.2.1 節で述べた「変動後の脱退率」を作成できる。

5.3.1 信頼区間の定式化

真の脱退率 q_x を推定するモデルとして、基底関数を $\{B_j(x)\}$ とする GAM を考える。

$$\text{logit } q_x = \sum_{j=1}^K \beta_j B_j(x) =: \eta(x)$$

この推定結果 $\text{logit } \hat{q}_x = \sum_{j=1}^K \hat{\beta}_j B_j(x) =: \hat{\eta}(x)$ に対し、 $V_{\hat{\beta}}$: 推定された係数 $\hat{\beta}$ に対する共分散行列として、各年齢 x での標準誤差を

$$SE_{\hat{\eta}}(x) := \sqrt{b(x)^T V_{\hat{\beta}} b(x)}$$

$$b(x) := [B_1(x), \dots, B_K(x)]^T$$

と定義すると、大標本の仮定の下、漸近正規性 $\hat{\beta} \sim N(\beta, V_{\hat{\beta}})$ から、

$$\hat{\eta}(x) = b(x)^T \hat{\beta} \sim N(\eta(x), SE_{\hat{\eta}}(x)^2)$$

を得る。よって信頼水準 α の信頼区間は次式で与えられる。

$$\text{logit}^{-1}(\hat{\eta}(x) \pm z \cdot SE_{\hat{\eta}}(x)); z = \text{VaR}(\alpha) \quad (5.1)$$

ここで VaR は正規分布に対する指標である。例えば 95%信頼区間（片側）の場合、 $z \approx 1.6449$ とすればよく、リスク尺度を TVaR(95%) とする場合、 $z \approx 2.0627$ とすればよい。

5.3.2 信頼区間の性質

信頼区間は、脱退率の推定値に対する推定精度を表し、係数推定誤差に起因するパラメータ不確実性

を定量化するものである。多数の標本を独立に繰り返して同じ推定手順を適用すると仮定したとき、各年齢 x について、構成した区間内に真の脱退率 q_x が含まれる頻度がおおむね設定した信頼水準（例えば 95%）となるように定義される。

5.3.3 「変動後脱退率」の作成方法

本論文では、リスク算定上のストレスシナリオとして、「全ての年齢における脱退が同一方向に変動する（完全正の相関を持つ）」というケースを想定する。この考え方は、資産クラスごとに完全な正の相関を仮定する価格変動リスク（標準算定方法）の考え方と整合的である。そして、各年齢に対する信頼区間の上側（または下側）を全ての年齢について結んだ曲線を「変動後脱退率」として用いる。つまり、

- ・数理債務が脱退率に関して単調増加である場合：

$$\text{logit}^{-1}(\hat{\eta}(x) + 2.06 \cdot SE_{\hat{\eta}}(x)) \quad (5.2)$$

- ・数理債務が脱退率に関して単調減少である場合：

$$\text{logit}^{-1}(\hat{\eta}(x) - 2.06 \cdot SE_{\hat{\eta}}(x)) \quad (5.3)$$

とする。

注意 5-2.

脱退率の信頼区間は特定の年齢に着目したときの脱退率の不確実性の幅を表すものであり、信頼水準を α としたとき、全ての年齢にわたって脱退率が上限（または下限）を超える確率は $1 - \alpha$ よりもずっと小さいことに留意しなければならない。

5.3.4 信頼区間に関する R コード

信頼区間は次の R コードで計算可能である。これは 3.2.1 節の R コードの続きを想定している。

```
# 信頼区間 (model_base に限定)
pred <- predict(
  model_base, newdata = newdata,
  type = "link", se.fit = TRUE
)
z <- dnorm(qnorm(0.95)) / 0.05 # TVaR(95%)
```

```
# 変動後脱退率の計算
lower_conf <- model_base$family$linkinv(
  pred$fit - z * pred$se.fit
)
upper_conf <- model_base$family$linkinv(
  pred$fit + z * pred$se.fit
)
```

3.2.3 節の例で実行した結果は図 5-2 の通りである。

```
> lower_conf
      1      2      3      4      5      6
0.064958418 0.112018776 0.128854261 0.108705918 0.085072955 0.071826681
      7      8      9     10     11     12
0.071144874 0.084887169 0.118516145 0.180727566 0.272734604 0.362480882
      13     14     15     16     17     18
0.392743807 0.334068607 0.211137346 0.098601838 0.040525200 0.019443246
      19     20     21     22     23     24
0.013465963 0.014218314 0.020774003 0.035059379 0.056543300 0.074323811
      25     26     27     28     29     30
0.070501539 0.046045751 0.023596150 0.012005081 0.007130120 0.005136714
      31     32     33     34     35     36
0.004331849 0.004045923 0.003915070 0.003635156 0.003118112 0.002683825
      37     38     39     40     41
0.002799907 0.004072558 0.007616502 0.013189358 0.017196981

> upper_conf
      1      2      3      4      5      6
0.68071526 0.61199266 0.60987520 0.62998556 0.60957572 0.54167859
      7      8      9     10     11     12
0.44860153 0.37641796 0.35975880 0.40425011 0.48817331 0.56850162
      13     14     15     16     17     18
0.60431927 0.56873564 0.45548386 0.30455996 0.18445770 0.12018913
      19     20     21     22     23     24
0.09845310 0.10794519 0.14918092 0.21968190 0.28739442 0.30615118
      25     26     27     28     29     30
0.26982688 0.21171716 0.15671605 0.11534707 0.09200807 0.08523944
      31     32     33     34     35     36
0.09009335 0.09799292 0.09911384 0.09175119 0.08347169 0.08042212
      37     38     39     40     41
0.08407127 0.09765416 0.13943230 0.27842627 0.59772139
```

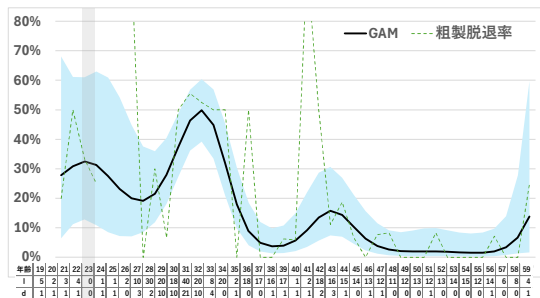


図 5-2 R 実行結果と変動後脱退率（帯の上下）

GAM による推定脱退率（実線）に合わせて、帯の上側が変動後脱退率（上側）、下側が変動後脱退率（下側）を表す。

20 歳台前半の加入者数が少ない部分の帯の幅が広く、20 歳台後半～30 歳台の加入者数が多い部分については狭くなっていることから、年齢ごとの脱退リスクの違いを表現できた方法であると言える。

5.4 具体例

第 4 章で用いた人員（大規模集団）を使用して、具体的にリスク量を計算する。引き続き $bs = "tp"$ ($m = 2$), $method = "REML"$ とする。

真の脱退率 q_{01} の場合に、第 4 章と同様の方法で推定脱退率を 1 つ作成し、同時に (5.2), (5.3) を用いて TVaR (95%) の場合の変動後脱退率を作成し、それぞれの数値債務を計算する。

5.4.1 計算の前提

給付設計は定額制を想定し、年齢ごとの要支給額を表す給付カーブとして 1 次式（後加重性が小さい）、3 次式（後加重性が大きい）を想定する（図 5-3）。一般に後加重性が大きいほど、脱退水準が下がると数値債務は大きくなる傾向にある。これは給付額が大きい年齢への到達率が上がるためである。

その他の計算基礎として、予定利率 2.0%、死亡率 0%、一時金選択率 100% と設定した。

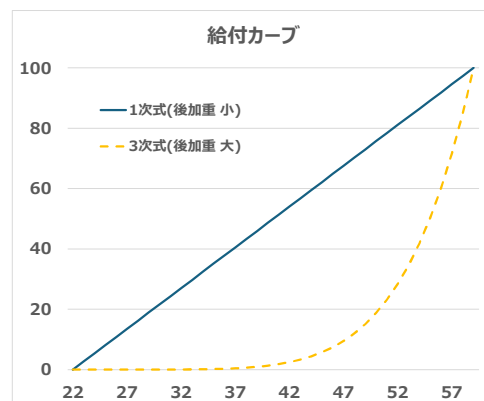


図 5-3 計算に使用する給付カーブ

5.4.2 計算結果

変動後脱退率は図 5-4 のようになった。以下では (5.2) に基づく変動後脱退率を upper, (5.3) に基づく変動後脱退率を lower と表記する。

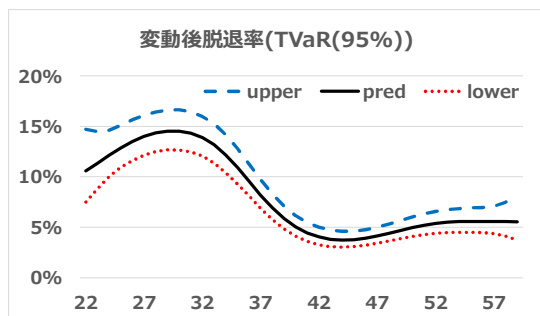


図 5-4 変動後脱退率の計算結果（大規模集団）

それぞれの変動後脱退率に対する数理債務の計算結果を表 5-1 に記載した。

表 5-1 数理債務の計算結果

脱退率	1 次式 (後加重:小)	3 次式 (後加重:大)
upper (5. 2)	158,741	131,867
pred (推定結果)	157,019	138,426
lower (5. 3)	155,755	143,697

給付カーブが 1 次式の場合，確かに脱退水準に対して数理債務は単調増加の傾向にあり，変動後脱退率を upper として，

$$\text{リスク} = 158,741 - 157,019 = 1,722$$

と計算できる．一方，給付カーブが 3 次式の場合，確かに脱退水準に対して数理債務は単調減少の傾向にあり，変動後脱退率を lower として，

$$\text{リスク} = 143,697 - 138,426 = 5,271$$

と計算できる．

なお，pred(推定脱退率)に対する upper (5. 2)，lower (5. 3) の割掛率を計算すると図 5-5 のようになり，概ね 2 割程度となった．このように変動後脱退率を求めることで，脱退率に割掛率を設定する際の根拠にも使用することができる．

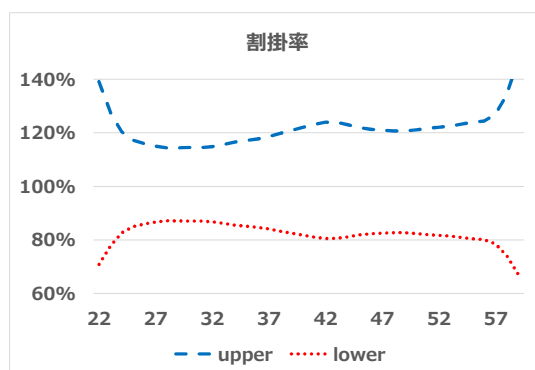


図 5-5 割掛率（大規模集団）

6 結論

本研究では，DB における脱退率推定の現行手法に存在する問題点を明らかにし，その解決策として一般化加法モデル（GAM）を用いた新たな推定手法を提案した．提案手法は統計ソフトウェア R の「mgcv パッケージ」により容易に実装可能であり，数値実験の結果を通じて，現行手法に比べて，集団の規模に関わらず安定的かつ合理的な推定結果が得られることを確認した．

特に，スプラインの基底関数の種類として bs=“tp”（微分平滑度の次数 $m = 2, 3$ ），平滑化パラメータの選択方法として method=“REML”を使用することが，真の脱退率との近似性や推定の安定性の観点から最適であることが示された．さらに，信頼区間を用いることで，財政悪化リスク相当額の算定に応用可能であることが示唆された．

微分平滑度の次数 $m = 2, 3$ の選択に関する判断基準について，本研究の数値実験の範囲では $m = 2$ と $m = 3$ の性能差は限定的であるため，現行手法との連続性，小標本における曲線形状 (3.5.1 節)，補外要件 (3.5.2 節) などの実務上の要請が選択の主たる決定要因となる．

まず，Greville の補整法が 3 回差分を使用することとの整合性を重視する場合には $m = 3$ が自然な選択である．また，人数規模が小さく情報量が乏しい場合，推定結果は $m - 1$ 次式（の逆ロジット変換）に近づきやすく， $m = 2$ を採用すると，これは 1 次

式（つまり直線）となり、解釈しやすい滑らかな形状が得られ、 $m = 3$ を採用すると、曲率の変化を許容した滑らかな形状となる。最後に補外の推定を行う場合、補外部分が $m - 1$ 次式（の逆ロジット変換）として表現されることを踏まえ、所望する補外形状に応じて $m = 2$ と $m = 3$ のいずれかを選択するという判断基準が得られる。なお、実務上の要請が特に無い場合、mgcv パッケージの既定値である $m = 2$ とすることは標準的であると考えられる。

一方で、本研究は人工データを用いた数値実験に基づくものであり、今後は実データを用いた実証的研究が求められる。また、実務への導入に際しては、関連する実務基準や実務ガイダンスへの手当てが求められる。

以上の成果は、DB における脱退率推定の改善とリスク管理の高度化に資するものであり、今後の実務への展開に向けた基盤を提供するものである。

謝辞

大阪大学大学院基礎工学研究科の鈴木讓教授、三井住友信託銀行の小田直人氏には本稿のチェックをいただき、非常に有意義なコメントおよび助言を多数いただいたことに深く感謝する。三菱 UFJ 信託銀行の白土和志氏、野田陽平氏をはじめ、公益社団法人日本アクチュアリー会データサイエンス関連基礎調査部会、ICA2026 研究会の各関係者等、多数の方々からコメントおよび助言をいただいたことに深く感謝する。最後に、有益な助言をくださった査読者の方々に心より感謝する。

参考文献

- 岩沢宏和[1991], 「小規模集団における脱退率の作成方法についての考察」, 『日本アクチュアリー会会報』, 第 44 号, 第 2 分冊, 105-130 頁.
- 小川裕之[1992], 「企業年金制度における脱退率のモデル化」, 『日本アクチュアリー会会報』, 第 45 号, 第 2 分冊, 261-291 頁.
- 桜井明[1981], 『スプライン関数入門』, 東京電機大学出版局.
- 鈴木讓[2021], 『機械学習のためのカーネル 100 問

with R』, 共立出版.

日本年金数理人会[2014], 『厚生年金基金実務基準 第 1 号（平成 26 年 11 月改訂）代行保険料率の計算に関する細目』.

日本年金数理人会[2025], 『確定給付企業年金に関する数理実務ガイダンス（最終改定 2025 年 2 月 19 日）』.

Duchon, J. [1977]. “Splines minimizing rotation-invariant semi-norms in sobolev spaces”. In: W. Schempp and K. Zeller (Eds.), *Constructive Theory of Functions of Several Variables*, Springer.

Greville, T. N. E. [1979],

“Moving-weighted-average smoothing extended to the extremities of the data”, *MRC Technical Summary Report # 2025*.

Hastie, T., and R. Tibshirani [1986],

“Generalized additive models”, *Statistical Science*, Vol. 1, No. 3, pp. 297-310.

Hastie, T., R. Tibshirani, and J. Friedman [2009], “*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*”, 2nd Edition, Springer.

Schoenberg, I. J. [1964], “Spline functions and the problem of graduation”, *Proceedings of the National Academy of Sciences of the U. S. A.*, Vol. 52, pp. 947-950.

Whittaker, E. T. [1923], “On a new method of graduation”, *Proceedings of the Edinburgh Mathematical Society*, Vol. 41, pp. 63-75.

Wood, S. N. [2017], “*Generalized Additive Models: An Introduction with R*”, 2nd Edition, Chapman & Hall/CRC.

Wood, S. N. [2025], “*mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*”, <https://cran.r-project.org/web/packages/mgcv/index.html> (2025/8/1).

Appendix A 命題 2-1 の証明

$a \leq x_1 < \dots < x_K \leq b$ とする. 以下, k 次スプラインとは, $k-1$ 回微分まで連続な区分 k 次多項式を指す. 特に自然 $2m-1$ 次スプラインは $2m-1$ 次スプラインである.

$\{x_1, \dots, x_K\}$ を knot とする任意の自然 $2m-1$ 次スプライン s について, m 回微分 $s^{(m)}$ は: $m-1$ 次スプラインであり, 両端 $(a, x_1), (x_K, b)$ では恒等的に 0 である. $s^{(m)}$ は区間 $[x_j, x_{j+1}]$ ($j = 1, \dots, K-1$) で $m-1$ 次多項式 $q_j(x)$ と等しいとする.

まず, ある実数 $\{d_1, \dots, d_K\}$ が一意に存在して, 区間 $[x_j, x_{j+1}]$ ($j = 1, \dots, K-1$) で,

$$q_j(x) = \sum_{i=1}^j d_i (x - x_i)^{m-1} \quad (\text{A.1})$$

$$\sum_{i=1}^K d_i x_i^r = 0 \quad (r = 0, 1, \dots, m-1) \quad (\text{A.2})$$

と表せることを示す. $s^{(m)}(x)$ は区間 $[a, x_1]$ では恒等的に 0 であり, $x = x_1$ における $m-1$ 回以下の微分は連続であるから, ある実数 d_1 が存在して, 区間 $[a, x_1]$ で $q_1(x) = d_1(x - x_1)^{m-1}$ と表せる. この実数 d_1 は区分定数関数 $s^{(2m-1)}$ の値から決まるため, s に対して一意である. 次に, $x = x_2$ において $m-1$ 回以下の微分が連続であるから, ある実数 d_2 が一意に存在して, 区間 $[x_1, x_2]$ で $q_2(x) = d_1(x - x_1)^{m-1} + d_2(x - x_2)^{m-1}$ と表せる. この議論を繰り返すことで, (A.1) と表せることが確かめられる. また, この議論から区間 $[x_K, b]$ で $s^{(m)}(x)$ は $\sum_{i=1}^K d_i(x - x_i)^{m-1}$ と表せるが, 仮定よりこれは恒等的に 0 であるため, (A.2) が成立する.

次に, 実数 $\{c_1, \dots, c_K\}$ を

$$c_i := \frac{d_i}{2 \binom{2m-1}{m} m!} \quad (i = 1, \dots, K)$$

と定めると, (A.2) から

$$\sum_{i=1}^K c_i x_i^r = 0 \quad (r = 0, 1, \dots, m-1) \quad (\text{A.3})$$

を満たす. (A.3) を用いると, 区間 $[x_j, x_{j+1}]$ で,

$$\begin{aligned} & \sum_{i=1}^K c_i |x - x_i|^{m-1} \\ &= \sum_{i=1}^j c_i (x - x_i)^{m-1} - \sum_{i=j+1}^K c_i (x - x_i)^{m-1} \\ &= \sum_{r=0}^{m-1} \binom{m-1}{r} x^{m-1-r} (-1)^r \left[\sum_{i=1}^j c_i x_i^r - \sum_{i=j+1}^K c_i x_i^r \right] \\ &= \sum_{r=0}^{m-1} \binom{m-1}{r} x^{m-1-r} (-1)^r \cdot 2 \sum_{i=1}^j c_i x_i^r \\ &= 2 \sum_{i=1}^j c_i (x - x_i)^{m-1} \end{aligned}$$

であるから, (A.1) より, 区間 $[x_j, x_{j+1}]$ で,

$$\begin{aligned} q_j(x) &= 2 \binom{2m-1}{m} m! \sum_{i=1}^j c_i (x - x_i)^{m-1} \\ &= \binom{2m-1}{m} m! \sum_{i=1}^K c_i |x - x_i|^{m-1} \\ &= \sum_{i=1}^K c_i \frac{d^m}{dx^m} |x - x_i|^{2m-1} \end{aligned}$$

と表せる. 同様に (A.3) を用いて計算すると, 両端 $(a, x_1), (x_K, b)$ で,

$$\sum_{i=1}^K c_i |x - x_i|^{m-1} \equiv 0$$

であることも確かめられるので, 区間 $[a, b]$ で

$$s^{(m)}(x) = \sum_{i=1}^K c_i \frac{d^m}{dx^m} |x - x_i|^{2m-1}$$

と表せることが示された. この m 回不定積分を取ることで, $m-1$ 次多項式 $p_{m-1}(x)$ を用いて,

$$s(x) = p_{m-1}(x) + \sum_{i=1}^K c_i |x - x_i|^{2m-1}$$

と表せる. よって命題 2-1 は示された.

Appendix B 命題 2-2 の証明

以下, $\{x_1, \dots, x_n\} = \{\xi_1, \dots, \xi_K\}; \xi_1 < \xi_2 < \dots < \xi_K$ で $\{\xi_1, \dots, \xi_K\}$ を定義する.

命題 2-2 の証明には, 次の 2 つの良く知られた事

実を使用する。

命題 B-1. 桜井[1981, 定理 2.7]

$m < K$ のとき, 指定した点 $(\xi_1, \zeta_1), \dots, (\xi_K, \zeta_K)$ を全て通る関数の中で, $\{\xi_1, \dots, \xi_K\}$ を knot とする自然 $2m-1$ 次スプライン s が一意に存在する。

命題 B-2. 桜井[1981, 定理 2.8]

$m < K$ のとき, $\{\xi_1, \dots, \xi_K\}$ を knot とする自然 $2m-1$ 次スプライン s と $f \in W_m[a, b]$ は $f(\xi_i) = s(\xi_i)$ ($i = 1, \dots, K$) を満たすとする。このとき, 次が成立する,

$$\int_a^b \{s^{(m)}(x)\}^2 dx \leq \int_a^b \{f^{(m)}(x)\}^2 dx$$

これらを用いて命題 2-2 を証明する。任意の $f \in W_m[a, b]$ に対して, 命題 B-1 より, $f(\xi_i) = s(\xi_i)$ ($i = 1, \dots, K$) を満たす, $\{\xi_1, \dots, \xi_K\}$ を knot とする自然 $2m-1$ 次スプライン s が一意に存在し, 命題 B-2 より

$$\int_a^b \{s^{(m)}(x)\}^2 dx \leq \int_a^b \{f^{(m)}(x)\}^2 dx$$

が従う。

$L(f(x_1), y_1, \dots, f(x_n), y_n) = L(s(x_1), y_1, \dots, s(x_n), y_n)$ であるから, $J(f) \geq J(s)$ が示された。

注意 B-1.

命題 2-1 と命題 2-2 を合わせると, 汎関数

$$L(f(x_1), y_1, \dots, f(x_n), y_n) + \lambda \int_a^b \{f^{(m)}(x)\}^2 dx$$

の最小化問題の解 f は

$$p_{m-1}(x) + \sum_{i=1}^K c_i |x - \xi_i|^{2m-1} \quad (B.1)$$

$$\sum_{i=1}^K c_i x_i^r = 0 \quad (r = 0, \dots, m-1) \quad (B.2)$$

なる関数のクラスに限定される。同様の事実は一般の d 次元の場合にも成立する。本論文ではこの事実を 1 次元の場合に限定して初等的な計算のみで証明したが, 一般の d 次元の場合には, Sobolev 空間 $W_m[a, b]$ に再生核 Hilbert 空間の構造を入れ, カー

ネル法の表現定理 (Representer theorem) を用いて証明することができる。例えば Hastie, Tibshirani and Friedman[2009], 鈴木[2021]を参照。

Appendix C UBRE, REML, AIC, BIC

GAM において平滑化パラメータ λ の選択に使用される UBRE, REML, AIC といった概念について説明する。(詳細は Hastie, Tibshirani and Friedman[2009], Wood[2017]を参照)

まず, mgcv パッケージにおいて, λ の選択方法を表す gam 関数の引数 method の既定値は “GCV.Cp” であるが, この意味するところは, 指数型分布族におけるスケールパラメータ s が既知の場合 (二項分布は $s = 1$) は UBRE, 未知の場合 (正規分布など) は一般化交差検証 (GCV) とされており, 本論文では二項分布を使用するため, UBRE を用いて平滑化パラメータが選択されることとなる。

UBRE (Un-Biased Risk Estimator) criterion は, 最小化問題 (3.3) の解である $\hat{\beta}$ における逸脱度 $D(\hat{\beta})$ と有効自由度 EDF, 指数型分布族におけるスケールパラメータ s を用いて,

$$\text{UBRE} := \frac{D(\hat{\beta})}{n} + 2s \cdot \frac{\text{EDF}}{n} - s$$

で定義される。これは, 真のリスク (期待逸脱度)

$$R(\lambda) := \frac{1}{n} \mathbb{E}[D(\beta_\lambda)]$$

の漸近的な不偏推定量となることが知られている。つまりこれは予測精度という観点から λ を選択することを目指す指標であると言える。なお, 本論文で使用する二項分布の場合, $s = 1$ であり,

$$\text{UBRE} = \frac{D(\hat{\beta})}{n} + 2 \cdot \frac{\text{EDF}}{n} - 1 \quad (C.6)$$

となる。

次に赤池情報量規準 (Akaike's Information criterion; AIC) とは, 最小化問題 (3.3) の解である $\hat{\beta}$ と有効自由度 EDF を用いて,

$$\text{AIC} := D(\hat{\beta}) + 2 \cdot \text{EDF} \quad (C.7)$$

で定義される。明らかに UBRE は AIC の線形変換で

表せるため、両者によるモデル選択は同値である。

AIC の第 1 項はデータの負の対数尤度であり、小さいほどデータへの当てはまりが良いモデルである。一方、AIC の第 2 項は有効自由度、つまりモデルの複雑度を表し、小さいほどシンプルなモデルとなる。つまり AIC が小さいモデルを選ぶことで、データへの当てはまりとモデルのシンプルさを両立したモデルを選択することができる。

一方、ベイズ情報量規準 (Bayesian information criterion ; BIC) とは、サンプルサイズ n を用いて

$$\text{BIC} := D(\hat{\beta}) + \text{EDF} \cdot \log n \quad (\text{C.8})$$

で定義される。これは AIC と比べると、第 2 項の罰則が強いいため、よりシンプルなモデルが選ばれる傾向にある。さらに BIC は、モデル集合の中に真のモデルが含まれていると仮定したとき、漸近的には真のモデルを確率 1 で選択できるという一致性 (consistency) を満たすことが知られている。

最後に REML (Restricted Maximum Likelihood) とは、平滑項をランダム効果とみなし、その分散成分を最尤推定 (restricted) で求める方法である。具体的には、尤度から λ 以外のパラメータを積分消去した周辺尤度を最大化する λ を選ぶこととなる。周辺尤度の計算にはラプラス近似が使用される。この手法の罰則項は GCV, Cp の罰則項 (edf) よりも強く、大きめの λ (=より滑らか) を選ぶ傾向があり、過少平滑 (undersmoothing) を抑制しやすい手法であると言える。

Appendix D クロスエントロピーの性質

CE が最小となるとき、全ての年齢 x で $q_x = \hat{q}_x$ が成り立つことを確認する。ベルヌーイ分布における Kullback-Leibler divergence の基本的な性質より、各 x において、

$$q_x \log \frac{q_x}{\hat{q}_x} + (1 - q_x) \log \frac{1 - q_x}{1 - \hat{q}_x} \geq 0$$

$$q_x \log \frac{q_x}{\hat{q}_x} + (1 - q_x) \log \frac{1 - q_x}{1 - \hat{q}_x} = 0 \Leftrightarrow q_x = \hat{q}_x$$

が成立する。したがって、

$$KL(q_x \parallel \hat{q}_x) := \sum_{x=x_e}^{x_r-1} \left\{ q_x \log \frac{q_x}{\hat{q}_x} + (1 - q_x) \log \frac{1 - q_x}{1 - \hat{q}_x} \right\}$$

$$E := \frac{1}{x_r - x_e} \sum_{x=x_e}^{x_r-1} \{ q_x \log q_x + (1 - q_x) \log (1 - q_x) \}$$

と定義すると、 E は真の脱退率 $\{q_x\}$ のみに依存する定数であり、

$$(x_r - x_e)CE = KL(q_x \parallel \hat{q}_x) + (x_r - x_e)E$$

と表せることから、

$$CE \geq E$$

$$CE = E \Leftrightarrow q_x = \hat{q}_x \quad (\forall x)$$

が成り立つ。よって CE が最小になるのは、全ての年齢 x で $q_x = \hat{q}_x$ が成り立つことが確認された。

Improving Withdrawal Rate Estimation Using Generalized Additive Models

Yuki Kurumadani^{*†}

Hirokazu Iwasawa[‡]

Abstract

Accurate estimation of withdrawal rates is fundamental to the actuarial valuation of defined benefit pension plans. Traditional estimation methods, however, suffer from several limitations. One such issue is the inability to adjust the degree of smoothing according to the amount of observed data at each age. To address these shortcomings, we propose a statistically rigorous approach based on generalized additive models (GAMs) assuming a Bernoulli distribution. This method enables smooth and flexible estimation of age-specific withdrawal rates using individual-level data on age and withdrawal status, while automatically modulating the smoothness of the fitted curve in accordance with the volume of observations. Empirical analysis demonstrates that the proposed method achieves higher stability and accuracy than conventional techniques, demonstrating its practical relevance in actuarial practice. Furthermore, our results suggest that incorporating confidence intervals enables the assessment of uncertainty in liability valuation arising from withdrawal rate estimation.

^{*} Graduate School of Engineering Science, Osaka University Email: kurumadani@sigmath.es.osaka-u.ac.jp

[†] Resona Bank, Ltd. Email: Yuki.Kurumadani@resonabank.co.jp

[‡] Faculty of Commerce, Graduate School of Accountancy, Waseda University Email: iwahiro@bb.mbn.or.jp