

ノンパラメトリック回帰分析と生命表*

—我が国生保標準生命表における補整の考察—

小 暮 厚 之[†]

Life Table and Nonparametric Regression—An Inquiry into the Graduation
of the Standard Life Table for Japanese Life Insurance Companies—

Atsuyuki Kogure*

生命表は死亡率の経験データそのものではなく、オリジナルな粗死亡率データに「補整」を施した結果である。本論文では、補整を粗死亡率データから真の死亡率を推定するプロセスであると捉え、ノンパラメトリック回帰分析の視点から生命表の作成について考察する。日本アクチュアリー会による「生保標準生命表」を例に取り、従来の標準的な移動平均法に代わり、最小2乗法や尤度法という統計モデルを局所的にあてはめる新たなアプローチを試みる。

The life table is not a mere summary of the raw mortality data, but represents the adjusted mortality rate by the graduation. In this article I consider the graduation as a process of estimating the true mortality rate from the raw mortality data and examine its properties from a nonparametric regression view. In particular, I make an inquiry into the Standard Life Table for Japanese Life Insurance Companies and apply a new class of new nonparametric techniques such as the local polynomial regression and the local likelihood.

1. はじめに

生命表は、対象となる人口集団の死亡率を年齢別に表にまとめたものである。我が国では、国民全体あるいは都道府県別など様々な人口集団に対する生命表が公表されている。また、日本アクチュアリー会による「生保標準生命表」は、保険契約者を人口集団とする生命表であり、生命保険料率の計算基礎として広く採用されている。

このような生命表は経験データそのものではなく、オリジナルな粗死亡率データに「補整」を施した結果である。いかなる補整を採用するかによって生命表に大きな差異が生じるのにも拘らず、補整の意義や意味について統計学の立場からの議論は少ないように思われる。本論文では、補整を粗死亡率データから真の死亡率を推定するプロセスであると捉え、ノンパラメトリック回帰分析の視点から生命表の作成について考察する。例として、日本アクチュアリー会による「生保標準生命表」を取り上げる。まず、「生保標準生命表」で用いられている Greville

* 本稿は、2003年度日本統計学会特別セッション「統計学と保険」（2003年9月4日於名城大学）及び21世紀COEプログラム「日本・アジアにおける総合政策学先導拠点」（慶應義塾大学大学院政策・メディア研究科）によるワークショップ（2003年11月8日於慶應義塾大学、日本保険・年金リスク学会との共催）において発表した内容に基づいてる。本稿を作成するに当たり、小島茂氏（ジブラルタ生命）より生保標準生命表に関してご教示頂きました。また、本誌査読者から大変有益なコメントを頂きました。ここに感謝致します。

[†] 慶應義塾大学総合政策学部, E-mail: kogure@sfc.keio.ac.jp

(1974) の3次13項式による補整がカーネル法の特例であることを見る。さらに、最近のノンパラメトリック手法の発展を踏まえて、局所回帰モデルと局所尤度モデルを用いた新たな補整へのアプローチを試みる。

2. 生命表

2.1 粗死亡率データと予定死亡率

表1は「生保標準生命表」から一部を抜粋したものである。それは、男子の各年齢（単位：歳）に対する予定死亡率（単位：パーミル）を表す。図1はこの予定死亡率（曲線で表示）とその元になっている粗死亡率データ（円で表示）の年齢に対するプロットである。離散的な粗

表1 男子の生保標準生命表による予定死亡率

年齢	死亡率	年齢	死亡率	年齢	死亡率	年齢	死亡率	年齢	死亡率
x	$1000 q_x$	x	$1000 q_x$	x	$1000 q_x$	x	$1000 q_x$	x	$1000 q_x$
0	1.10	10	0.15	20	1.14	30	0.84	40	1.56
1	0.76	11	0.14	21	1.07	31	0.85	41	1.71
2	0.50	12	0.15	22	0.99	32	0.88	42	1.88
3	0.33	13	0.22	23	0.92	33	0.92	43	2.08
4	0.24	14	0.34	24	0.88	34	0.98	44	2.29
5	0.22	15	0.52	25	0.86	35	1.05	45	2.51
6	0.22	16	0.73	26	0.85	36	1.13	46	2.73
7	0.21	17	0.94	27	0.85	37	1.22	47	2.96
8	0.19	18	1.09	28	0.84	38	1.33	48	3.21
9	0.17	19	1.15	29	0.84	39	1.44	49	3.48

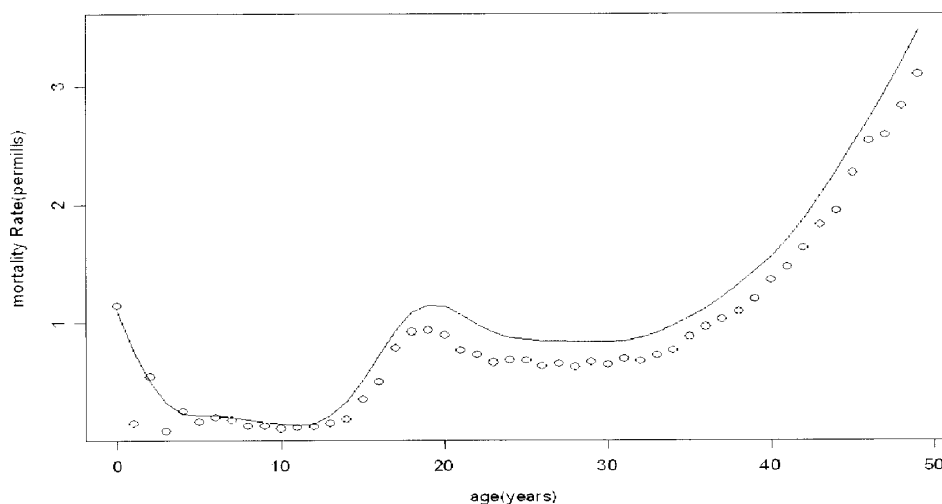


図1 粗死亡率と予定死亡率

死亡率データから予定死亡率のような滑らかな死亡曲線を作成することを補整という。

生命表を作成する場合、年齢は1年間隔に離散化されるのが一般的である。 x 歳の年齢の人を e_x 人観測したとき1年以内に d_x 人が死亡したとすると、 x 歳の人の粗死亡率は $\dot{q}_x = d_x / e_x$ と算出される¹。

2.2 死亡率の確率モデル

x 歳の人の余命時間を確率変数 T_x で表すと、真の死亡率 $q(x)$ は

$$q(x) \equiv \Pr(T_x < 1) \quad (1)$$

である。粗死亡率 \hat{q}_x は $q(x)$ の推定値である。死亡法則を表す様々な T_x の分布が古くから提案されている。2つの代表的な死亡法則の場合には、死亡率は

- Weibul 法則

$$q(x) = 1 - \exp\left(-\frac{k}{n}[(x+1)^{n+1} - x^{n+1}]\right) \quad (2)$$

- Makeham-Gompertz 法則

$$q(x) = 1 - \exp(-A - mc^{x(c-1)}) \quad (3)$$

とモデル化される。

しかし、このようなパラメトリック・モデルによって実際のデータを完全に説明できるとは限らない。場合によっては、パラメトリック・モデルをフィットさせる代わりに、粗死亡率をノンパラメトリックに平滑化して、真の死亡利率を推定する「補整」(graduation)が適用される²。

3. 生保標準生命表における補整

本節では、まず我が国保険会社とその保険料率算定の基礎として用いている生保標準生命表(1996)における補整について簡潔に説明する。

3.1 粗死亡率

生保標準生命表の基礎データは我が国生命保険会社 22 社(当時)の契約者のデータを合算したものである。観察期間は、15 歳以上(有審査)については 1989 年から 1991 年までの 3 年間であり、契約者数の少ない 14 歳以下(有審査と無審査の合計)については、1986 年から 1991 年の 6 年間である。

すべての契約者の寿命が x 歳から $x+1$ 歳まで完全に観察されることはない。これは図 2 のいわゆるレキシス図(Lexi Diagram)によって例示される。レキシス図の横軸は観察期間、縦軸は年齢を表わす。図の中の各斜線は、各契約者の寿命を表わす。記録が完全でないケースは

- 観察期間の開始時点より前に x 歳になった契約者の寿命
- x 歳になってから保険契約を結んだ契約者の寿命
- 死亡以外の理由で $x+1$ 歳になる前に保険契約が途中で打ち切られた契約者の寿命

である。粗死亡率は契約件数 e_x に対する死亡者数 d_x の比として

$$\hat{q}_x \equiv \frac{d_x}{e_x}$$

と定義される。以上のデータの不完全性を考慮に入れ、 e_x はレキシス図の各線分の縦方向の

¹ 3.1 節で述べるように、実際には「打ち切り」などがあるためより複雑である。

² 別のアプローチとして、Weibull 法則を世代ごとに別々にあてはめるような方法も提案されている(荒井, 2001)

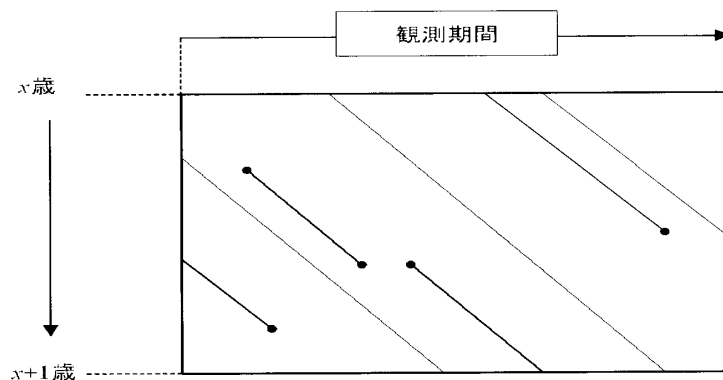


図2 レキシス図

長さの和である経過契約件数であり、 d_x はレキシス図で観測される死亡数である³。

3.2 補 整

補整とは英語の *graduation* に充てた言葉であり、本来の意味は「平滑化」であろう。死亡率を算出する際の重要なステップが「平滑化」であることから、粗死亡率データから生命表を作成するプロセス一般を補整と呼ぶことも多い。生保標準生命表における「補整」は次の3つのステップからなる：

(1) 1次補整

「数学的危険論」による調整

(2) 2次補整

移動平均による平滑化 (Greville の補整)

(3) 3次補整

高年齢者層の死亡率の平滑化 (52歳以上の粗死亡率データに1.2節で述べた Makeham-Gompertz 法則をあてはめる)

本論文では、2次補整における移動平均に代替する方法としてノンパラメトリック法を考察する。

3.2.1 1次補整

粗死亡率 \dot{q}_x の1次補整値は、

$$q_x^{(1)} = \dot{q}_x + \min \left(\sqrt[2]{\frac{\dot{q}_x(1-\dot{q}_x)}{n_x}}, 0.34\dot{q}_x \right) \quad (2)$$

と与えられる。ここで、 n_x は、

$$n_x = 4,000,000 \times \phi(x|41.4, 14.9^2)$$

であり、 $\phi(x|\mu, \sigma^2)$ は平均が μ 、分散が σ^2 の正規分布の密度関数を表す。1次補整値は粗死亡率の信頼区間の上限を与える。1次補整は、個々の被保険者の生死を互いに独立なベルヌーイ試行分布と見なし、中心極限定理が成立するという集团的危険理論に基づいて得られると考えられる。しかし、(2)では契約者数 e_x ではなく、 n_x が用いられることに注意されたい。

³ これ以外にも粗死亡率の定義として、

$$\dot{q} = \frac{d_x}{e_x + \frac{d_x}{2}} \text{ または } \dot{q} = 1 - \exp\left(-\frac{d_x}{e_x}\right)$$

を用いる場合もある。詳細は Gerber(1991) を見よ。

n_x は、契約者の年齢の分布が 1991 年度の経過契約件数から計算した平均 41.4 と標準偏差 14.9 を持つ正規分布に従うとき、400 万人の保険契約者集団における x 歳の被保険者の人数の期待値である。ここで、保険契約者数をなぜ 400 万人としているかは不明であるが、標準的な保険契約者集団の大きさとして 400 万件を想定したと類推できる⁴。

3.2.2 2 次補整

1 次補整値の系列を $\{q_j^{(1)}\}$ とするとき、 x 歳の死亡率の 2 次補整値 $\{q_j^{(2)}\}$

$$q_x^{(2)} = \sum_{j=x-6}^{x+6} a_{j-x} q_j^{(1)}$$

と与えられる。ここで、 $\{a_j\}$ は移動平均の加重であり

$$a_j \equiv a_{-j}, \quad \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{pmatrix} \equiv \begin{pmatrix} 0.240058 \\ 0.214337 \\ 0.147356 \\ 0.065492 \\ 0 \\ -0.027864 \\ -0.019350 \end{pmatrix}$$

で与えられる。これは、Greville (1974) の 3 次 13 項式と呼ばれる加重平均である。この加重は、

$$\sum_{j=-6}^6 a_j = 1; \quad \sum_{j=-6}^6 j^s a_j = 0 \quad (s=1, 2, 3) \quad (3)$$

という性質を満たす。

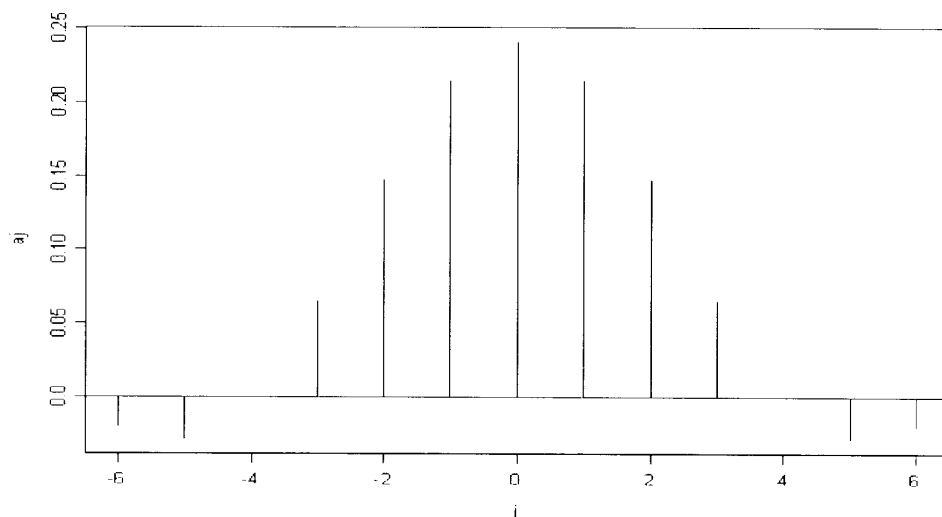


図3 Greville の 3 次 13 項式の加重

⁴ 統計理論の立場からは、このような正規近似には以下のような問題が残ると思われる。まず、 e_x 自体が確率変数であるため、 n_x は 2 項分布には従わない。また、たとえその点が正当化されたとしても、 $q(x)$ の値が小さいため正規分布による近似の誤差が十分に小さいかどうか疑問が残る。

4. 補整とノンパラメトリック回帰分析

4.1 移動平均の意味

生保標準表で用いられる2次補整は

$$\begin{aligned} q_x^{(2)} &= \sum_{j=x-6}^{x+6} a_j - q_x^{(1)} \\ &= \sum_{j=-6}^6 a_j \dot{q}_{x+j} + \sum_{j=-6}^6 a_j \min \left(\sqrt[2]{\frac{\dot{q}_{x-j}(1-\dot{q}_{x+j})}{n_{x+j}}}, 0.3 \dot{q}_{x+j} \right) \end{aligned}$$

と表わせる. 右辺の第1項

$$\dot{q}(x) \equiv \sum_{j=-6}^6 a_j \dot{q}_{x+j} \quad (4)$$

は粗死亡率に移動平均を施した平滑化である. これが本来の意味の補整であり, 以下ではこれを考察する. 多くの場合, 1.2節で述べたようなパラメトリック・モデルによる補整は, 複雑な死亡率の構造を必ずしも正しくは反映しない. そのため, (4)のような移動平均による補整が行われる.

移動平均による補整は, 粗死亡率データから真の死亡率を推定することに他ならない. これを見るために, 粗死亡率データの構造を考えよう. 年齢と粗死亡率に対するデータ $\{(x_i, \dot{q}_i), i = 1, 2, \dots, n)\}$ が与えられているとしよう. 粗死亡率データの構造として

$$\dot{q}_i = q(x_i) + \epsilon_i \quad (5)$$

を想定する⁵. ここで, $q(x)$ は (1) で定義される x 歳の人の死亡率を表す確定的な関数とする. また, ϵ_i は誤差項であり, 簡単化のために互いに独立であると仮定する. すべての i に対して

$$x_i - x_{i-1} = \Delta$$

とし, $\Delta \rightarrow 0$ という状況を考える. q が十分に滑らかであるとすれば, その期待値は,

$$\begin{aligned} E[\dot{q}(x)] &= \sum_{j=-6}^6 a_j E[\dot{q}_{x+j}] = \sum_{j=-6}^6 a_j q(x+j\Delta) \\ &= \sum_{j=-6}^6 a_j \left(q(x) + j\Delta q'(x) + \frac{(j\Delta)^2}{2!} q''(x) + \frac{(j\Delta)^3}{3!} q'''(x) \right) + O(\Delta^4) \\ &= \left(\sum_{j=-6}^6 a_j \right) q(x) + \left(\sum_{j=-6}^6 j a_j \right) \Delta q'(x) \\ &\quad + \left(\sum_{j=-6}^6 j^2 a_j \right) \frac{\Delta^2}{2!} q''(x) + \left(\sum_{j=-6}^6 j^3 a_j \right) \frac{\Delta^3}{3!} q'''(x) + O(\Delta^4) \end{aligned}$$

となる. ここで, (3) に注意すると,

$$E[\dot{q}(x)] = q(x) + O(\Delta^4)$$

を得る. (3) の条件によって, 移動平均によるバイアスが $O(\Delta^4)$ にまで減少できたことになる. また, (5) の分散は

$$\begin{aligned}\text{Var}(\hat{q}(x)) &= \text{Var}\left(\sum_{j=-6}^6 a_j \hat{q}_{x+j}\right) = \sum_{j=-6}^6 a_j^2 \text{Var}(\epsilon_{x+j}) \\ &\leq \left(\sum_{j=-6}^6 a_j^2\right) \max_j [\text{Var}(\epsilon_{x+j})] \approx 0.2038 \max_j [\text{Var}(\epsilon_{x+j})]\end{aligned}$$

となる。従って、もしも誤差分散 $\text{Var}(\epsilon_{x+j})$ が一定ならば分散は約 5 分の一に減少することになる。

以上から、粗死亡率の代わりに Greville の移動平均を用いることによって、バイアスを小さく抑えながら分散を著しく減少できることが分かる。すなわち、補整とは、粗死亡率データからより効率的に真の死亡率を推定するプロセスに他ならない。一般の移動平均によっても同様な効果を実現できる⁶。その中で、なぜアクチュアリーが Greville の補整を採用するに至ったかは明らかではない。Greville は補整以外にも様々な分野でアクチュアリー界に貢献した研究者であり、そのことが関係しているのかもしれない⁷。

このような補整の意義に対しては、粗死亡率はそもそも膨大な観測値から計算された値であるから誤差分散は無視しうるほど小さく、そもそも補整する意味はないという批判がある。しかし、粗死亡率は各契約者の様々な異質性を無視して年齢だけによって集計した「マクロ」変数であり、その誤差が十分に小さいかどうかは必ずしも明らかではない。アクチュアリーが生命表作成において補整を用いてきたひとつの理由であろう。

4.2 回帰モデル

(5) のデータの構造は、より一般に、粗死亡率を被説明変数、年齢を説明変数とする非線形回帰モデル

$$\hat{q}_i = q(x_i) + v(x_i)^{1/2} Z_i \quad (6)$$

と見る事ができる。真の死亡率 $q(x_i)$ は、この回帰モデルの回帰関数 q の $x=x_i$ における値を表す。ここで、 $\{Z_i\}$ は平均 0、分散 1 の IID 列であり、 $v(x_i)$ は死亡率 \hat{q}_i の分散不均一性を表す未知の関数である。

移動平均による補整は、回帰関数 q の関数形を決めることなく q を推定するという意味で、ノンパラメトリック推定法である。このような立場から、Copas and Haberman (1983), Gavin, Haberman and Verral (1994) ではノンパラメトリック推定法による補整の研究がなされている。彼らが用いている代表的な推定法は、Nadaraya=Watson 法と呼ばれるカーネル回帰法である⁸。以下で見るように、Nadaraya=Watson 法は移動平均の自然な拡張であり、その特殊ケースと考えることにより、前述の Greville の移動平均による補整の持つ意味を理解できる。しかし、Nadaraya=Watson 法には、「境界バイアス」として知られている問題点があり、今日ではそれに代わる手法として、局所回帰法や局所尤度法という新しい手法が用いられるようになってきた。本節では Nadaraya=Watson 法による補整を概観し、4 節以降で新たなノンパラメトリック手法による補整について述べる。

⁵ $q(x)$ が 0 以上 1 以下でなければならないという制約を考慮すると、 \hat{q}_i の対数変換やロジット変換に対してこのような構造を仮定することも考えられる。

⁶ 例えば、Anderson (1971) Chapter 3 を参照されたい。

⁷ Greville の様々な貢献については、山口他 (1994) を参考にされたい。

⁸ カーネル法によるノンパラメトリック回帰については、Jones, Davies and Park (1994) や小暮・寒河江 (2000) 見られたい。

4.3 Nadaraya=Watson 法と移動平均

ノンパラメトリック回帰では、離散的なデータを平滑化することにより回帰関数を推定する。その平滑化を行う際に、カーネル関数による畳み込みを用いる手法を一般にカーネル法という。カーネル関数は、図4に示されるような原点0を中心とした左右対称でかつ面積が1となる実数値関数

$$K(z) = K(-z); \quad \int_{-\infty}^{\infty} K(z) dz = 1$$

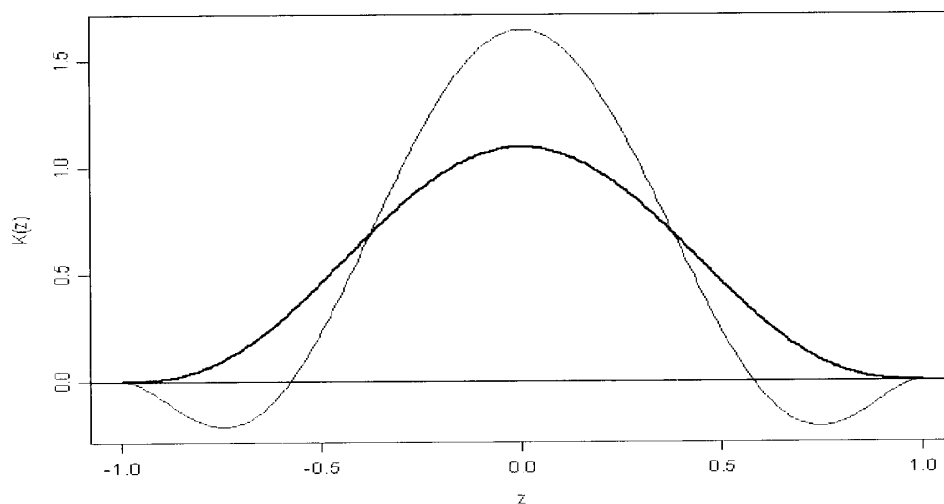


図4 カーネル関数の例

である。さらに、

$$\int_{-\infty}^{\infty} z^j K(z) dz = 0 \quad j=1, 2, \dots, p-1; \quad \int_{-\infty}^{\infty} z^p K(z) dz \neq 0$$

という条件を満たすカーネル関数を p 次カーネル関数という。カーネル法による最も代表的な推定量は、

$$\hat{q}(x) \equiv E[\hat{q}|x] \equiv \sum_{j=1}^n \hat{q}_j \times \frac{K_h(x-x_j)}{\sum_{i=1}^n K_h(x-x_i)} \quad (7)$$

である。これを Nadaraya=Watson 推定量という、 K_h は

$$K_h(\cdot) \equiv K(\cdot/h)/h$$

であり、 h はバンド幅と呼ばれる正値の平滑化パラメータである。特に

$$K(z) = 0 \text{ for } |z| \geq 1; \quad h = r\Delta$$

と選ぶと、(7)の右辺は

$$\sum_{j=x-r}^{x+r} \frac{K_h(j)}{\sum_{i=1}^n K_h(x-x_i)} \hat{q}_i = \sum_{j=x-r}^{x+r} a_j \hat{q}_{x+i}$$

となる。ただし

$$a_j \equiv \frac{K_h(j)}{\sum_{i=1}^n K_h(x-x_i)}$$

である.

移動平均をカーネル平滑化の特殊ケースと考えることによって, カーネル法の理論を用いることができる. 特に, Greville の 3 次 13 項式移動平均 (生保標準生命表) は, 「 K として 4 次カーネル関数を用いたカーネル推定量, 13 項式は, 最近傍の 13 個の観測値を用いるバンド幅に対応する」と理解できる.

5. 局所回帰法による補整

Nadaraya=Watson 法による推定では, x が 0 に近い境界部分ではバイアスが大きくなるという境界バイアス問題が存在する⁹. これを防ぐために, 境界部分では, 通常のカーネル関数と異なる「境界修正カーネル関数」を用いることも提案されている (Gabin, Haberman and Verrall, 1995). しかし, このような修正はいささかアドホックであり, より自然で統一的な平滑化が望まれる. 本節では, Nadaraya=Watson 法に代わる新たなノンパラメトリック手法である局所回帰法による補整について述べる.

5.1 局所回帰法

点 x における平均死亡関数 $q(x)$ のテイラー近似

$$q(x) \approx \beta_0 + \beta_1(t-x) + \cdots + \beta_p(t-x), \quad z-h < t < x+h \quad (8)$$

を考える. ここで, $\beta_j \equiv q^{(j)}(x)/j! (j=0, 1, \dots, p)$ である. 局所回帰法では, (8) の右辺の多項式を局所的にあてはめる. 以下の最小化問題:

$$\sum_{i=1}^n K_h(x_i-x) [\hat{q}_i - (\beta_0 + \beta_1(x_i-x) + \cdots + \beta_p(x_i-x)^p)]^2$$

を $\beta \equiv (\beta_0 \ \beta_1 \ \cdots \ \beta_p)'$ に関して最小化せよ.

の解を $\hat{\beta} \equiv (\hat{\beta}_0 \ \hat{\beta}_1 \ \cdots \ \hat{\beta}_p)'$ とするとき, 局所回帰推定量は

$$\hat{q}(x) = \hat{\beta}_0$$

で与えられる. Nadaraya=Watson 推定量が「移動平均」の拡張であったのに対して, 局所回帰推定量は「移動最小 2 乗推定量」の一般化であるといえる. $p=0$ ならば, 局所回帰推定量は Nadaraya-Watson 推定量に一致する. $p \geq 1$ の局所回帰推定量では, データに適応して β 's が選ばれるため境界バイアスが自動的に修正される.

5.2 通常の回帰モデルとの類似

局所回帰推定量は漸近的には通常のカーネル推定量と同値な表現が可能であり, カーネル法に比べて漸近的により有効であるとは主張できない. しかし, 境界バイアス問題を自動的に解決するという利点に加え, 通常のパラメトリック回帰分析の概念を用いることができるという実的な利便性を持つ.

被説明変数ベクトルと説明変数行列を

⁹ 境界バイアス問題については, Wand and Jones (1995) の 5.5 節を参照されたい.

$$Q \equiv \begin{bmatrix} \dot{q}_1 \\ \vdots \\ \dot{q}_0 \end{bmatrix}, \quad X_x \equiv \begin{bmatrix} 1 & (x_1 - x) & \cdots & (x_1 - x)^p \\ \vdots & \vdots & & \vdots \\ 1 & (x_n - x) & \cdots & (x_n - x)^p \end{bmatrix}$$

とすると, $\hat{\beta}$ は, 通常の線形回帰推定量と同じように,

$$\hat{\beta} = P(x)Q$$

と表わせる. ここで, $P(x)$ は射影行列

$$P(x) \equiv (X_x' W_x X_x)^{-1} X_x' W_x$$

である. ただし

$$W_x \equiv \begin{bmatrix} K_h(x_1 - x) & 0 & \cdots & 0 \\ 0 & K_h(x_2 - x) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & K_h(x_n - x) \end{bmatrix}$$

である. 従って, $e_1' \equiv [1 \ 0 \ \cdots \ 0]$ とすると, 局所回帰推定量は

$$\hat{q}(x) = \hat{\beta}_0 = e_1' P(x)Q$$

と表わせる.

このとき回帰ベクトルは, 線形回帰分析の自然な拡張として

$$\hat{Q} \equiv \begin{bmatrix} \hat{q}(x_1) \\ \vdots \\ \hat{q}(x_n) \end{bmatrix} = \begin{bmatrix} e_1' P(x_1) \\ \vdots \\ e_1' P(x_n) \end{bmatrix} Q$$

と定義できるから, ハット行列 H と残差ベクトル ϵ を

$$H \equiv \begin{bmatrix} e_1' P(x_1) \\ \vdots \\ e_1' P(x_n) \end{bmatrix}, \quad \epsilon \equiv Q - \hat{Q} = (I - H)Q$$

と定義できる. さらに, 局所回帰モデルの自由度は, パラメータの個数の概念の拡張として

$$m \equiv \text{tr}(H) \quad (9)$$

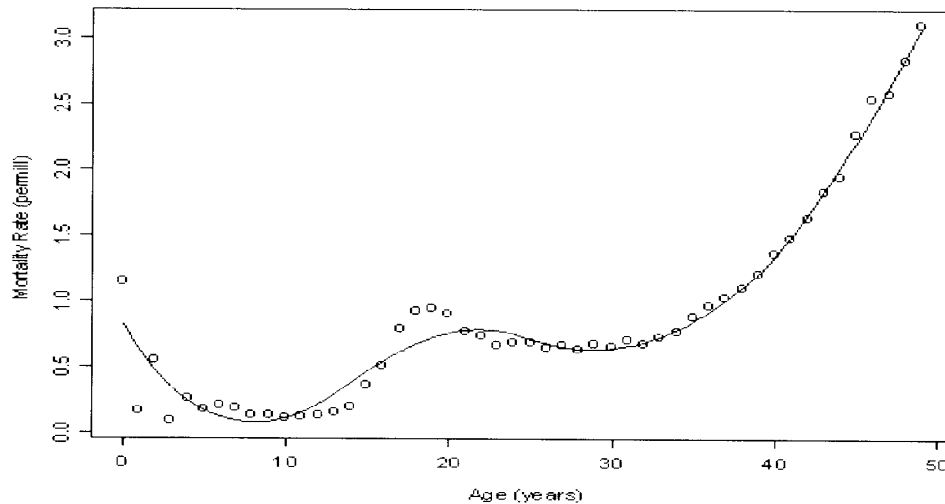
と定義できる.

5.3 クロスバリデーション

局所回帰モデルを用いるには, カーネル関数 K , バンド幅 h , 次数 p の3種類の選択を行わなければならない. 図5は, $K = \text{Epanechnikov}$, $p = 5$, $h =$ 最近傍の65%の観測値と選択したときの局所回帰推定を表している¹⁰. 近年いかなる選択が望ましいかについては様々な研究成果が報告されているが, 多くは漸近理論に基づくものである¹¹. ここでは, 通常の回帰分析における選択基準としてよく用いられる GCV 基準を用いる. ある選択を行ったときの平均平方

¹⁰ 本稿の統計計算は, Loader (1999) による Locfit を用いた.

¹¹ 例えば, Jones, Marron and Sheather (1996) を見よ.

図5 局所回帰推定 ($K=Triweight$, $p=3$, $h=$ 最近傍の 50)

誤差のクロスバリデーションによる推定量は

$$CV \equiv \frac{1}{n} \sum_{i=1}^n (\hat{q}_i - \hat{q} - i(x_i))^2$$

である。ここで、 $\hat{q}_{-i}(x)$ は i 番目の観測値を除いた $(n-1)$ 個の観測値からの $q(x_i)$ の推定値である。このとき、 n が十分大きいときは、 CV は以下の GCV とほぼ等しい。

$$\begin{aligned} CV \approx GCV &\equiv \frac{n}{(n-m)^2} \sum_{i=1}^n (\hat{q}_i - \hat{q}(x_i))^2 \\ &= \frac{n}{(n-m)^2} \epsilon' \epsilon \end{aligned}$$

となる。

図 6a, 6b は、 GCV による次数の選択を示している。 K には Epanechnikov を用いている。横軸は、(9) で与えられるフィットされたモデルの自由度を指し、異なるバンド幅 h に対応している。ここでは、 h は、最近傍の観測値の $\alpha=20\%$ から $\alpha=80\%$ までの 5% 刻みで動かしたときの GCV の値が示されている。 h の上昇とともに、 m は増加する。 GCV は、次数 $p=5$ における $h=45\%$ に対応する自由度 m に対して最小値を取る。

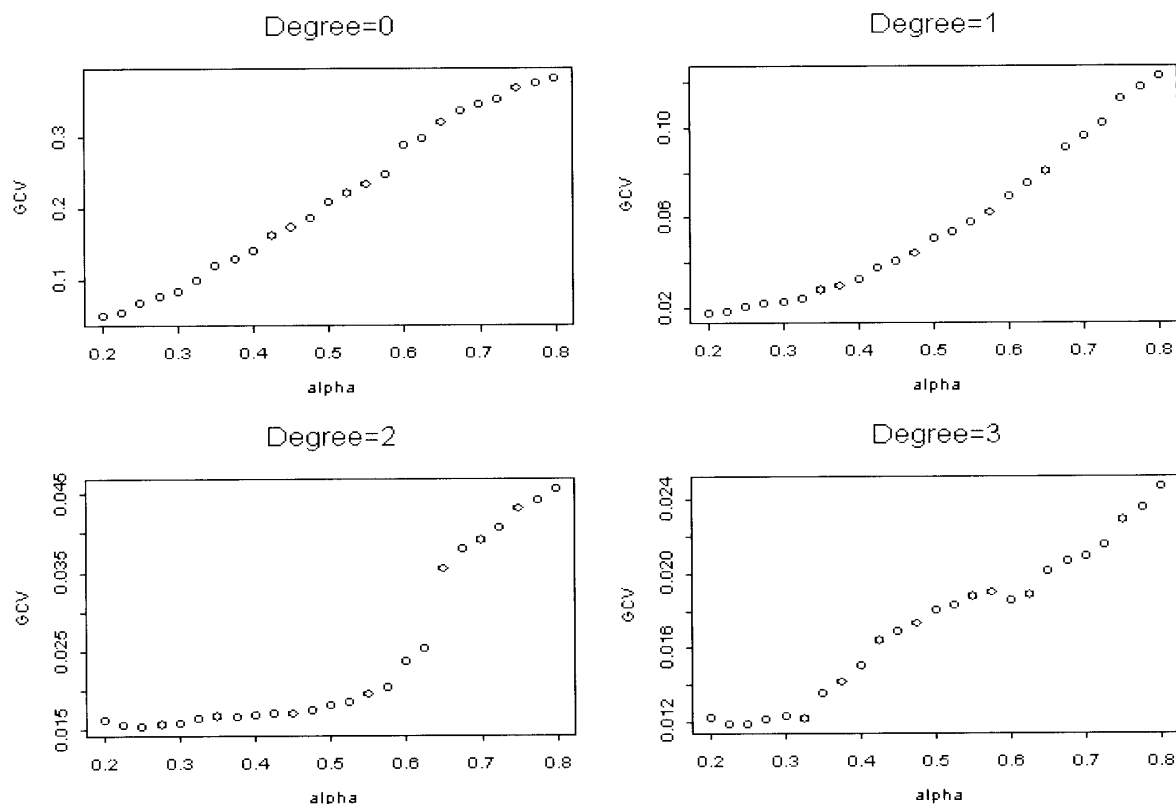
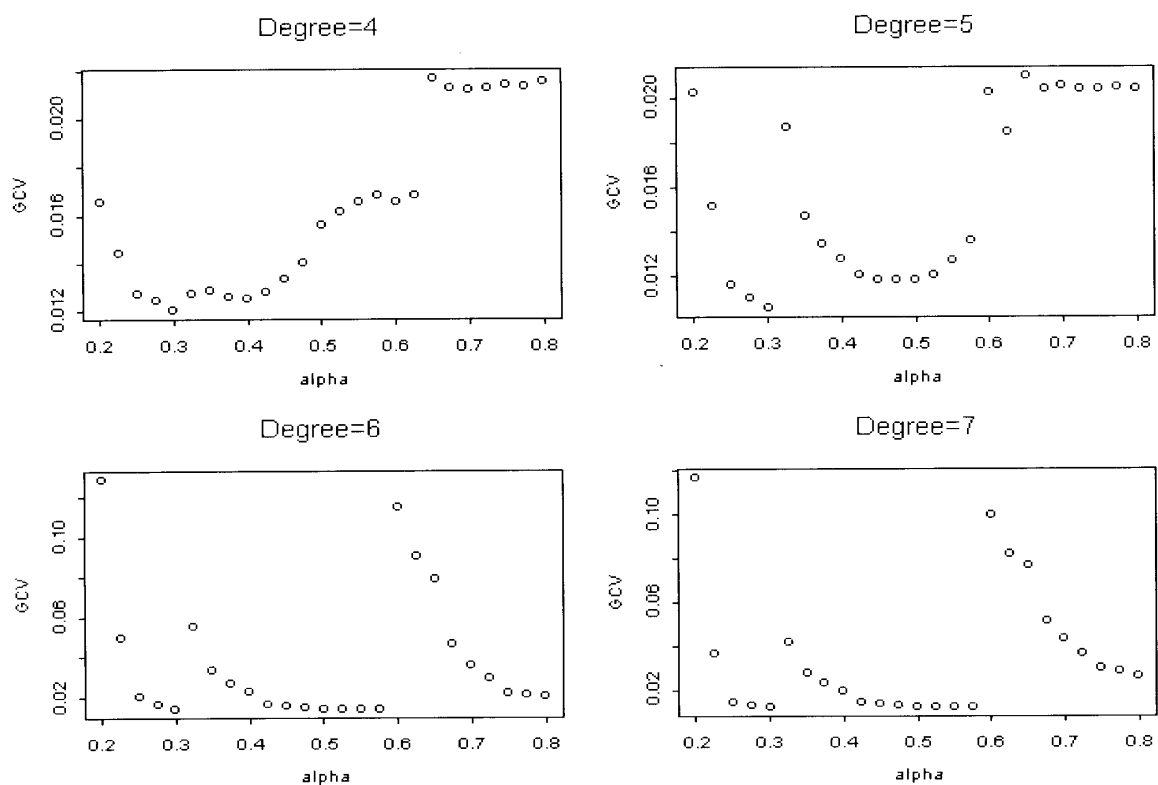
次にカーネル関数の選択を考える。図 7 は以下の 4 種類の代表的なカーネル関数に対する GCV の値を示している。次数は $p=5$ と固定している。

- 一様分布

$$K(z) = \frac{1}{2} 1_{\{|z| < 1\}}$$

- 標準正規分布

$$K(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}$$

図 6a GCV による次数とバンド幅の選択 ($p=0, 1, 2, 3$)図 6b GCV による次数とバンド幅の選択 ($p=4, 5, 6, 7$)

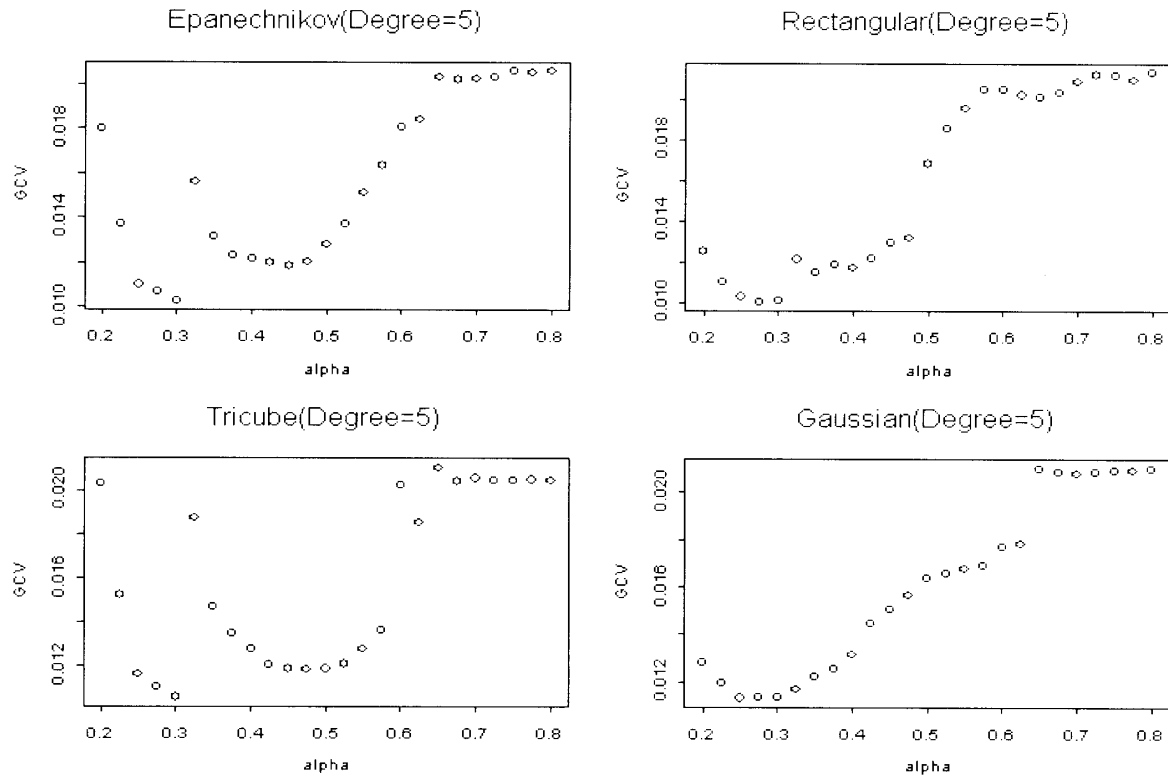


図7 GCV によるカーネル関数の選択

- Epanechnikov

$$K(z) = \frac{3}{4} (1 - z^2) 1_{\{|z| < 1\}}$$

- Triweight

$$K(z) = \frac{35}{32} (1 - z^2)^3 1_{\{|z| < 1\}}$$

図7を見る限り大きな差はないように思われる．これは， h の選択に比べ K の選択の推定効率にはロバストであるというよく知られた理論的結果とも一致している．

以上の結果， $K = \text{Epanechnikov}$, $p = 5$, $h = \text{NN}(0.45)$ が選択された．ここで， $\text{NN}(0.45)$ は最近傍の 45% の観測値を用いるという意味である．図8は，選択されたモデルを生保標準死亡率と比較したものである．ここで，局所回帰推定値は， $\hat{q}(x)$ そのものではなく，それに生保標準表における 1 次補正に対応する調整を加えた

$$\hat{q}(x) + \min\left(\sqrt[2]{\frac{\hat{q}(x)(1 - \hat{q}(x))}{n_x}}, 0.3 \hat{q}(x)\right)$$

を用いている．18 歳前後及び 5 歳以下を除いて，2 つの曲線はほぼ一致している．特に，生保標準死亡率における 18 歳におけるピークは局所回帰推定値でも明確に現れている．

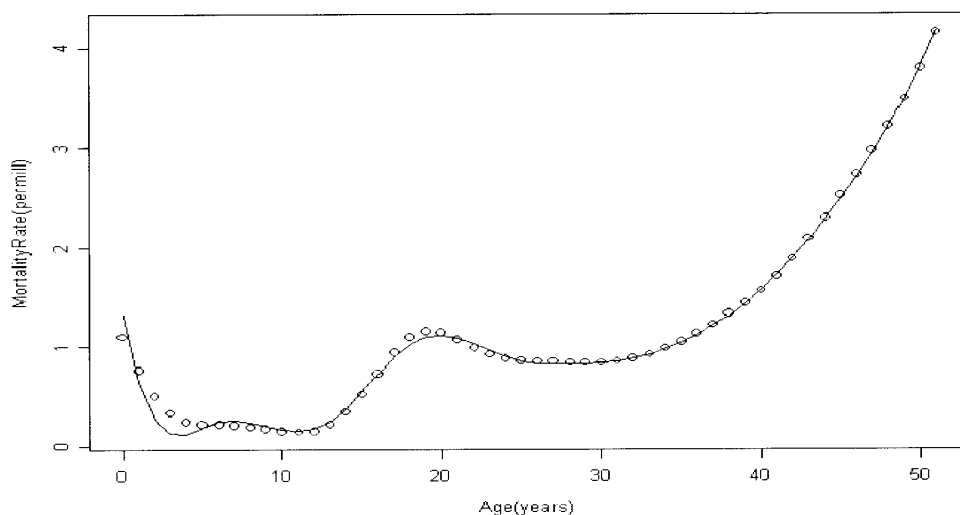


図8 標準生命表死亡率 VS 局所回帰補整値

6. 加重局所回帰法による補整

図9は、選択されたモデルの残差のプロットである。残差は明白な分散不均一性を示している¹²。図10は、オリジナルな粗死亡率の標本標準偏差のプロットである。このような分散不均一性に対処するために加重局所回帰法を考察する。

死亡率データは、 $\hat{q}_i \equiv d_i/e_i$ と定義された。ここで、死亡者数 d_i が、成功回数が e_i 、成功確率が $q(x_i)$ の2項分布に近似的に従うとみなせるから

$$\text{Var}(\hat{q}_i) \approx \frac{q(x_i)(1-q(x_i))}{e_i} \quad (10)$$

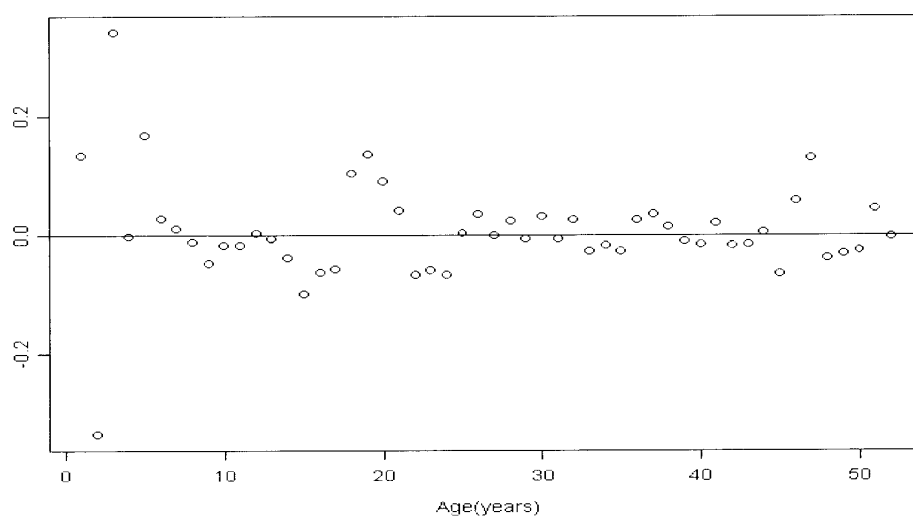


図9 推定されたモデルの残差

¹² 図9からは、隣接する年齢間の正の相関関係の可能性も示されているが、本稿ではこれについては考察しない。

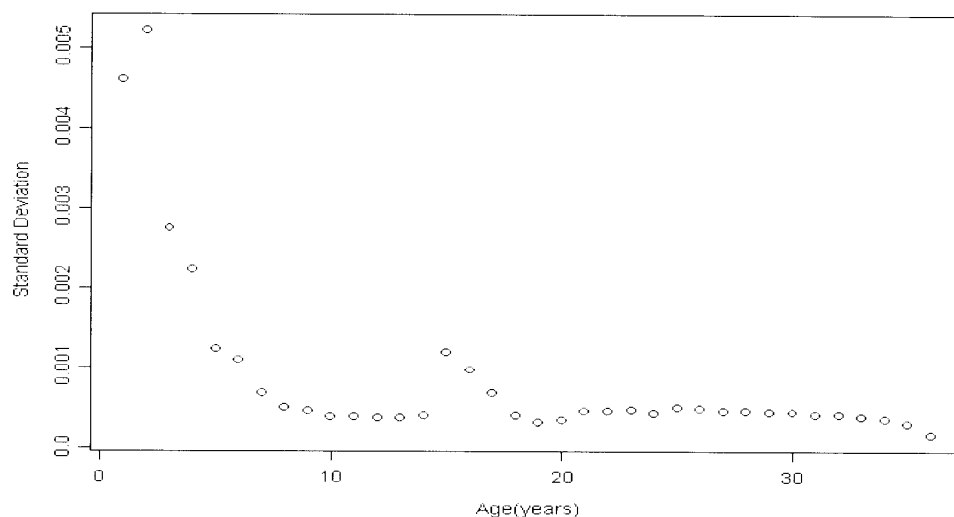


図 10 粗死亡率の標準偏差

となる．従って，この分散不均一性を考慮に入れた加重を用いた局所回帰法を用いることが考えられる．ここでは，直接的に (10) の右辺の分散を加重として用いるのではなく，まず被説明変数である死亡率を

$$Y_i \equiv \sin^{-1} \sqrt{\hat{q}_i}$$

と変換してから加重最小 2 乗法を行う．この変換の結果

$$\text{Var}(Y_i) \approx \frac{1}{e_i}$$

となり，被説明変数がパラメータの値に依存しなくなる（分散安定化変換）．以下の最小化問題：

$$\sum_{i=1}^n e_i K_h(x_i - x) [Y_i - (\beta_0 + \beta_1(x_i - x) + \cdots + \beta_p(x_i - x)^p)]^2$$

を β に関して最小化せよ．

の解を $\hat{\beta}$ とするとき，加重局所最小 2 乗法による推定値は

$$\hat{q}(x) \equiv \sin^2(\hat{\beta}_0)$$

と与えられる．図 11 は，この加重局所最小 2 乗法の推定値を示している．クロスバリデーションによって， $K = \text{Epanechnikov}$, $p = 6$, $h = \text{NN}(0.30)$ が選択された．

7. 局所尤度法による補整

加重局所モデルは，死亡者数 d_i が，成功回数が e_i ，成功確率が $q(x_i)$ の 2 項分布に近似的に従うという想定の下で分散の調整を行った．このアプローチの自然な拡張は，この 2 項モデルの情報を分散だけに限定せずに尤度全体を用いることである． d_i が試行回数 e_i ，成功（死亡）確率 $q(x_i)$ の 2 項確率分布に従うとき，死亡率の対数尤度は

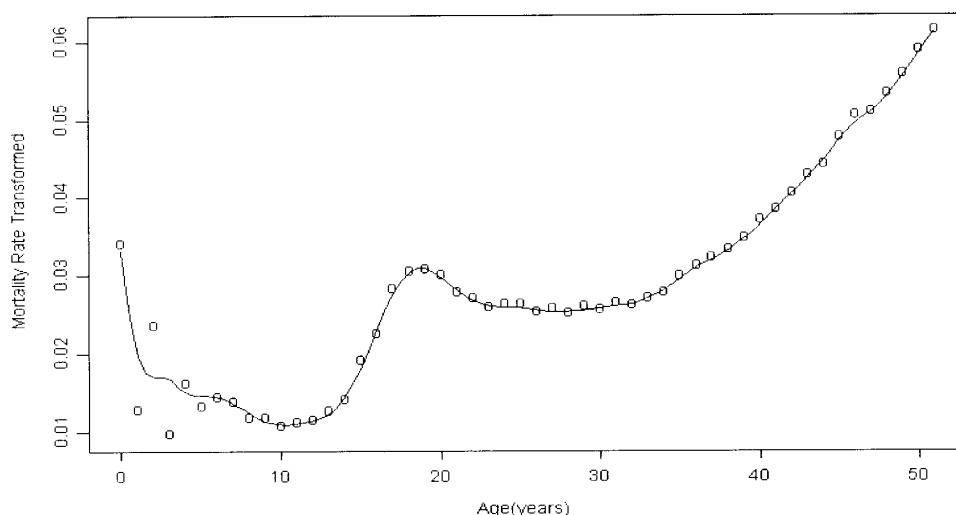


図 11 加重局所最小 2 乗法によって選択されたモデル

$$\begin{aligned}\log \Pr\left(\hat{q}_i = \frac{d_i}{e_i}\right) &\propto d_i \log \frac{q(x_i)}{1 - q(x_i)} + e_i \log (1 - q(x_i)) \\ &= d_i \theta(x_i) - e_i \log (1 - \exp \{\theta(x_i)\})\end{aligned}$$

となる。ただし、一般化線形モデルに倣って、 $\theta(x)$ は $q(x)$ のロジット変換変換

$$\theta(x) = \log \frac{q(x)}{1 - q(x)}$$

である。

点 x における θ のテイラー展開

$$q(x) \approx \beta_0 + \beta_1(t-x) + \cdots + \beta_p(t-x)^p = B'x$$

を考える。以下の最大化問題：

$$\sum_{i=1}^n K_h(x_i - x) [d_i \beta' x - e_i \log(1 + \exp \{\beta' x\})]$$

を β に関して最小化せよ。

の解を $\hat{\beta}'(x) \equiv [\hat{\beta}_0 \ \hat{\beta}_1 \ \cdots \ \hat{\beta}_p]$ とするとき、局所最尤推定値は

$$\hat{q}(x) = \frac{\exp \{\hat{\beta}_0\}}{1 + \exp \{\hat{\beta}_0\}}$$

と与えられる。

尤度を基準とするときの自然な損失関数はカルバック＝ライブラー情報量である。ある観測値 (x_i, \hat{q}_i) に対するある選択を行ったときのカルバック＝ライブラー情報量に基づく乖離度を

$$D(\hat{q}_i, \theta(x_i)) \equiv 2 \left(\sup_{\theta} l(\hat{q}_i, \theta(x_i)) - l(\hat{q}_i, \theta(x_i)) \right)$$

と定義する．ここで，

$$l(d_i|\theta) \equiv d_i \theta - e_i \log(1 + \exp\{\theta\})$$

とする．このとき，期待乖離度のクロスバリデーションによる推定量は

$$CV(\hat{\theta}(x)) \equiv \sum_{i=1}^n D(\dot{q}_i, \hat{\theta}_{-i}(x_i))$$

である．ここで， $\hat{\theta}_{-i}(x_i)$ は， i 番目の観測値を除いた残りの $n-1$ 個の観測値からの $\theta(x_i)$ の局所尤度推定値である．このとき，

$$CV(\hat{\theta}(x)) \approx AIC \approx \sum_{i=1}^n D(\dot{q}_i, \hat{\theta}(x_i)) + 2m$$

である．この基準の結果， $K = \text{Epanechnikov}$, $p = 3$, $h = \text{NN}(0.45)$ が選ばれた．生保生命表の補整値では対応する値はその結果は図 12 に与えられている．これは加重局所最小 2 乗法よりも局所最小 2 乗法に近い結果となっている．

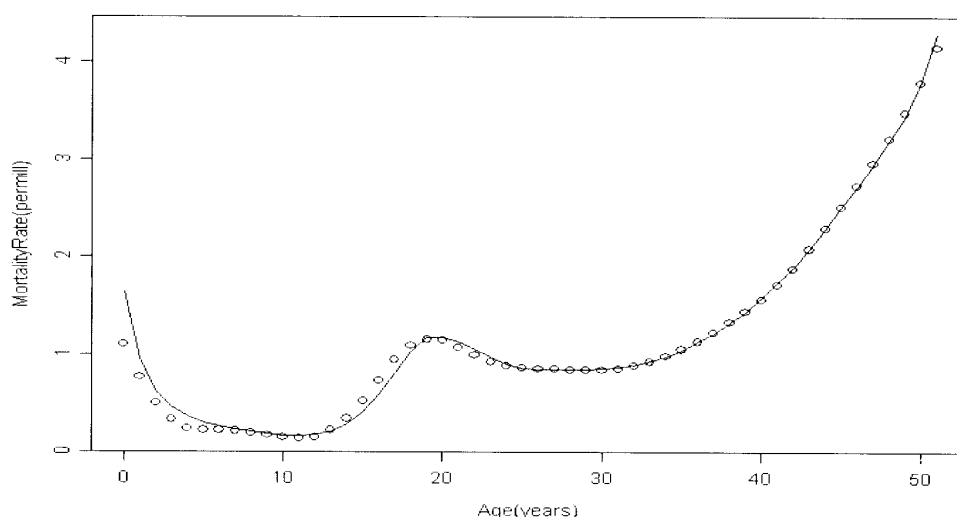


図 12 局所尤度法による補整

8. 終わりに

本論文では，ノンパラメトリック回帰の立場から生命表における補整について考えた．男子の生保標準生命表を例として取り上げ，新たなノンパラメトリック手法による補整を試みた．20 以上の期間では生保標準生命表の予定死亡率と一致する傾向が見られたが，18 歳前後と 5 歳以下の 2 つの期間で一定の乖離が見られた．男子 18 歳の時期における死亡率の特徴とされる「こぶ」は，ノンパラメトリック補整では低めに推計され，一方 5 歳以下の期間では高めに推計された．本稿ではバンド幅の選択を（生保標準生命表の場合と同様に）年齢 x に依らず一定として行った．そのため，18 歳の近傍の他の年齢のデータの影響を強く受けすぎているのかもしれない．この点については，今後の研究において，バンド幅を局所的にあてはめるアプローチを試みたい．

また，理論的改良点としては，局所尤度法において用いた 2 項分布の尤度が適切でないかも

しれないという指摘ができる．例えば， d_x が平均が $\lambda_x \equiv \mu_{x+0.5} \times e_x$ のポアソン分布に従い， e_x がパラメータが $\alpha = d_x$, $\beta = \mu_{x+0.5}$ のガンマ分布に従う．というポアソン-ガンマ・モデルの方が適切かもしれない．ここで， $\mu_{x+1/2}$ は $x+0.5$ 歳の死力

$$\mu_{x+1/2} \approx \log(1 - q(x))$$

を表わす．この点も将来の研究課題としたい．

参 考 文 献

- [1] Anderson, T. W. (1971). *The Statistical Analysis of Time Series*, Wiley.
- [2] 荒井 昭 (2001). 生命に関する一考察—生命関数の数式近似—, 日本アクチュアリー会会報 **54**, 85-121.
- [3] Copas, J and Haberman (1983). Non-parametric Graduation Using Kernel Methods, *Journal of the Institute of Actuaries*, **118**, 135-156.
- [4] Fan J. and Gijbels, I (1996). *Local Polynomial Modelling and Its Applications*, Chapman and Hall.
- [5] 二見 隆 (1992). 生命保険数学, 生命保険文化研究所.
- [6] Gavin, J. B., Haberman, S., AND Verral, R. J. (1994). On the Choice of Bandwidth for Kernel Graduation, *Journal of the Institute of Actuaries* **121**, 119-34.
- [7] Gavin, G., Haberman, S. and Verrall, R. (1995). Graduation by Kernel and Adaptive Kernel Methods with a Boundary Correction, *Transactions of Society of Actuaries*, **47**, 173-209.
- [8] Gerber, H. (1991). *Life Insurance Mathematics*, Springer.
- [9] Greville, T. (1974). *Graduation*. Part 5 Study Note 53. 1. 73. Chicago: Society of Actuaries.
- [10] Jones, M. C., Davies, S. J., and Park, B. U. (1994). Versions of Kernel-Type Regression Estimators, *Journal of the American Statistical Association*, **89**, 825-32.
- [11] Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A Brief Summary of Bandwidth Selection for Density Estimation, *Journal of the American Statistical Association*, **91**, 401-407.
- [12] 小暮厚之, 寒河江雅彦 (2000). ノンパラメトリック統計モデル推定の最近の展開, 日本統計学会誌, **30**, 265-280.
- [13] Loader, C. (1999). *Local Regression and Likelihood*, Springer.
- [14] Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*, Chapman and Hall.
- [15] 山口喜一他 (1994). 生命表研究, 古今書院.