Link, W. A. and R. J. Barker (2010) *Bayesian Inference: with ecological applications*
輪読会 2010/12/18-19
担当： 深谷肇一 kfukaya@ees.hokudai.ac.jp

Chapter 13: Autoregressive Smoothing

パラメーターの非線形な時間変化を平滑的に表す解析がしたい。

→ **Autoregressive model**

パラメーター値そのものでなく、パラメーター値の時間変化を独立な確率変数と考える。

13.1 DOVE DATA AND PRELIMINARY ANALYSES

ナゲキバト(mourning dove)の標識採捕データ (Table 13.1)

1968 年から 1973 年にかけて標識された個体の採捕状況を 1968 年から 1976 年まで観測。
生存率の時間変化を知りたい。

モデル

i = 1, 2, …, 6 : コホートに関連する添字

j = 1, 2, …, 9 : 再捕獲年に関連する添字

M_i: i 年コホートの個体数

X_{ij}: j 年に捕獲された i 年コホートの個体数

f_j: j 年に個体が生存している場合に、再捕獲される条件付き確率

S_j: j 年に個体が生存している場合に、j+1 年まで生存する条件付き確率

π_{ij}: i 年コホートの個体が j 年に再捕獲される確率

1. 生存 1 年目の再捕獲率

$$\pi_{ii} = f_i$$

2. 生存 2 年目以降の再捕獲率

$$\pi_{ij} = f_j \times \prod_{k=i}^{j-1} S_k \qquad (13.1)$$

尤度

$$\prod_{i=1}^{6}\prod_{j=i}^{9}\pi_{ij}^{X_{ij}} \times \prod_{i=1}^{6}\left(1-\sum_{j=i}^{9}\pi_{ij}\right)^{M_i-\sum_{j=i}^{9}X_{ij}}$$

（深谷：本文では Σ の添字に誤植？）

パラメーターの最尤推定値 (Table 13.2)

$S_1, S_2, \ldots, S_5$ の推定はけっこうひどい。

だいたい 25 回のベルヌーイ試行からの推定と同じくらいの精度。

f_7, f_8, f_9 の推定値がないので、S_6, S_7, S_8 の推定値は得られない。

f_7, f_8, f_9 を $\bar{f}$ （$f_1, \ldots, f_6$ の平均）などと仮定して推定することもできるが、推定結果は仮定に敏感。

→ S_6, S_7, S_8 の推定には何らかの事前情報が必要。

予備解析モデル (Panel 13.1)

f_i がある共通の分布からの標本であると仮定。

→ $\text{logit}(f_i) \sim \text{N}(\mu_f, 1/\sigma_f^2)$

モデルへのコメント

1. π_{ij} を再帰的に定義(p. 290 参照)することもできる

→ 掛け算の数が減って一見有効なコーディングに見えるが、実際は実行時間が 2 倍になるほど効率が悪い。

経験的に、BUGS でパラメーターの再帰的な表記方法は避けたほうがよい。

2. モデルの表記で用いられないパラメーター($\pi_{3,1}$ や $\pi_{4,3}$)は定義しなくてもよい。

結果 (Figure 13.1)

生存率は減少していると言えるだろうか？

S4 > S5 の事後確率 → 24%

S7 > S8 の事後確率 → 25%

S1 < S2 の事後確率 → 94%

→ 最初増加して、その後減少が続いていたのかもしれない。

13.2 MODELING DIFFERENCES IN PARAMETER VALUES

生存率の時間変化を表すための変数変換

生存率の時間変化を表すのに、変化の絶対値(absolute change)も相対値(proportional change)も便利ではない。

ロジット変換された生存率(θ_j)における変化を考える。

$$\theta_j = \text{logit}(S_j) = \log(S_j/(1-S_j))$$

θ_j の一次差分(Δ_j)は以下のとおり。

$$\Delta_j = \theta_j - \theta_{j-1}, \quad j = 2, 3, \ldots, n \qquad (13.2)$$

Δ_j が独立に、平均 0、分散 σ_Δ^2 の同一の分布に従う(iid)確率変数であると仮定。

→ 時間 j と k ($j > k$)におけるロジット生存率の差は

$$\theta_j - \theta_k = \Delta_{k+1} + \cdots + \Delta_j$$

となり、その分散は $(j-k)\sigma_\Delta^2$。

→ σ_Δ^2 がそれほど大きくないとき、隣り合う年の生存率は類似。時間が経つにつれて生存率に大きな違いが出てくる。

生存率の差分モデル (Panel 13.2)

Δ_j が独立に、平均 0、分散 σ_Δ^2 の同一の分布に従う(iid)確率変数であるということは、

→ $[\theta_j \mid \theta_{j-1}] = \mathrm{N}(\theta_{j-1}, 1/\sigma_\Delta^2)$, for j = 2, 3, …, 8

（深谷：BUGS コードに誤植？）

技術的なノート：なぜ Panel13.2 で $0.0001\tau_\theta$ を使うのか

省略します。

13.3 RESULTS FOR DOVE ANALYSIS

差分モデルの結果(Figure 13.2)

特徴 1. 変動のパターンがかなり平滑化された。

特徴 2. 信用区間はずっと短くなった。

Noisy なデータと Bayesian shrinkage

データがノイジーな場合におけるベイズ推定

→ 事前分布の影響を強く受ける。

ナゲキバトのノイジーな生存率データ (Figure 13.2)

→ **Bayesian shrinkage**: 事前分布によって表される類似性が強調され、パラメーター推定値が互いに収縮。

データの情報量が増えると shrinkage は解消する (Figure 13.3)

→ データ量が 3 倍（緑）、10 倍（青）になった場合の事後分布

13.4 HIGHER ORDER DIFFERENCES

一次差分について見てきたので、二次差分（Δ_j^2）についても考えてみる (Panel 13.3)

$$\Delta_j^2 = \Delta_j - \Delta_{j-1} = (\theta_j - \theta_{j-1}) - (\theta_{j-1} - \theta_{j-2}) = (\theta_j - 2\theta_{j-1} + \theta_{j-2}), \text{ for } j = 3, 4, \ldots, n \tag{13.3}$$

Δ_j^2 が独立に、平均 0、精度 τ の同一の分布に従う(iid)確率変数であると仮定。

→ $[\theta_j \mid \theta_k] = \mathrm{N}(2\theta_{j-1} - \theta_{j-2}, \tau), \quad (k < j)$

（深谷：$k < j$ の意味がよくわかりませんでした。）

二次差分を使った解析の例

Breslow and Clayton (1993)

アイスランドにおける出生年と乳がん発症率の関係を解析

高次の差分についても考えてみる

θ の高次差分における、θ_j の分布の一般形

$$[\theta_j \mid \theta_k] = \mathrm{N}\left(\sum_{k=1}^{m}\binom{m}{k}(-1)^{k+1}\theta_{j-k}, \tau\right), \text{ for } j = m+1, m+2, \ldots, n \tag{13.4}$$

最初の m 個の θ_j について無情報な事前分布を与える。

高次差分モデルの問題点

1. 解釈が難しい
2. 推定値が不安定 (Figure 13.4)

$(1, \theta_{j-1}), \ldots, (m, \theta_{j-m})$を通る自由度 m-1 の多項式 $f(x)$ (cf., Fig. 13.4)を考える。

→ $f(0)$は θ_j の平均(cf., eq. 13.4)と一致する。

→ m 次の差分モデルは、θ_j を、$f(x)$を用いて外挿的に推定することに等しい。

（深谷：意訳しています）

Fig. 13.4 を見ると分かるとおり、6 次の多項式による推定($f(0)$)はかなり精度が低い。

MCMC のサンプリング効率も低い。

標本には強い自己相関。

ナゲキバトデータの例 (Figure 13.5)

（深谷：高次モデルでもあんまり変わらないような？？）

13.5 AFTERWORD

パラメーターの傾向(signal)を推定するモデルの選択

1. 注意深く行うこと
2. モデル選択に関連する事項は慎重に報告すること

モデルの仮定が推定に及ぼす影響

Inferential basis = Data + Prior knowledge

→ 1. 良質なデータの場合、パラメーターの事後分布はモデル化された傾向（深谷：平滑化の仮定など）の影響を受けにくい
2. データの質が悪い場合、パラメーターの事後分布はモデル化された傾向の影響を受けやすい
3. パラメーターの事後分布は、モデル化された傾向の仮定が強い（事前分布がinformative）ほど影響を受けやすい。

パラメーターのパターンを知ることは、データのパターンを知ることより難しい

1. パラメーター数はデータ数より少ない。
2. パラメーター値を完全に知ることはできない

→ パラメーターのパターンの検証は、推定に影響する検証不可能な仮定の上で行われている可能性がある。

パラメーターのパターンに誤った仮定を適用してしまった場合どうなるか？

1. データからの情報が十分に得られているパラメーターの推定にはあまり影響しない。
2. データからの情報が少ないパラメーターの事後分布や、新しいパラメーターの予測分布に影響(false confidence)する。

パラメーターのパターンを検証する場合は、データがどれくらい推定に強く影響するかをはっきりと示す必要がある。(cf., Fig. 13.2)