

Outline	
上野	9.1 Introduction
	9.2 Mark-Recapture Models and Missing data
	9.3 Time and Behavior Models
平岩さん	9.4 Individual Heterogeneity Models
岨さん	9.5 Example: Koalas
	9.6 Afterword

9.1 Introduction

Mark-recapture 法とは、捕獲と放逐を繰り返すことによって個体数を推定する方法。

cf. Seber(1982), Williams et al. (2002)

☆Mark-recapture には、欠損値 (Missing data) が存在する！

一回目のサンプリング

二回目のサンプリング

.....



☆この本では、最尤推定法に対応したベイズ的方法を示すということではなく、階層モデルとベイズ推定を使用することによる possibilities(展望?)を重要視する。

☆この章では、(1)8.1で議論された Complete Data Likelihood (CDL)の便利さを強調する。(2)データ拡大法(data augmentation method)と複数モデル推定のための階層モデリングの構築方法を提示する。

9.2 Mark-Recapture Models and Missing Data

個体番号 i \ サンプリング j 回目		1	2	3	4	5
X^{obs} u 行	1	1	0	0	1	0
	2	0	0	1	0	0
	3	0	0	1	0	1
	4	0	0	0	0	1
	5	0	0	0	0	1
X^{miss} $N-u$ 行	6	0	0	0	0	0
	7	0	0	0	0	0
	8	0	0	0	0	0
	9	0	0	0	0	0
	10	0	0	0	0	0
	.					
	N	0	0	0	0	0

9.2.1 Completing the Data

Complete Data Likelihood (CDL) で推定作業を行うのがいい。

$$\text{Complete Data} : \{X^{obs}, X^{miss}, I\}$$

$$I: \text{N-dimension inclusion vector, } I = (\underbrace{1,1,1,1,1}_{u \text{次元}}, \underbrace{0,0,0,0,0}_{N-u \text{次元}}, \dots)$$

CDL のパラメータとは I の次元数 N と、 X^{obs}, X^{miss} である。 N の条件付き確率として、 I (そして X^{miss}) をまるで観察したかのように扱うことができる。一度、 N と u が特定されれば、 I や X^{miss} に特に情報は含まれていない。

したがって、CDL は、

$$\begin{aligned} [X^{obs}, X^{miss}, I | \theta, N] &\propto [X^{obs}, u. | \theta, N] \\ &= [X^{obs} | u., \theta^{obs}] [u. | \theta, N] \end{aligned} \quad (9.1)$$

となる。

ここでは θ はパラメータの行列のことを指し、 X^{obs} については θ^{obs} 、 X^{miss} については θ^{miss} に分けられる。

本来ならば CDL から Observed Data Likelihood (ODL) に行く際に、missing data を積分する必要がある(8.1)。しかし N 依存的に X^{miss} は一定次元数のゼロマトリックスになるので、尤度から除外される。したがって ODL も (9.1) 式の右辺に比例する。

(9.1)にある式は一般的であり、実際には θ に制約をかけてモデルを特定する必要がある。

モデルは、 $[X^{obs} | u., \theta^{obs}]$ や $[u. | \theta, N]$ の項を特定し、かつ θ に制約をかけることで構築される。

閉鎖系では、移出入はゼロである。パラメータは $\theta = \{p^{obs}, p^{mis}\}$ であらわされ、 p とは捕獲率のことである。

CDL の簡易式は、

$$[X | N, p^{obs}, p^{mis}] \propto \binom{N}{u.} \prod_{i=1}^N \prod_{j=1}^k p_{ij}^{x_{ij}} (1 - p_{ij})^{(1-x_{ij})} \quad (9.2)$$

N : 推定したいもの。

p : nuisance parameters (直接の興味にはないが推論で必要なパラメータ)。

x_{ij} : {1(捕まっていた), 0(捕まっていない)}

Mark-recapture モデルは2つのモデルに大別される。

1. Time and behavior model

個体は共通のパラメータを持ついくつかのグループに明確に分けられる(捕獲された、されな
いに関わらず)。パラメータは、いつのサンプリング期だったのか、捕獲履歴、観察された共変
量 (etc. 年齢や性別) に依存する。

2. Heterogeneity model

パラメータは個体によって全て異なると考えるものの、捕獲率を同じとしたメンバーからなるグ
ループで明確に区分すべきか、捕獲率を何らかの共変量の関数として表すべきか、情報が十
分でない。捕獲しやすいグループとしにくいグループに分けて考える。

9.3 Time and Behavior Model

このモデルでは、観察された個体も、観察されなかった個体も、パラメータは共通。

つまり、 ρ^{miss} と ρ^{obs} は同じ。

ただし、 ρ^{obs} は、サンプリング期 (time) や、捕獲履歴 (behavior) によって異なるかもしれない。

閉鎖系 Mark-recapture モデルにおける4つの標準モデル

- M_0 : 捕獲率は一定 (サンプリング時期や、過去の捕獲履歴に関わらない)。
- M_t : 捕獲率は、サンプリング時期によって異なる。
- M_b : 捕獲率は、過去の捕獲履歴によって異なる。
- $M_{t,b}$: 捕獲率は、サンプリング時期や過去の捕獲履歴によって異なる。

M_b (捕獲率は過去の捕獲履歴によって異なる) モデルの場合:

個体番号 i \ サンプリング j 回目	1	2	3	4	5
2	0	0	1	0	0
捕獲率	p	p	p	c	c

個体番号 i \ サンプリング j 回目	1	2	3	4	5
5	0	0	0	0	1
捕獲率	p	p	p	p	p

捕獲率 $p >$ 捕獲率 c : trap shy

捕獲率 $p <$ 捕獲率 c : trap happy

上のように、i 番の個体が初めて捕獲されたサンプリング期を i 番目の要素とするベクトル y を設定すると、

$$[X^{obs} | u., \rho^{obs}] = [X^{obs} | y, u., c][y | u., \rho].$$

9.3.1 Gibbs Sampler for Model M_t

M_t (捕獲率はサンプリング期によって異なる) モデルを使って、どのように Gibbs サンプルングするか考えてみる。

$$[X | N, \theta] \propto \binom{N}{u.} \prod_{j=1}^k p_j^{n_j} (1 - p_j)^{N - n_j} \quad (9.3)$$

$$\theta = (p_1, p_2, p_3, p_4, \dots, p_k)$$

n_j が j 回目に捕獲された動物数。

p_j の事前分布で便利なのは、 $(\alpha_\rho, \beta_\rho)$ のベータ分布である。N の事前分布が p に依存しないならば、フル条件付き分布を導く (他の全てのパラメータの条件下での分布)。

$$[p_j | \cdot] \propto \rho_j^{n_j + \alpha_\rho - 1} (1 - \rho_j)^{N - n_j + \beta_\rho - 1} \quad \text{これは } \beta \text{ 分布。}$$

N の事前分布は、 (α_N, β_N) 負の二項分布が一般的である。そもそもは平均が (α_N, β_N) のガンマ分布から生成されたもののポアソン分布であるべきだが、ポアソン分布だと、平均と分散が等しいという厳しすぎる制約がかかるため、負の二項分布にした。

N のフル条件付き分布を導き出すために、9.3式から始めて、それを事前分布と掛け合わせる。

$$[N|\cdot] \propto [X|N, \theta][N|\alpha, \beta] \\ \propto \binom{N}{u} \prod_{j=1}^k p_j^{n_j} (1-p)^{N-n_j} \times \frac{\Gamma(N+\alpha)}{\Gamma(N+1)\Gamma(\alpha)} \left(\frac{\beta}{1+\beta}\right)^\alpha \left(\frac{1}{1+\beta}\right)^N$$

ここで、 $N=u+U$ と置き換える。 U は非マーク個体の総数。変数変換の定理(2.2.4)によって、そして U を含まない項を無視することで、 $[U|\cdot]$ のフル条件付き分布を得る。

$$[U|\cdot] \propto \frac{\Gamma(u+U+\alpha)}{\Gamma(U+1)} \left(\frac{\prod_{j=1}^k (1-p_j)}{1+\beta}\right)^U \quad (9.4)$$

もう少し努力すると、(9.4)が U の (a,b) の負の二項分布の核心部になっていることが分かる。

$$a = u + \alpha_N \\ b = \frac{1 + \beta_N - \prod_{j=1}^k (1-p_j)}{\prod_{j=1}^k (1-p_j)}$$

それゆえ、 (a,b) の負の二項分布から U を取り出すことで、フル条件付きの $[N|\cdot]$ から乱数を生成をする。そして、 N は $u+U$ から得られる。

ギブズサンプリングは以下のように行う。

Step1: 初期値のセットから始める。 $\{p_j^{(0)}\}$ を使って、 $U^{(1)}$ を負の二項分布 (a,b) から生成する。

$a = u + \alpha_N$ 、 $b = \frac{1 + \beta_N - \pi_0}{\pi_0}$ である。

Step2: つぎに、 $p_j^{(1)}$ をベータ分布 $(n_j + \alpha_p, N^{(1)} - n_j + \beta_p)$ から生成する。

Step3: 1と2を繰り返して、 $N^{(i-1)}$ に条件依存的な $p^{(i)}$ と $p^{(i)}$ に条件依存的な $N^{(i)}$ を生成する。

ギブズサンプリングを実施するためには、 p と N に関する事前情報を特徴づける、 $\alpha_p, \beta_p, \alpha_N, \beta_N$ について決定する必要がある。objective 分析では、 $\alpha_p = \beta_p = 1$ (一様事前分布)か、 $\alpha_p = \beta_p = 1/2$ (Jeffreys 事前分布)を設定することができる。


N に関する Jeffreys 事前分布では、すべてのパラメータが固定されている場合、 $[N] \propto 1/N$ で与えられ、 $\alpha_N = \beta_N = 0$ になる。一方、 N に関する一様事前分布は $\alpha_N = 1, \beta_N = 0$ になる。どちらの事前分布も非正則であるが、それらは正則な事後分布を導き、無情報事前分布としては、よい候補である。正則な一定の一様分布を考慮することで、一様分布を使うことができる、その場合、 $f(N)=1/v$ である ($N=1,2,\dots,v$)。 v については、 N に比較して大きく、 N のフル条件付き分布は上に示した二項分布に近似される。 v を大きくするほど、近似できる。

非正則: 積分しても1にならない(第6章参照)

9.3.2 Example: Adult Female Meadow Voles

Williams et al. (2002)は meadow voles(アメリカハタネズミ)を対象とした研究から分析方法を提示している。

捕獲個体数 $u. = 52$
 capture vector, $n = (27, 23, 26, 22, 23)$
 burn-in. = 1000
 Markov chain 数 = 100,000
 2種類の初期値: $p_j^{(0)} = 0.9, p_j^{(0)} = 0.3$
 事前分布: 非正則 Jeffreys 分布 $\alpha_N = \beta_N = 0$, p については $Be(0.5, 0.5)$



結果 (Fig. 9.3)

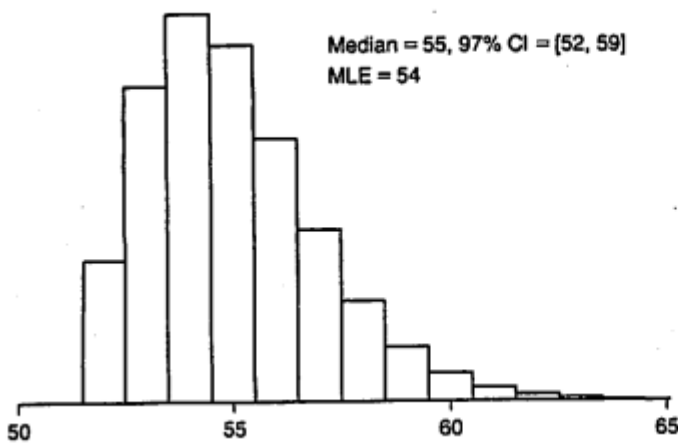


FIGURE 9.2 Posterior distribution for N in the meadow vole example with $Be(1/2, 1/2)$ priors for each p_j and an improper negative binomial $NB(0, 0)$ prior on N .

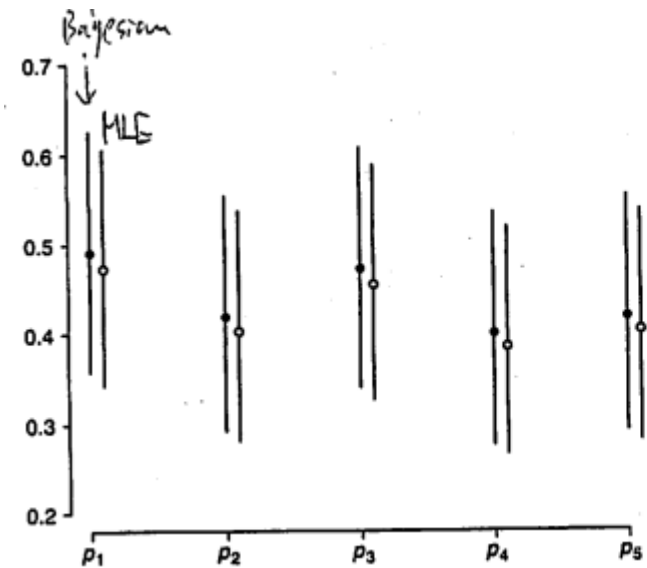


FIGURE 9.3 Posterior summaries for the capture probabilities p_j in the meadow vole example with $Be(1/2, 1/2)$ priors for each p_j and an improper negative binomial $NB(0, 0)$ prior on N . The closed circles are the posterior medians and the lines denote 95% credible intervals. The open circles denote the MLE's and the associated lines asymptotic 95% confidence intervals.

Goodness of fit

モデルの仮定に対する推定結果の感度を調べるためには、モデルが観測データにどれくらいあてはまっているか調べておくことが重要である。多数回サンプリングを仮定した場合、頻度主義者が使うテストでは、モデルからの乖離度合に対して、漸近カイ二乗分布を用いる。

$$T = \sum_{\omega \in \Omega} \frac{(x_{\omega} - \hat{e}_{\omega})^2}{\hat{e}_{\omega}}$$

ここでは、 x_{ω} は ω という捕獲履歴 (e.g. 0,1のデータ{1,00,1,000,,0})を持つ個体の総数、 \hat{e}_{ω} はモデルからの期待値、 Ω は捕獲履歴の総セットである (null historyは除く)。

このテストの問題→ \hat{e} から推定される e_{ω} は十分大きいことが想定されていること。

解決→:事後分布予測アセスメント (Section 5.6)を使うことである。このアプローチでは、期待値はMarkov チェインで生成された各パラメータセットごとに計算される。

$$T_h^{obs} = \sum_{\omega \in \Omega} \frac{(x_{\omega} - e_{\omega}^{(h)})^2}{e_{\omega}^{(h)}}$$

$$T_h^{rep} = \sum_{\omega \in \Omega} \frac{(x_{hw}^{(rep)} - e_{\omega}^{(h)})^2}{e_{\omega}^{(h)}}$$

このようにして、観察データは、モデルに基づいて作られたデータと比較することができる。

posterior predictive goodness of fit アセスメントをするために、MCMCで作られた各 p に対して繰り返しデータを生成した。この例に関しては、モデルに対するデータのあてはまりは悪い (Fig. 9.4)。

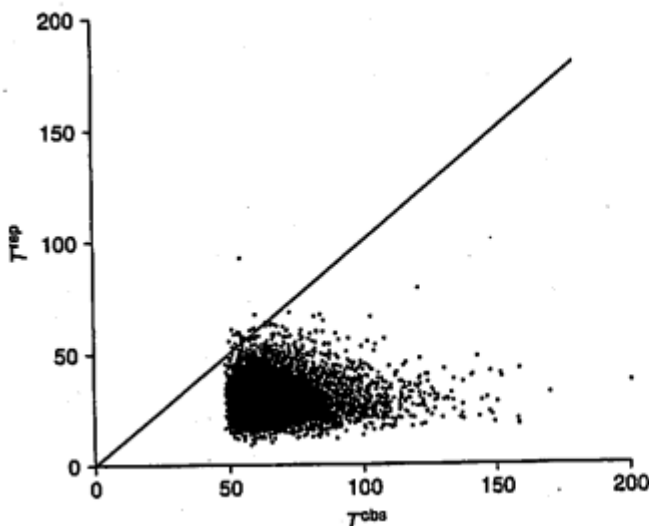


FIGURE 9.4 Scatterplot of predictive T^{rep} versus observed T^{obs} values of the discrepancy statistic for the female meadow vole data fitted to model M_1 . The Bayesian p -value is estimated by the proportion of points above the 45° line and has a value of 0.002.