

## 統計モデリング入門 2016 (f)

### 階層ベイズモデル Hierarchical Bayesian Model

久保拓弥 kubo@ees.hokudai.ac.jp

北大環境科学院の講義 <http://goo.gl/76c4i>

2016-07-27

ファイル更新時刻: 2016-07-27 16:06

kubostat2016f (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (f) 2016-07-27 1 / 66

## 今日の統計モデル: 階層ベイズモデル

### The development of linear models

そして **Markov Chain Monte Carlo (MCMC)** を使った Bayesian Estimation (ベイズ推定)

kubostat2016f (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (f) 2016-07-27 2 / 66

## Why? GLM is not enough ...

### 階層ベイズモデルが必要!

Apply Hierarchical Bayesian Model (HBM)!

kubostat2016f (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (f) 2016-07-27 3 / 66

## 今日のハナシ

- ① MCMC サンプリングのための **例題**  
logistic regression: binomial distribution
- ② 同じような推定を MCMC でやってみる  
最尤推定と Markov chain Monte Carlo (MCMC) はちがう!
- ③ Softwares for MCMC sampling  
“Gibbs sampling” などが簡単にできるような.....
- ④ GLMM と階層ベイズモデル  
GLMM のベイズモデル化
- ⑤ 階層ベイズモデルの **推定**  
ソフトウェア JAGS を使ってみる

kubostat2016f (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (f) 2016-07-27 4 / 66

MCMC サンプリングのための **例題**    logistic regression: binomial distribution

## example

### 1. MCMC サンプリングのための 例題

logistic regression: binomial distribution

and logit link function

kubostat2016f (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (f) 2016-07-27 5 / 66

MCMC サンプリングのための **例題**    logistic regression: binomial distribution

## example    seed survivorship, again

### 例題: 植物の種子の生存確率

- 架空植物の種子の生存を調べた
- 種子: 生きていれば発芽する
  - どの個体でも 8 個の種子を調べた
- 生存確率: ある種子が生きている確率
- データ: 植物 20 個体, 合計 160 種子の生存の有無を調べた
- 73 種子が生きていた — このデータを統計モデル化したい

kubostat2016f (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (f) 2016-07-27 6 / 66

MCMC サンプリングのための 例題 logistic regression: binomial distribution

たとえばこんなデータが得られたとしましょう

個体ごとの生存数	0	1	2	3	4	5	6	7	8
観察された個体数	1	2	1	3	6	6	1	0	0

観察された植物の個体数

生存していた種子数  $y_i$

これは個体差なしの均質な集団

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 7 / 66

MCMC サンプリングのための 例題 logistic regression: binomial distribution

生存確率  $q$  と 二項分布 の関係

- 生存確率を推定するために**二項分布** という確率分布を使う
- 個体  $i$  の  $N_i$  種子中  $y_i$  個が生存する確率

$$p(y_i | q) = \binom{N_i}{y_i} q^{y_i} (1-q)^{N_i-y_i}$$

- ここで仮定していること
  - 個体差はない
  - つまり すべての個体で同じ生存確率  $q$

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 8 / 66

MCMC サンプリングのための 例題 logistic regression: binomial distribution

ゆづど

尤度: 20 個体ぶんのデータが観察される確率

- 観察データ  $\{y_i\}$  が確定しているときに
- パラメータ  $q$  は値が自由にとりうると考える likelihood
- 尤度 は 20 個体ぶんのデータが得られる確率の積, パラメータ  $q$  の関数として定義される

$$L(q|\{y_i\}) = \prod_{i=1}^{20} p(y_i | q)$$

個体ごとの生存数	0	1	2	3	4	5	6	7	8
観察された個体数	1	2	1	3	6	6	1	0	0

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 9 / 66

MCMC サンプリングのための 例題 logistic regression: binomial distribution

対数尤度方程式と最尤推定

- この尤度  $L(q | \text{データ})$  を最大化するパラメータの推定量  $\hat{q}$  を計算したい
- 尤度を対数尤度になおすと

$$\log L(q | \text{データ}) = \sum_{i=1}^{20} \log \binom{N_i}{y_i} + \sum_{i=1}^{20} \{y_i \log(q) + (N_i - y_i) \log(1 - q)\}$$

- この対数尤度を最大化するように未知パラメーター  $q$  の値を決めてやるのが**最尤推定**

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 10 / 66

MCMC サンプリングのための 例題 logistic regression: binomial distribution

maximum likelihood estimation

最尤推定 (MLE) とは何か

- 対数尤度  $L(q | \text{データ})$  が最大になるパラメーター  $q$  の値をさがしだすこと
- 対数尤度  $\log L(q | \text{データ})$  を  $q$  で偏微分して 0 となる  $\hat{q}$  が対数尤度最大  $\partial \log L(q | \text{データ}) / \partial q = 0$
- 生存確率  $q$  が全個体共通の場合の最尤推定量・最尤推定値は

$$\hat{q} = \frac{\text{生存種子数}}{\text{調査種子数}} = \frac{73}{160} = 0.456 \text{ ぐらい}$$

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 11 / 66

MCMC サンプリングのための 例題 logistic regression: binomial distribution

二項分布で説明できる 8 種子中  $y_i$  個の生存

$\hat{q} = 0.46$  なので  $\binom{8}{y} 0.46^y 0.54^{8-y}$

観察された植物の個体数

生存していた種子数  $y_i$

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 12 / 66

同じような推定を MCMC でやってみる 最尤推定と Markov chain Monte Carlo (MCMC) はちがう

## 2. 同じような推定を MCMC でやってみる

最尤推定と Markov chain Monte Carlo (MCMC) はちがう!

そして“なんとなく”ベイズ統計モデルと関連づけ

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 13 / 66

同じような推定を MCMC でやってみる 最尤推定と Markov chain Monte Carlo (MCMC) はちがう

ここでやること: 尤度と MCMC の関係を考える

- さきほどの簡単な例題 (生存確率) のデータ解析を
- 最尤推定ではなく
- Markov chain Monte Carlo (MCMC) 法のひとつである**メトロポリス法** (Metropolis method) であつかう
- 得られる結果: 「パラメーターの値の分布」.....??

MCMC をもちださなくてもいい簡単すぎる問題  
説明のためあえてメトロポリス法を適用してみる

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 14 / 66

同じような推定を MCMC でやってみる 最尤推定と Markov chain Monte Carlo (MCMC) はちがう

### メトロポリス法を説明するための準備

連続的な対数尤度関数  $\log L(q)$

離散化:  $q$  がとびとびの値をとる

説明を簡単にするため  
生存確率  $q$  の軸を離散化する  
(実際には離散化する必要などない)

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 15 / 66

同じような推定を MCMC でやってみる 最尤推定と Markov chain Monte Carlo (MCMC) はちがう

### 試行錯誤による $q$ の最尤推定値の探索

ちょっと効率の悪い「試行錯誤の最尤推定」

- ①  $q$  の値の「行き先」を「両隣」どちらかにランダムに決める
- ② 「行き先」が現在の尤度より高ければ、 $q$  の値をそちらに変更
- ③ 尤度が変化しなくなるまで (1), (2) をくりかえす

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 16 / 66

同じような推定を MCMC でやってみる 最尤推定と Markov chain Monte Carlo (MCMC) はちがう

### メトロポリス法のルール: この例題の場合

- ① パラメーター  $q$  の初期値を選ぶ  
(ここでは  $q$  の初期値が 0.3)
- ②  $q$  を増やすか減らすかをランダムに決める  
(新しく選んだ  $q$  の値を  $q_{new}$  としましょう)
- ③  $q_{new}$  における尤度  $L(q_{new})$  ともとの尤度  $L(q)$  を比較
  - $L(q_{new}) \geq L(q)$  (あてはまり改善):  $q \leftarrow q_{new}$
  - $L(q_{new}) < L(q)$  (あてはまり改悪):
    - 確率  $r = L(q_{new})/L(q)$  で  $q \leftarrow q_{new}$
    - 確率  $1-r$  で  $q$  を変更しない
- ④ 手順 2. にもどる  
( $q = 0.01$  や  $q = 0.99$  でどうなるんだ、といった問題は省略)

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 17 / 66

同じような推定を MCMC でやってみる 最尤推定と Markov chain Monte Carlo (MCMC) はちがう

### メトロポリス法のルールで $q$ を動かす

最尤推定法

メトロポリス法 (MCMC)

「単調な山のぼり」にはならない

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 18 / 66

同じような推定を MCMC でやってみる 最尤推定と Markov chain Monte Carlo (MCMC) はちがう

### 対数尤度関数の「山」でうろうろする $q$ の値

メトロポリス法 (そして一般の MCMC) は  
最適化ではない

ときどきはでに落ちこちる  
何のためにこんなことをやるのか?  
 $q$  の変化していく様子を記録してみよう

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 19 / 66

同じような推定を MCMC でやってみる 最尤推定と Markov chain Monte Carlo (MCMC) はちがう

### ステップごとに $q$ の値をサンプリング

この曲線、何の分布?  
サンプルされた  $q$  のヒストグラム

もっと試行錯誤してみたほうがいいのか?

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 20 / 66

同じような推定を MCMC でやってみる 最尤推定と Markov chain Monte Carlo (MCMC) はちがう

### もっと長くサンプリングしてみる

この曲線、何の分布?  
サンプルされた  $q$  のヒストグラム

まだまだ.....?

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 21 / 66

同じような推定を MCMC でやってみる 最尤推定と Markov chain Monte Carlo (MCMC) はちがう

### もっともっと長くサンプリングしてみる

じつはこれは「 $q$  の確率分布」  
.....このあと説明

サンプルされた  $q$  のヒストグラム

なんだか、ある「山」のかたちにとまとまったぞ?

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 22 / 66

同じような推定を MCMC でやってみる 最尤推定と Markov chain Monte Carlo (MCMC) はちがう

### MCMC は何をサンプリングしている?

対数尤度  $\log L(q)$

尤度  $L(q)$  に  
比例する確率分布

尤度に比例する確率分布からのランダムサンプル

最尤推定はパラメーターの値の点推定  
MCMC は「パラメーターの事後分布」(推定したいこと)  
はこういう分布ですよと推定している

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 23 / 66

同じような推定を MCMC でやってみる 最尤推定と Markov chain Monte Carlo (MCMC) はちがう

### MCMC の結果として得られた $q$ の経験分布

- データと統計モデル (二項分布) を決めて、MCMC サンプルすると、 $p(q)$  からのランダムサンプルが得られる
- このランダムサンプルをもとに、 $q$  の平均や 95% 区間などがわかる— 便利じゃないか!

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 24 / 66

同じような推定を MCMC でやってみる 最尤推定と Markov chain Monte Carlo (MCMC) はちがう

## ベイズ統計モデルの推定

**統計モデルとデータにもとづいて事後分布の推定**

- パラメータ数の少ないベイズモデルであれば、尤度の数値計算やメトロポリス法で可能
- パラメータ数の多い複雑な統計モデルであれば、あとで説明する サンプリングソフトウェアを使用する

事後分布  $p(q|Y)$     尤度  $L(q)$     事前分布  $p(q)$

生存確率  $q$

kubostat2016f (<http://goo.gl/76c4s>)    統計モデリング入門 2016 (f)    2016-07-27    25 / 66

Softwares for MCMC sampling “Gibbs sampling” などが簡単にできるような.....

## 3. Softwares for MCMC sampling

“Gibbs sampling” などが簡単にできるような.....

事後分布から効率よくサンプリングしたい

kubostat2016f (<http://goo.gl/76c4s>)    統計モデリング入門 2016 (f)    2016-07-27    26 / 66

Softwares for MCMC sampling “Gibbs sampling” などが簡単にできるような.....

## 統計ソフトウェア R

<http://www.r-project.org/>

kubostat2016f (<http://goo.gl/76c4s>)    統計モデリング入門 2016 (f)    2016-07-27    27 / 66

Softwares for MCMC sampling “Gibbs sampling” などが簡単にできるような.....

## 簡単な GLMM なら R だけで推定可能

- R にはいろいろな GLMM の最尤推定関数が準備されている.....
  - library(glmML) の glmML()
  - library(lme4) の lmer()
  - library(nlme) の nlme() (正規分布のみ)
- しかし もうちょっと複雑な GLMM, たとえば個体差 + 地域差をいれた統計モデルの最尤推定は **かなり難しい** (へんな結果が得られたりする)
- 積分がたたくさん入っている尤度関数の評価がしんどい

kubostat2016f (<http://goo.gl/76c4s>)    統計モデリング入門 2016 (f)    2016-07-27    28 / 66

Softwares for MCMC sampling “Gibbs sampling” などが簡単にできるような.....

## どのようなソフトウェアで MCMC 計算するか?

- 自作プログラム
  - 利点: 問題にあわせて自由に設計できる
  - 欠点: 階層ベイズモデル用の MCMC プログラミング, けっこうめんどろ
- R のベイズな package
  - 利点: 空間ベイズ統計など便利な専用 package がある
  - 欠点: 汎用性, とぼしい
- “BUGS” で “Gibbs sampler” なソフトウェア
  - 利点: 幅ひろい問題に適用できて, 便利
  - 欠点というほどでもないけど, 多少の勉強が必要
  - えーっと “Gibbs sampler” って何?

kubostat2016f (<http://goo.gl/76c4s>)    統計モデリング入門 2016 (f)    2016-07-27    29 / 66

Softwares for MCMC sampling “Gibbs sampling” などが簡単にできるような.....

## さまざまな MCMC アルゴリズム

### いろいろな MCMC

- メトロポリス法:** 試行錯誤で値を変化させていく MCMC
  - メトロポリス・ヘイスティングス法: その改良版
- ギブス・サンプリング:** 条件つき確率分布を使った MCMC
  - 複数の変数 (パラメータ・状態) を効率よくサンプリング

kubostat2016f (<http://goo.gl/76c4s>)    統計モデリング入門 2016 (f)    2016-07-27    30 / 66

Softwares for MCMC sampling "Gibbs sampling" などが簡単にできるような.....

### Gibbs sampling とは何か?

- MCMC アルゴリズムのひとつ
- 複数のパラメータの MCMC サンプリングに使う
- 例: パラメータ  $\beta_1$  と  $\beta_2$  の Gibbs sampling
  - $\beta_2$  に何か適当な値を与える
  - $\beta_2$  の値はそのままにして、その条件のもとでの  $\beta_1$  の MCMC sampling をする (条件つき事後分布)
  - $\beta_1$  の値はそのままにして、その条件のもとでの  $\beta_2$  の MCMC sampling をする (条件つき事後分布)
  2. - 3. をくりかえす
- 教科書の第 9 章の例題で説明

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 31 / 66

Softwares for MCMC sampling "Gibbs sampling" などが簡単にできるような.....

### 図解: Gibbs sampling (統計モデリング入門の第 9 章)

MCMC  $\beta_1$  のサンプリング  $\beta_2$  のサンプリング

step 1

step 2

step 3

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 32 / 66

Softwares for MCMC sampling "Gibbs sampling" などが簡単にできるような.....

### 便利な "BUGS" 汎用 Gibbs sampler たち

- BUGS 言語 (+ っぽいもの) でベイズモデルを記述できるソフトウェア
  - WinBUGS — 歴史を変えて.....さようなら?
  - OpenBUGS — 予算が足りなくて停滞?
  - JAGS — お手軽で良い, どんな OS でも動く
  - Stan — いま一番の注目
    - 今日では紹介しませんが.....
- リンク集: <http://hosho.ees.hokudai.ac.jp/~kubo/ce/BayesianMcmc.html>

えーと.....BUGS 言語って何?

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 33 / 66

Softwares for MCMC sampling "Gibbs sampling" などが簡単にできるような.....

### このベイズモデルを BUGS 言語で記述したい

データ Y[i]  
種子数8個のうちの生存数

二項分布  
dbin(q, 8)

生存確率 q

無情報事前分布

BUGS 言語コード

```
for (i in 1:N.sample) {
  Y[i] ~ dbin(q, 8)
}
q ~ dunif(0.0, 1.0)
```

矢印は手順ではなく、依存関係をあらわしている  
BUGS 言語: ベイズモデルを記述する言語

Spiegelhalter et al. 1995. BUGS: Bayesian Using Gibbs Sampling version 0.50.

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 34 / 66

Softwares for MCMC sampling "Gibbs sampling" などが簡単にできるような.....

### いろいろな OS で使える JAGS3.4.x

- R core team のひとり Martyn Plummer さんが開発
  - Just Another Gibbs Sampler
- C++ で実装されている
  - R がインストールされていることが必要
- Linux, Windows, Mac OS X バイナリ版もある
- 開発進行中
- R からの使う: library(rjags)

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 35 / 66

Softwares for MCMC sampling "Gibbs sampling" などが簡単にできるような.....

### JAGS を R の "したうけ" として使う

モデルの構造

データとパラメータの初期値

サンプリングの詳細

Input

JAGS

事後分布からのランダムサンプル

Trace of beta(1)

Density of beta(1)

Trace of beta(2)

Density of beta(2)

Output

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 36 / 66

Softwares for MCMC sampling "Gibbs sampling" などが簡単にできるような.....

## R から JAGS にこんなかんじで仕事を命じる (1 / 3)

```
library(rjags)
library(R2WinBUGS) # to use write.model()

model.bugs <- function()
{
  for (i in 1:N.data) {
    Y[i] ~ dbin(q, 8) # 二項分布にしたがう
  }
  q ~ dunif(0.0, 1.0) # q の事前分布は一様分布
}
file.model <- "model.bug.txt"
write.model(model.bugs, file.model) # ファイル出力
```

# 次につづく.....

kubostat2016f (<http://goo.gl/76c4s>)

統計モデリング入門 2016 (f)

2016-07-27

37 / 66

Softwares for MCMC sampling "Gibbs sampling" などが簡単にできるような.....

## R から JAGS にこんなかんじで仕事を命じる (2 / 3)

```
load("mcmc.RData") # (data.RData ではなく mcmc.RData!!)
list.data <- list(Y = data, N.data = length(data))
inits <- list(q = 0.5)
n.burnin <- 1000
n.chain <- 3
n.thin <- 1
n.iter <- n.thin * 1000

model <- jags.model(
  file = file.model, data = list.data,
  inits = inits, n.chain = n.chain
)
```

# まだ次につづく.....

kubostat2016f (<http://goo.gl/76c4s>)

統計モデリング入門 2016 (f)

2016-07-27

38 / 66

Softwares for MCMC sampling "Gibbs sampling" などが簡単にできるような.....

## R から JAGS にこんなかんじで仕事を命じる (3 / 3)

```
# burn-in
update(model, n.burnin) # burn in

# サンプリング結果を post.mcmc.list に格納
post.mcmc.list <- coda.samples(
  model = model,
  variable.names = names(inits),
  n.iter = n.iter,
  thin = n.thin
)
# おわり
```

kubostat2016f (<http://goo.gl/76c4s>)

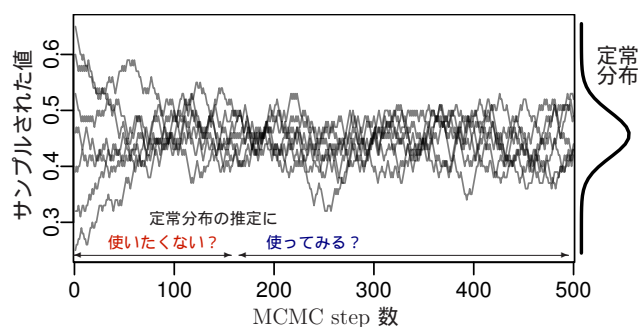
統計モデリング入門 2016 (f)

2016-07-27

39 / 66

Softwares for MCMC sampling "Gibbs sampling" などが簡単にできるような.....

## burn in って何? → 「使いたくない」長さの指定

kubostat2016f (<http://goo.gl/76c4s>)

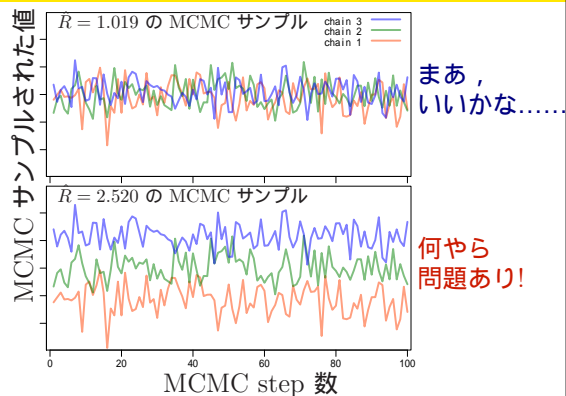
統計モデリング入門 2016 (f)

2016-07-27

40 / 66

Softwares for MCMC sampling "Gibbs sampling" などが簡単にできるような.....

## 試行間で差がないかを「診断」する

kubostat2016f (<http://goo.gl/76c4s>)

統計モデリング入門 2016 (f)

2016-07-27

41 / 66

Softwares for MCMC sampling "Gibbs sampling" などが簡単にできるような.....

収束診断の  $\hat{R}$  指数

- `gelman.diag(post.mcmc.list)` → 実演表示
- $\hat{R}$  は Gelman-Rubin の収束判定用の指数
  - $\hat{R} = \sqrt{\frac{\text{var}^+(\psi|y)}{W}}$
  - $\text{var}^+(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$
  - $W$ : サンプル列内の variance の平均
  - $B$ : サンプル列間の variance
  - Gelman et al. 2004. Bayesian Data Analysis. Chapman & Hall/CRC

kubostat2016f (<http://goo.gl/76c4s>)

統計モデリング入門 2016 (f)

2016-07-27

42 / 66

Softwares for MCMC sampling "Gibbs sampling" などが簡単にできるような.....

### Gibbs sampling → 事後分布の推定

- plot(post.mcmc.list)

**Trace of q**

Iterations

**Density of q**

N = 1000 Bandwidth = 0.0083f

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 43 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

## 4. GLMM と階層ベイズモデル

### GLMM のベイズモデル化

hierarchical Bayesian  
階層ベイズモデルとなる

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 44 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

### 二項分布では説明できない観測データ!

100 個体の植物の合計 800 種子中 **403 個** の生存が見られたので、平均生存確率は 0.50 と推定されたが.....

観察された植物の個体数

生存した種子数  $y_i$

二項分布による予測

ぜんぜんうまく表現できてない!

さっきの例題と同じようなデータなのに?  
(「統計モデリング入門」第 10 章の最初の例題)

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 45 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

### 個体差 → 過分散 (overdispersion)

極端な過分散の例

観察された植物の個体数

生存した種子数  $y_i$

- 種子全体の平均生存確率は 0.5 くらいかもしれないが.....
- 植物個体ごとに種子の生存確率が異なる: 「個体差」
- 「個体差」があると overdispersion が生じる
- 「個体差」の原因は観測できない・観測されていない

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 46 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

### モデリングやりなおし: 個体差を考慮する

- 生存確率を推定するために **二項分布** という確率分布を使う
- 個体  $i$  の  $N_i$  種子中  $y_i$  個が生存する確率は二項分布

$$p(y_i | q_i) = \binom{N_i}{y_i} q_i^{y_i} (1 - q_i)^{N_i - y_i}$$

- ここで仮定していること
  - 個体差があるので個体ごとに生存確率  $q_i$  が異なる

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 47 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

### GLM わざ: ロジスティック関数で表現する生存確率

- 生存確率  $q_i = q(z_i)$  をロジスティック関数  $q(z) = 1 / \{1 + \exp(-z)\}$  で表現

$q(z)$

- 線形予測子  $z_i = a + r_i$  とする
  - パラメーター  $a$ : 全体の平均
  - パラメーター  $r_i$ : 個体  $i$  の個体差 (ずれ)

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 48 / 66



GLMM と階層ベイズモデル GLMM のベイズモデル化

個々の個体差  $r_i$  を最尤推定するのはまずい

パラメーター数 > サンプルサイズ

- 100 個体の生存確率を推定するためにパラメーター 101 個 ( $a$  と  $\{r_1, r_2, \dots, r_{100}\}$ ) を推定すると.....
- 個体ごとに生存数 / 種子数を計算していることと同じ! (「データのみあげ」と同じ)

そこで、次のように考えてみる

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 49 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

suppose  $\{r_i\}$  follow the Gaussian distribution  
 $\{r_i\}$  のばらつきは正規分布だと考えてみる

$$p(r_i | s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{r_i^2}{2s^2}\right)$$

この確率密度  $p(r_i | s)$  は  $r_i$  の「出現しやすさ」をあらわしていると解釈すればよいでしょう。  $r_i$  がゼロにちかい個体はわりと「ありがち」で、  $r_i$  の絶対値が大きな個体は相対的に「あまりいない」。

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 50 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

ひとつの例示: 個体差  $r_i$  の分布と過分散の関係

(A) 個体差のばらつきが小さい場合 (B) 個体差のばらつきが大きい場合

確率  $q_i = \frac{1}{1 + \exp(-r_i)}$  の二項乱数を発生させる

観察された個体数  $y_i$  対して  $p(y_i | q_i)$  が生成した生存種子数の一例

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 51 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

これは  $r_i$  の事前分布の指定、ということ

前回の講義で  $\{r_i\}$  は正規分布にしたがうと仮定したが  
 ベイズ統計モデリングでは「100 個の  $r_i$  たちに  
 共通する事前分布として正規分布を指定した」  
 ということになる

$$p(r_i | s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{r_i^2}{2s^2}\right)$$

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 52 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

ベイズ統計モデルでよく使われる三種類の事前分布

たいいていのベイズ統計モデルでは、ひとつのモデルの中で複数の種類の事前分布を混ぜて使用する。

(A) 主観的な事前分布 (できれば使いたくない!)  
 信じる!

(B) 無情報事前分布  
 わからない?

(C) 階層事前分布  
 $s$  によって変わる...

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 53 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

$r_i$  の事前分布として階層事前分布を指定する

階層事前分布の利点  
 「データにあわせて」事前分布が変形!

$$p(r_i | s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{r_i^2}{2s^2}\right)$$

kubostat2016f (http://goo.gl/76c4i) 統計モデリング入門 2016 (f) 2016-07-27 54 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

### 統計モデルの大域的・局所的なパラメーター

全データ  
 個体 1 のデータ  
 個体 2 のデータ  
 個体 3 のデータ

$\{r_1, r_2, r_3, \dots, r_{100}\}$  ←  $s$   $a$

local parameter random effects      global parameter fixed effects

データのどの部分を説明しているのか?

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 55 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

### パラメーターごとに適切な事前分布を選ぶ

(B) 無情報事前分布 (C) 階層事前分布

$a, s$   
わからない?

$\{r_i\}$   
 $s$ によって変わる...

パラメーターの種類	説明する範囲	事前分布
全体に共通する平均・ばらつき	global 大域的	無情報事前分布
個体・グループごとのずれ	local 局所的	階層事前分布

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 56 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

### 個体差 $\{r_i\}$ のばらつき $s$ の無情報事前分布

- $s$  はどのような値をとってもかまわない
- そこで  $s$  の事前分布は **無情報事前分布** (non-informative prior) とする
- たとえば一様分布, ここでは  $0 < s < 10^4$  の一様分布としてみる

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 57 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

### 全個体の「切片」 $a$ の無情報事前分布

「生存確率の (logit) 平均  $a$  は何でもよい」と表現している

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 58 / 66

GLMM と階層ベイズモデル GLMM のベイズモデル化

### 階層ベイズモデル: 事前分布の階層性

超事前分布 → 事前分布という階層があるから

データ 種子8個のうち  $Y[i]$  が生存

二項分布 生存確率  $q[i]$  ←  $r[i]$  ← 植物の個体差

事前分布 hyper  $s$  個体差のばらつき parameter

無情報事前分布 (超事前分布)

全体共通の「平均」  $a$  無情報事前分布

矢印は手順ではなく、依存関係をあらわしている

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 59 / 66

階層ベイズモデルの 推定 ソフトウェア JAGS を使ってみる

estimation

### 5. 階層ベイズモデルの 推定

ソフトウェア JAGS を使ってみる

R の “したうけ” として JAGS を使う

kubostat2016f (http://goo.gl/76c4s) 統計モデリング入門 2016 (f) 2016-07-27 60 / 66

階層ベイズモデルの推定 ソフトウェア JAGS を使ってみる

### 階層ベイズモデルを BUGS コードで記述する

```

model
{
  for (i in 1:N.data) {
    Y[i] ~ dbin(q[i], 8)
    logit(q[i]) <- a + r[i]
  }
  a ~ dnorm(0, 1.0E-4)
  for (i in 1:N.data) {
    r[i] ~ dnorm(0, tau)
  }
  tau <- 1 / (s * s)
  s ~ dunif(0, 1.0E+4)
}
    
```

61 / 66

階層ベイズモデルの推定 ソフトウェア JAGS を使ってみる

### JAGS で得られた事後分布サンプルの要約

```

> source("mcmc.list2bugs.R") # なんとなく便利なので...
> post.bugs <- mcmc.list2bugs(post.mcmc.list) # bugs クラスに変換
    
```

62 / 66

階層ベイズモデルの推定 ソフトウェア JAGS を使ってみる

### bugs オブジェクトの post.bugs を調べる

- print(post.bugs, digits.summary = 3)
- 事後分布の 95% 信頼区間などが表示される

```

3 chains, each with 4000 iterations (first 2000 discarded), n.thin = 2
n.sims = 3000 iterations saved
    
```

	mean	sd	2.5%	25%	50%	75%	97.5%	Rhat	n.eff
a	0.020	0.321	-0.618	-0.190	0.028	0.236	0.651	1.007	380
s	3.015	0.359	2.406	2.757	2.990	3.235	3.749	1.002	1200
r[1]	-3.778	1.713	-7.619	-4.763	-3.524	-2.568	-1.062	1.001	3000
r[2]	-1.147	0.885	-2.997	-1.700	-1.118	-0.531	0.464	1.001	3000
r[3]	2.014	1.074	0.203	1.282	1.923	2.648	4.410	1.001	3000
r[4]	3.765	1.722	0.998	2.533	3.558	4.840	7.592	1.001	3000
r[5]	-2.108	1.111	-4.480	-2.775	-2.047	-1.342	-0.164	1.001	2300
...	(中略)								
r[99]	2.054	1.103	0.184	1.270	1.996	2.716	4.414	1.001	3000
r[100]	-3.828	1.766	-7.993	-4.829	-3.544	-2.588	-1.082	1.002	1100

63 / 66

階層ベイズモデルの推定 ソフトウェア JAGS を使ってみる

### 各パラメーターの事後分布サンプルを R で調べる

64 / 66

階層ベイズモデルの推定 ソフトウェア JAGS を使ってみる

### 得られた事後分布サンプルを組みあわせて予測

- post.mcmc <- to.mcmc(post.bugs)
- これは matrix と同じようにあつかえるので、作図に便利
- .....このあとごちゃごちゃと計算する必要あるけど、省略.....

65 / 66

階層ベイズモデルの推定 ソフトウェア JAGS を使ってみる

## 次回予告

The next topic

### 階層ベイズモデルと時間変化モデル

Hierarchical Bayesian Model (HBM) & Time Change Model

66 / 66