

## 統計モデリング入門 2017 (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>

2017-06-26

ファイル更新時刻: 2018-06-29 15:05

kubostat2017f (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (f) 2017-06-26 1 / 75

## 今日の統計モデル: 階層ベイズモデル

### The development of linear models

Be more flexible

↑

Hierarchical Bayesian Model

parameter estimation MCMC

↑

Generalized Linear Mixed Model (GLMM)

MLE

↑

Generalized Linear Model (GLM)

MSE

↑

Linear model

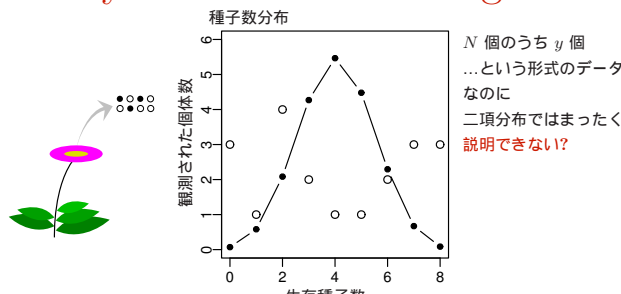
Incorporating random effects such as individuality

Always normal distribution? That's non-sense!

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

kubostat2017f (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (f) 2017-06-26 2 / 75

## Why? GLM is not enough ...



種子数分布

$N$  個のうち  $y$  個 ... という形式のデータなのに二項分布ではまったく説明できない!

### 階層ベイズモデルが必要!

Apply Hierarchical Bayesian Model (HBM)!

kubostat2017f (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (f) 2017-06-26 3 / 75

## 今日のハナシ

example

- ① MCMC sampling のための 例題  
logistic regression: binomial distribution
- ② The same data, but applying Markov Chain Monte-Carlo (MCMC)  
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
- ⑥ 階層ベイズモデル (HBM)? or GLMM?  
Model: HBM and GLMM are equivalent

kubostat2017f (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (f) 2017-06-26 4 / 75

MCMC sampling のための 例題    logistic regression: binomial distribution

### example

## 1. MCMC sampling のための 例題

logistic regression: binomial distribution

and logit link function

kubostat2017f (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (f) 2017-06-26 5 / 75

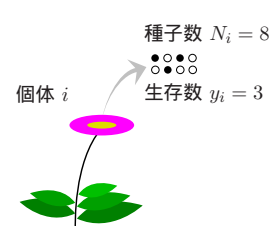
MCMC sampling のための 例題    logistic regression: binomial distribution

### 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



kubostat2017f (<http://goo.gl/76c4i>) 統計モデリング入門 2017 (f) 2017-06-26 6 / 75

MCMC sampling のための 例題 logistic regression: binomial distribution

### 二項分布を説明するための例題

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 (no individual difference!)

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 7 / 75

MCMC sampling のための 例題 logistic regression: binomial distribution

### 生存確率 $q$ と 二項分布 の関係

survivorship

- 生存確率  $q$  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

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

- In this example ...
  - 個体差はない** no individual difference
  - つまり **すべての個体で同じ生存確率  $q$**

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 8 / 75

MCMC sampling のための 例題 logistic regression: binomial distribution

### 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
- 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

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 9 / 75

MCMC sampling のための 例題 logistic regression: binomial distribution

### 対数尤度 方程式と 最尤推定

log likelihood MLE

- 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  $q$  の値を決めてやるのが**最尤推定**

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 10 / 75

MCMC sampling のための 例題 logistic regression: binomial distribution

### maximum likelihood estimation 最尤推定 (MLE) : 二項分布の場合

- 対数尤度  $L(q | \text{データ})$  が最大になるパラメータ  $q$  の値をさがすこと
- 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
- 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

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

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 11 / 75

MCMC sampling のための 例題 logistic regression: binomial distribution

### fitting 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  $y_i$

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 12 / 75

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

## 2. The same data, but applying Markov Chain Monte-Carlo (MCMC)

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

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 13 / 75

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

### Maximum likelihood Estimation (MLE) vs. MCMC

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 **Metropolis Method** (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 Metropolis Method を適用してみる

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 14 / 75

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

### An example for MCMC

#### MCMC 法を説明するための例題

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)

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 15 / 75

Monte-Carlo (MCMC) 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」

$q$ : seed srurvivorship

- 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

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 16 / 75

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

### この例題の Metropolis Method のルール

- 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)

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 17 / 75

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

### Metropolis Method のルールで $q$ を動かす

最尤推定法 (left) vs. メトロポリス法 (MCMC) (right)

Metropolis Method だと  
「単調な山のぼり」にはならない

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 18 / 75

Monte-Carlo (MCMC) 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

Metropolis Method (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

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 19 / 75

Monte-Carlo (MCMC) 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?

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 20 / 75

Monte-Carlo (MCMC) 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.....?

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 21 / 75

Monte-Carlo (MCMC) 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」のかたちにとまったぞ?

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 22 / 75

Monte-Carlo (MCMC) 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

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 23 / 75

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

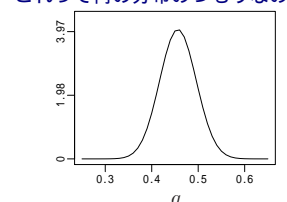
### MCMC の結果として得られた $q$ の経験分布

- 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!

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 24 / 75

Monte-Carlo (MCMC) 最尤推定と Markov chain Monte Carlo (MCMC) はちがう  
 (非ベイズな統計学)  
**しかし 普通の統計学 では  $q$  の分布 とかありえない**

これって何の分布のつもりなの?



- パラメーター  $q = 0.4500000 \dots$  といったスカラー値であり, 分布ではない!
- 信頼区間は “ $q$  の分布” ではない!

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 25 / 75

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

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 26 / 75

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... is it enough?**

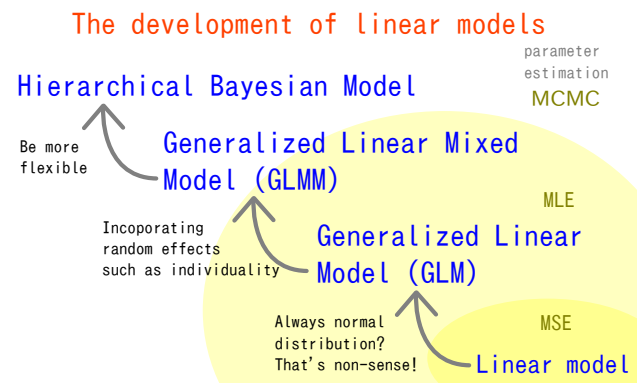
<http://www.r-project.org/>



kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 27 / 75

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

### The development of linear models



kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 28 / 75

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 (ヘンな結果が得られたりする)
- 積分がたくさん入っている尤度関数の評価がしんどい

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 29 / 75

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?**

- ① 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, けっこうめんどう
- ② 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
- ③ “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: 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?

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 30 / 75

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

- **Metropolis Method**: 試行錯誤で値を変化させていく MCMC
  - Metropolis-Hastings: その改良版
- **Gibbs sampling**: 条件つき確率分布を使った MCMC
  - 複数の変数 (パラメーター・状態) を効率よくサンプリング an efficient method for sampling of parameter values

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 31 / 75

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
- 複数のパラメーターの MCMC サンプリングに使う
- 例: パラメーター  $\beta_1$  と  $\beta_2$  の Gibbs sampling
  - ①  $\beta_2$  に何か適当な値を与える
  - ②  $\beta_2$  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)
  - ③  $\beta_1$  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)
  - ④ 2. - 3. をくりかえす
- 教科書の第 9 章の例題で説明

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 32 / 75

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

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 33 / 75

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?

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 34 / 75

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.

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 35 / 75

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 JAGS 4.3.0

- 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)`

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 36 / 75

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

MCMC sampling from posterior distributions  
事後分布からのランダムサンプル

Input

Output

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 37 / 75

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.....
```

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 38 / 75

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.....
```

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 39 / 75

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
```

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 40 / 75

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

定常分布

定常分布の推定に使いたくない? 使ってみる?

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 41 / 75

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

$R = 1.019$  の MCMC サンプル

まあ、いいかな.....

$R = 2.520$  の MCMC サンプル

何やら問題あり!

kubostat2017f (http://goo.gl/76c4s) 統計モデリング入門 2017 (f) 2017-06-26 42 / 75



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

### 収束診断の $\hat{R}$ (アールハット) 指数

- `gelman.diag(post.mcmc.list)` → 実演表示
- R-hat 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

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 43 / 75

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 = 0.0083E

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 44 / 75

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

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 45 / 75

GLMM と階層ベイズモデル GLMM のベイズモデル化

### Binomial distribution can NOT explain the DATA! 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)

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 46 / 75

GLMM と階層ベイズモデル GLMM のベイズモデル化

### individual difference 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 ぐ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

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 47 / 75

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: Modeling of individual difference

- 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

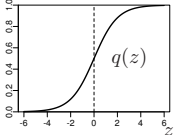
kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 48 / 75



GLMM と階層ベイズモデル GLMM のベイズモデル化

各個体の生存確率は logistic 回帰のモデルを使う

- 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



- 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)

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 49 / 75

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 = データのよみあげ

number of parameters **パラメーター数** > sample size **サンプルサイズ**

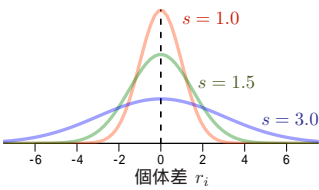
- 100 個体の生存確率を推定するためにパラメーター 101 個 ( $a$  と  $\{r_1, r_2, \dots, r_{100}\}$ ) を推定すると.....
- 個体ごとに生存数 / 種子数を計算していることと同じ! (“データのよみあげ” と等価)

そこで、次のように考えてみる

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 50 / 75

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



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

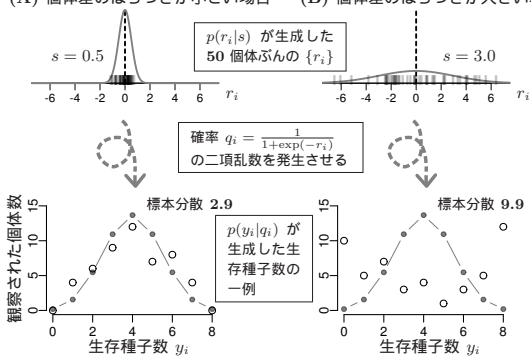
この確率密度  $p(r_i | s)$  は  $r_i$  の「出現しやすさ」をあらわしていると解釈すればよいでしょう。  $r_i$  がゼロにちかい個体はわりと「ありがち」で、  $r_i$  の絶対値が大きな個体は相対的に「あまりない」。

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 51 / 75

GLMM と階層ベイズモデル GLMM のベイズモデル化

overdispersion  
 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 vs 生存種子数  $y_i$

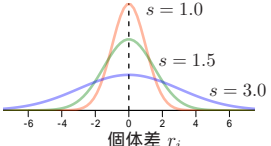
標本分散 2.9 (A) vs 標本分散 9.9 (B)

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 52 / 75

GLMM と階層ベイズモデル GLMM のベイズモデル化

prior  
 これは  $r_i$  の 事前分布 の指定, という事

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



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

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 53 / 75

GLMM と階層ベイズモデル GLMM のベイズモデル化

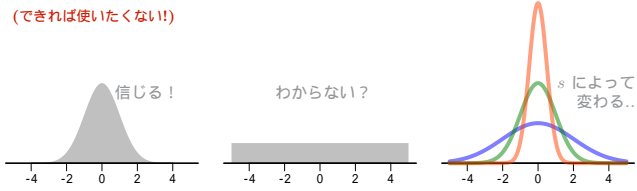
priors  
 ベイズ統計モデルでよく使われる三種類の 事前分布

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...



kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 54 / 75

GLMM と階層ベイズモデル GLMM のベイズモデル化

hierarchical prior  
 $r_i$  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!

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

2017-06-26 55 / 75

GLMM と階層ベイズモデル GLMM のベイズモデル化

Global and local parameters in the model  
 統計モデルの大域的・局所的なパラメーター

local parameter random effects  $\{r_1, r_2, r_3, \dots, r_{100}\}$   
 global parameter fixed effects  $a, s$

データのどの部分を説明しているのか?

2017-06-26 56 / 75

GLMM と階層ベイズモデル GLMM のベイズモデル化

choose proper priors: non-informative and hierarchical priors  
 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?

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

2017-06-26 57 / 75

GLMM と階層ベイズモデル GLMM のベイズモデル化

non-informative prior  
 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

2017-06-26 58 / 75

GLMM と階層ベイズモデル GLMM のベイズモデル化

non-informative prior  
 全個体の「切片」  $a$  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

2017-06-26 59 / 75

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: Hierarchical and non-informative priors

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  $q[i]$   
 植物の個体差  $r[i]$   
 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

2017-06-26 60 / 75

階層ベイズモデルの 推定 ソフトウェア JAGS を使ってみる

### 5. 階層ベイズモデルの 推定 <sup>estimation</sup>

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

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 61 / 75

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)
}
    
```

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]

植物の個体差 r[i]

事前分布 hyperparameter

全個体共通の「平均」 a

無情報事前分布

個体差のばらつき parameter

無情報事前分布 (超事前分布)

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 62 / 75

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
    
```

3 chains, each with 4000 iterations (first 2000 discarded)

80% integral for each chain,  $R_{j|k} > 2$ , medians and 80% intervals

\* array truncated for lack of space

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 63 / 75

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
    
```

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 64 / 75

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

Trace of a

Density of a

Iterations

N = 1000 Bandwidth = 0.06795

Trace of s

Density of s

Iterations

N = 1000 Bandwidth = 0.07627

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 65 / 75

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.....

観察された 植物の個体数

生存していた種子数

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 66 / 75

階層ベイズモデル (HBM)? or GLMM? Model: HBM and GLMM are equivalent

## 6. 階層ベイズモデル (HBM)? or GLMM?

Model: HBM and GLMM are equivalent

Estimation: NOT equivalent

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 67 / 75

階層ベイズモデル (HBM)? or GLMM? Model: HBM and GLMM are equivalent

## Generalized Linear Mixed Model 一般化線形混合モデル

The development of linear models

Hierarchical Bayesian Model (parameter estimation MCMC)

Generalized Linear Mixed Model (GLMM) (Be more flexible)

Generalized Linear Model (GLM) (Incorporating random effects such as individuality)

Linear model (Always normal distribution? That's non-sense!)

MLE, MSE

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 68 / 75

階層ベイズモデル (HBM)? or GLMM? Model: HBM and GLMM are equivalent

## Improve logistic regression model (GLM)!

### ロジスティック回帰のモデル

probability distribution binomial distribution

- 確率分布: 二項分布
- 線形予測子:  $\beta_1 + \beta_2 x_i + r_i$
- リンク関数: logit リンク関数

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 69 / 75

階層ベイズモデル (HBM)? or GLMM? Model: HBM and GLMM are equivalent

## 個体 $i$ の個体差を $r_i$ としてみよう

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 70 / 75

階層ベイズモデル (HBM)? or GLMM? Model: HBM and GLMM are equivalent

fixed effects random effects  
固定効果 と ランダム効果

## Generalized Linear Mixed Model (GLMM)

linear predictor  
で使う Mixed な 線形予測子:  $\beta_1 + \beta_2 x_i + r_i$

- fixed effects:  $\beta_1 + \beta_2 x_i$
- random effects:  $+r_i$

fixed? random? よくわからん.....?

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 71 / 75

階層ベイズモデル (HBM)? or GLMM? Model: HBM and GLMM are equivalent

## global parameter と local parameter

## Generalized Linear Mixed Model (GLMM)

linear predictor  
で使う Mixed な 線形予測子:  $\beta_1 + \beta_2 x_i + r_i$

- fixed effects:  $\beta_1 + \beta_2 x_i$ 
  - global parameter — for all individuals
- 全個体のばらつき  $s$  も global parameter
- random effects:  $+r_i$ 
  - local parameter — only for individual  $i$

kubostat2017f (http://goo.gl/76c4i) 統計モデリング入門 2017 (f) 2017-06-26 72 / 75

階層ベイズモデル (HBM)? or GLMM? Model: HBM and GLMM are equivalent

### maximum likelihood estimation of GLMM

データ  $y_i \sim$  binomial distribution

$$p(y_i | \beta_1, \beta_2) = \binom{8}{y_i} q_i^{y_i} (1 - q_i)^{8 - y_i}$$

個体差  $r_i \sim$  Gaussian distribution

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

Integrate out  $r_i$ !

$$L_i = \int_{-\infty}^{\infty} p(y_i | \beta_1, \beta_2, r_i) p(r_i | s) dr_i$$

likelihood for all data  
全データの尤度 —  $\beta_1, \beta_2, s$  の関数

$$L(\beta_1, \beta_2, s) = \prod_i L_i$$

kubostat2017f (<http://goo.gl/76c4s>) 統計モデリング入門 2017 (f) 2017-06-26 73 / 75

階層ベイズモデル (HBM)? or GLMM? Model: HBM and GLMM are equivalent

- Model: HBM and GLMM are same
- Estimation: **NOT** same
  - Hierarchical Bayesian model (HBM) is **better** because we can apply MCMC estimation.
  - Maximum likelihood estimation (MLE) is **NOT** easy!

kubostat2017f (<http://goo.gl/76c4s>) 統計モデリング入門 2017 (f) 2017-06-26 74 / 75

階層ベイズモデル (HBM)? or GLMM? Model: HBM and GLMM are equivalent

## 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

The development of linear models

kubostat2017f (<http://goo.gl/76c4s>) 統計モデリング入門 2017 (f) 2017-06-26 75 / 75