統計モデリング入門 2018 (g)

階層ベイズモデルと時間変化モデル

久保拓弥 kubo@ees.hokudai.ac.jp

北大環境科学院の講義 http://goo.gl/76c4i

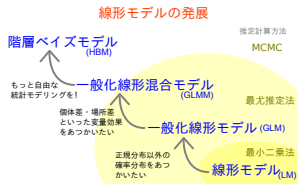
2018–07–09

ファイル更新時刻: 2018–07–06 15:45

この時間で説明したいこと

1. 複数ランダム効果の階層ベイズモデル
   - 個体差 + グループ差，など
2. 時間変化の階層ベイズモデル
   - 一回だけの変化: “対応のある” (paired) データセット

GLMM is a simplified Hierarchical Bayesian Model



一般化線形混合モデル (Generalized Linear Mixed Model) **は階層ベイズモデルのひとつ**

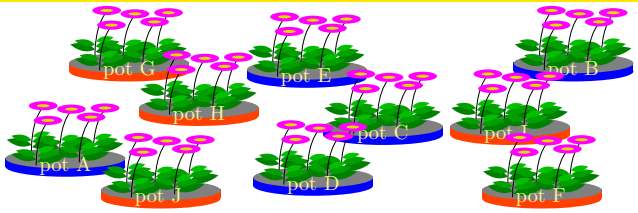
- GLMM: (individual differences) + (group differences) + $\cdots$
- HBM: to estimate posterior distributions

1. 複数ランダム効果の階層ベイズモデル

個体差 + グループ差，など

You can not neglect these “differences”

seed number data, complicated design



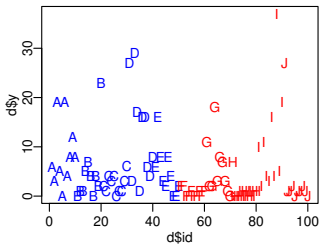
- **肥料をやったら個体ごとの種子数** y_i **が増えるかどうかを調べたい**
- **植木鉢** 10 **個，各鉢に** 10 **個体の架空植物 (合計** 100 **個体)**
  - **コントロール** ($f_j =$ **C**) 5 **鉢 (合計** 50 **個体)**
  - **肥料をやる処理** ($f_j =$ **T**) 5 **鉢 (合計** 50 **個体)**

y: number of seeds

```
> d <- read.csv("d1.csv")
> head(d)
  id pot f  y
1  1   A C  6
2  2   A C  3
3  3   A C 19
4  4   A C  5
5  5   A C  0
6  6   A C 19
```

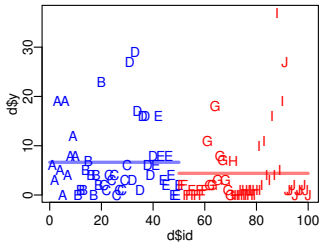
- `id` **列: 個体番号** $\{1, 2, 3, \cdots, 100\}$
- `pot` **列: 植木鉢名** {A, B, C, $\cdots$, J}
- `f` **列: 処理: コントロール** `C`, **肥料** `T`
- `y` **列: 種子数 (応答変数)**

データはとにかく図示する!!



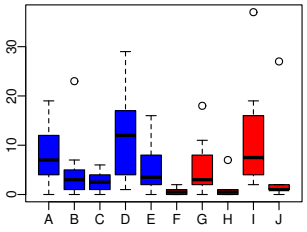
- `plot(d$id, d$y, pch = as.character(d$pot), ...)`
- **コントロール・処理 でそんなに差がない**?

Plot your data!



- **むしろ 処理 のほうが平均種子数が低い**?
- **(注) この架空データは 肥料の効果はゼロ と設定して生成した**

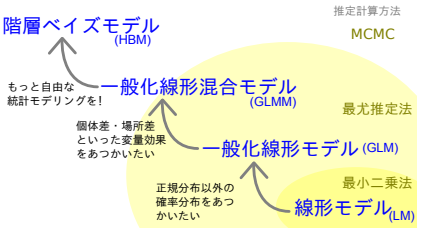
individual + pot differences



- `plot(d$pot, d$y, col = rep(c("blue", "red"), each = 5))`
- **植木鉢由来の** random effects **みたいなものはブロック差と呼ばれる**

(一般化な) 線形モデルのわくぐみで，とりあえず考えてみる

線形モデルの発展



GLM: 個体差もブロック差も無視

```
> summary(glm(y ~ f, data = d, family = poisson))
...(略)...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.8931     0.0549   34.49  < 2e-16
fT           -0.4115     0.0869   -4.73  2.2e-06
...(略)...
```

- **肥料をやる処理** (f) **をすると，平均種子数が下がる**?
- AIC **でモデル選択しても同じような結果に**

GLMM: 個体差だけ考慮，ブロック差は無視

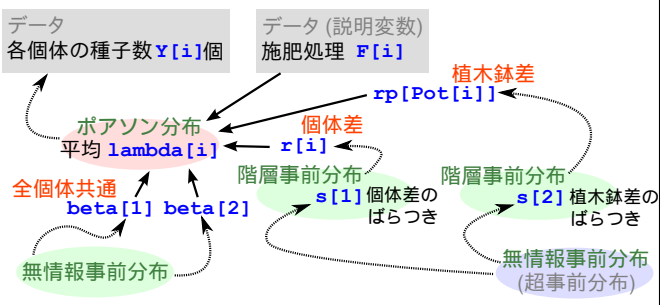
```
> library(glmmML)
> summary(glmmML(y ~ f, data = d, family = poisson,
+ cluster = id))
...(略)...
               coef se(coef)     z Pr(>|z|)
(Intercept)  1.351    0.192  7.05  1.8e-12
fT          -0.737    0.280 -2.63  8.4e-03
...(略)...
```

- **やっぱり同じ**?
- **むしろ肥料処理の悪影響が強い**?

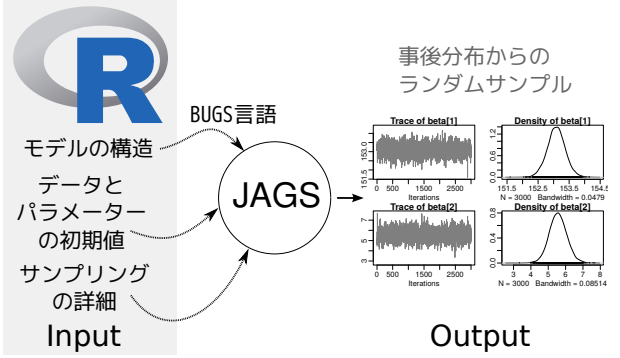
HBM: individual + block differences

- ここでは log リンク関数を使う
- 平均の対数 $\log(\lambda_i) = a + bf_i +$ (個体差) + (ブロック差)
- 事前分布の設定
  - 切片 a と f_i の係数 b は無情報事前分布 (すごく平らな正規分布)
  - 個体差とブロック差は階層的な事前分布 (それぞれ標準偏差 σ_1, σ_2 の正規分布，平均はゼロ)
  - 標準偏差 σ_* は無情報事前分布 ($[0, 10^4]$ の一様分布)

Diagram of the Hierarchical Bayesian Model



JAGS: to estimate posterior distributions



個体差 + ブロック差のあるポアソン回帰の BUGS code (1)

```
model
{
  for (i in 1:N.sample) {
    Y[i] ~ dpois(lambda[i])
    log(lambda[i]) <- a + b * F[i] + r[i] + rp[Pot[i]]
  }
  # 次のページの事前分布の定義につづく
```

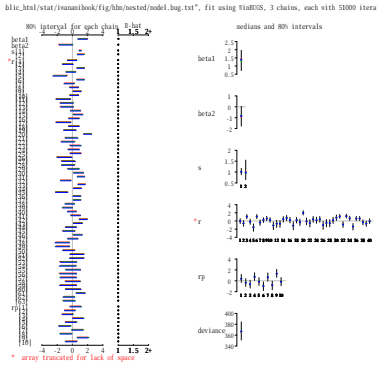
ここでの BUGS coding のポイント

- 因子型の説明変数 $f_i \in \{\mathtt{C}, \mathtt{T}\}$ は，それぞれ F[i] を 0, 1 と置きかえる
- Pot[i] は 1, 2, ..., 10 と数字になおした植木鉢名をいれておいて，植木鉢の効果 rp[...] を参照させる

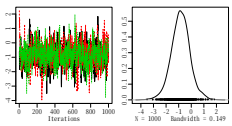
個体差 + ブロック差のあるポアソン回帰の BUGS code (2)

```
  # 前のページからのつづき
  a ~ dnorm(0, 1.0E-4) # 切片
  b ~ dnorm(0, 1.0E-4) # 肥料の効果
  for (i in 1:N.sample) {
    r[i] ~ dnorm(0, tau[1])  # 個体差
  }
  for (j in 1:N.pot) {
    rp[j] ~ dnorm(0, tau[2]) # 植木鉢の差（ブロック差）
  }
  for (k in 1:N.tau) {
    tau[k] <- 1.0 / (sigma[k] * sigma[k]) # 個体・植木鉢のばらつき
    sigma[k] ~ dunif(0, 1.0E+4)
  }
}
```


MCMC sampling



Yes! no fertalization effects (b)

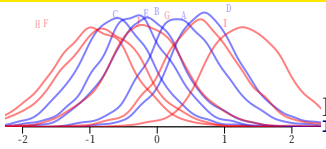


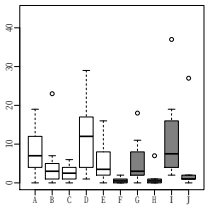
	mean	sd	2.5%	25%	50%	75%	97.5%	Rh
a	1.501	0.529	0.482	1.157	1.493	1.852	2.565	1.0
b	-1.016	0.706	-2.436	-1.476	-0.993	-0.565	0.395	1.0
sigma[1]	1.020	0.114	0.822	0.939	1.014	1.089	1.265	1.0

...(略)...

Example data was generated under "(fertalization effects = 0)"

Posteriors of Block (or Pot)





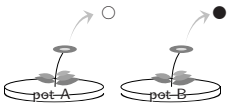
Do not neglect individual and groups differences!

- random effects つまり 個体差・ブロック差が大きい
- random effects の影響が大きいときには，fixed effects の大きさが見えにくくなる— ニセの「効果」が見えることもあれば，見えるはずの傾向が隠されることも
  - 個体差・ブロック差の階層ベイズモデルが必要!
- もしブロック差を人為的に小さくできないなら，ブロック数をもっと増やして，より正確な植木鉢の効果のばらつきを正確に推定するしかない

differences both in plants and pots

個体差 + 場所差の GLMM I

(A) 個体・植木鉢が反復

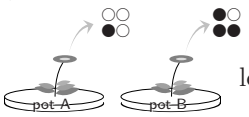


個体差も植木鉢差も推定できない

$\mathrm{logit} q_i = \beta_1 + \beta_2 x_i$ (GLM)

q_i: 種子の生存確率

(B) 個体は擬似反復，植木鉢は反復



個体差は推定できる
植木鉢差は推定できない

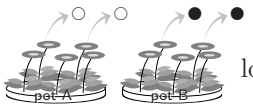
$\mathrm{logit} q_i = \beta_1 + \beta_2 x_i + r_i$

より正確にいうと (A) (B) は個体差と植木鉢差の区別がつかない

differences both in plants and pots

個体差 + 場所差の GLMM II

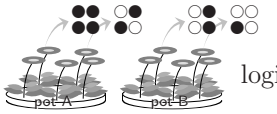
(C) 個体は反復，植木鉢は擬似反復



個体差は推定できない
植木鉢差は推定できる

$\mathrm{logit} q_i = \beta_1 + \beta_2 x_i + r_j$

(D) 個体・植木鉢が擬似反復



個体差も植木鉢差も推定できる

$\mathrm{logit} q_i = \beta_1 + \beta_2 x_i + r_i + r_j$

複雑なモデルほど最尤推定は困難，しかも多くのデータが必要

GLMM は階層ベイズモデル (HBM) で!

- 現実のデータ解析では個体差・場所差の効果を統計モデルに組みこまなければならない
- これらは歴史的には random effects とよばれてきた
- 用語の整理: 統計モデルには global parameter と local parameter があると考えればよい
- GLMM では global parameter を最尤推定する— local parameter は積分して消す
- local parameter が増えると (e.g. 個体差 + 場所差) 最尤推定が難しい → 階層ベイズモデル (Hierarchical Bayesian Model) で事後分布 (posterior) 推定!

2. 時間変化の階層ベイズモデル

一回だけの変化: “対応のある” (paired) データセット

架空の実験：給食タイプ→小学生の身長の伸び?

岩波データサイエンス vol.1

久保が書いた階層ベイズモデルの解説記事の例題

架空の実験：給食タイプ→小学生の身長の伸び?

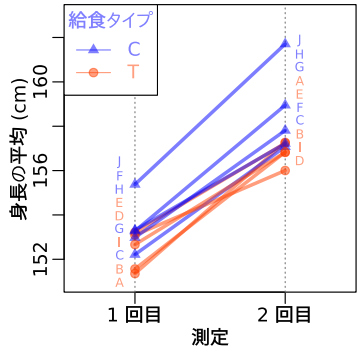
調査地	給食	標本サイズ		身長の平均 (cm)		身長の標準偏差	
(県)	タイプ	1 回目	2 回目	1 回目	2 回目	1 回目	2 回目
A	T	55	51	151.36	157.27	2.94	2.98
B	T	53	49	151.56	156.83	3.07	3.14
C	C	55	53	152.22	157.08	3.20	3.21
D	T	53	52	153.09	156.00	2.65	2.64
E	T	58	55	153.22	157.24	3.07	3.03
F	C	55	53	153.31	157.22	3.10	3.13
G	C	58	53	152.98	157.81	2.49	2.45
H	C	59	57	153.27	158.95	3.08	3.06
I	T	56	51	152.67	156.82	2.82	2.92
J	C	56	50	155.37	161.71	3.10	3.21

・給食タイプ T (新型) : A, B, D, E, I 県
・給食タイプ C (普通) : C, F, G, H, J 県

新型給食 f=T の真の効果は 0!

		Sample size		Mean		SD	
City	Treatment	year 1	year 2	y1	y2	y1	y2
A	T	55	51	151.36	157.27	2.94	2.98
B	T	53	49	151.56	156.83	3.07	3.14
C	C	55	53	152.22	157.08	3.20	3.21
D	T	53	52	153.09	156.00	2.65	2.64
E	T	58	55	153.22	157.24	3.07	3.03
F	C	55	53	153.31	157.22	3.10	3.13
G	C	58	53	152.98	157.81	2.49	2.45
H	C	59	57	153.27	158.95	3.08	3.06
I	T	56	51	152.67	156.82	2.82	2.92
J	C	56	50	155.37	161.71	3.10	3.21

(架空) データ： 給食と身長成長

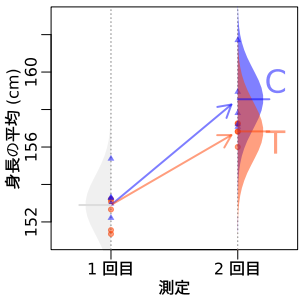


ダメな GLM: bad model 1

調査地	給食	標本サイズ		身長の平均 (cm)		身長の標準偏差	
(県)	タイプ	1 回目	2 回目	1 回目	2 回目	1 回目	2 回目
A	T	55	51	151.36	157.27	2.94	2.98
B	T	53	49	151.56	156.83	3.07	3.14
C	C	55	53	152.22	157.08	3.20	3.21
D	T	53	52	153.09	156.00	2.65	2.64
E	T	58	55	153.22	157.24	3.07	3.03
F	C	55	53	153.31	157.22	3.10	3.13
G	C	58	53	152.98	157.81	2.49	2.45
H	C	59	57	153.27	158.95	3.08	3.06
I	T	56	51	152.67	156.82	2.82	2.92
J	C	56	50	155.37	161.71	3.10	3.21

(例) fit ← glm(y ~ t + t:f, ……)
測定回数： t = **1** または **2** (1 回目, 2 回目)
給食タイプ： f = C または T

ダメな GLM: bad model 1

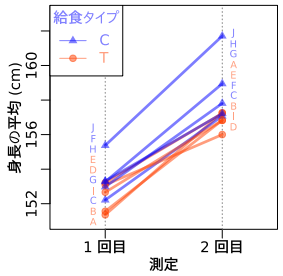


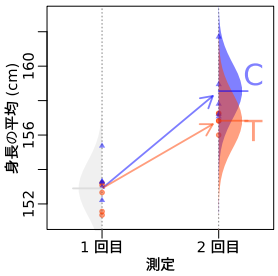
(例) fit ← glm(y ~ t + t:f, ……)

測定回数： t = **1** または **2** (1 回目, 2 回目)

給食タイプ： f = C または T

対応 (paired) が考慮されてない！

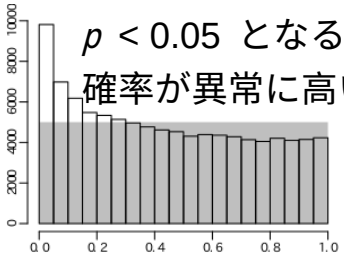




ダメな GLM: bad model 1

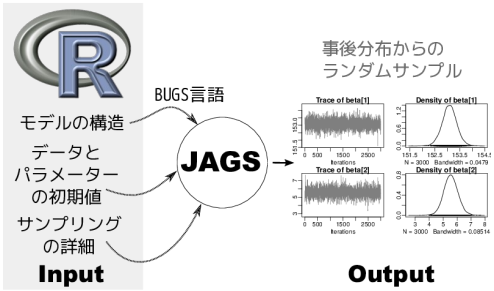
glm(y ~ t + t:f, ……)

bad model 1 による第一種の過誤の悪化

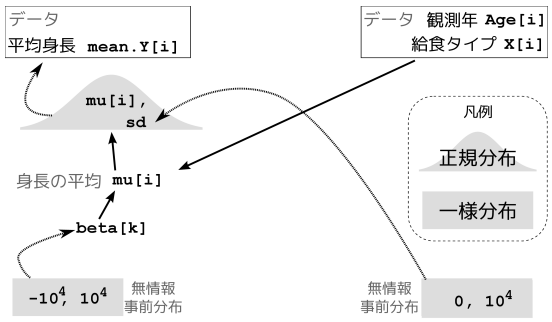


$p < 0.05$ となる
確率が異常に高い

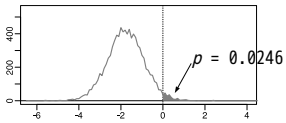
R と JAGS



bad model 1 を Bayes model 化



bad model 1 による第一種の過誤の悪化

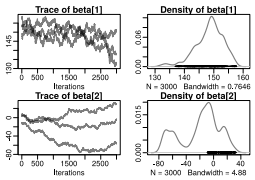


新給食が身長増加に与える効果

しかし…

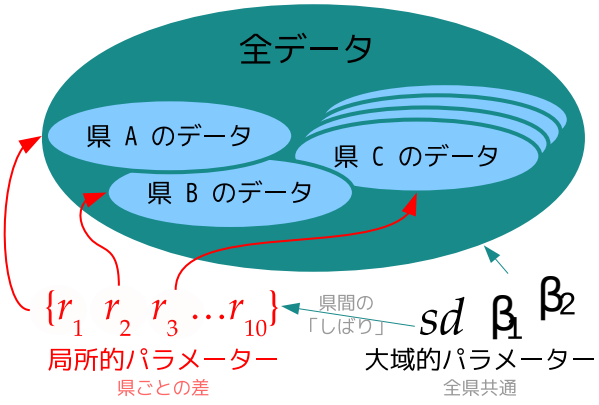
新給食 f=T の真の効果は 0!

bad model 2: 各県独立 Bayes 版

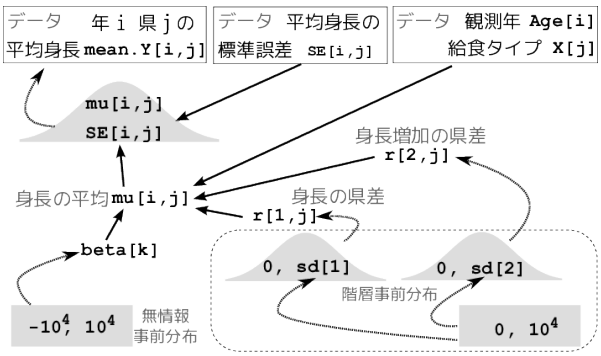


Bayes model でもダメなものはダメ…

Hierarchical Bayesian Model



Hierarchical Bayesian Model

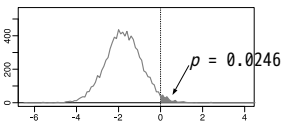


Hierarchical Bayesian Model

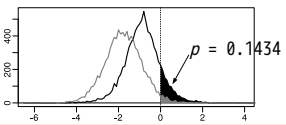
```
model
{
    for (i in 1:2) { # age
        for (j in 1:N.pref) {
            Y.mean[i, j] ~ dnorm(mu[i, j], Tau.se[i, j])
            mu[i, j] <- beta[1] + r[1, j] + (
                beta[2] + beta[3] * X[i, j] + r[2, j]
            ) * Age[i, j]
        }
    }
    for (k in 1:N.beta) {
        beta[k] ~ dunif(-1.0E+4, 1.0E+4)
    }
    for (i in 1:N.r) {
        for (j in 1:N.pref) {
            r[i, j] ~ dnorm(0, tau[i])
        }
        tau[i] <- 1 / (sd[i] * sd[i])
        sd[i] ~ dunif(0, 1.0E+4)
    }
}
```


Hierarchical Bayesian Model による推定結果

bad model 1



HBM



新型給食 f=T の真の効果は 0!

各県の local parameter