統計モデリング入門 2015 (f)

Generalized Linear Mixed Model (GLMM)
一般化線形混合モデル

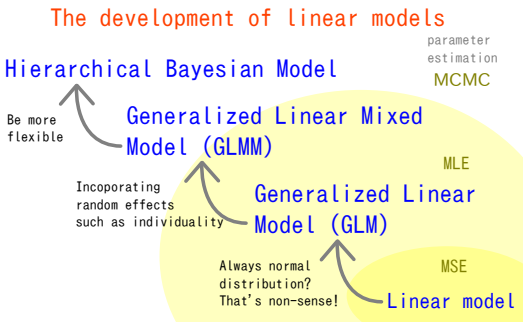
久保拓弥 kubo@ees.hokudai.ac.jp

北大環境科学院の講義 http://goo.gl/76c4i

2015–07–27

ファイル更新時刻: 2015–07–28 19:35

statistaical models appeared in the class
この授業であつかう統計モデルたち



Kubo Doctrine: "Learn the evolution of linear-model family, firstly!"

今日のハナシ

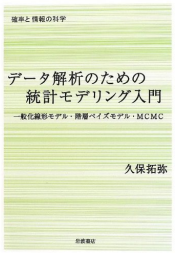
1. GLM では説明できない種子データ
   overdispersion data
   「ばらつき」が大きすぎる!
2. 過分散と個体差
   overdispersion caused by individual differences
   観測されていない個体差がもたらす過分散
3. 一般化線形混合モデル
   Genralized Liear Mixed Model
   個体差をあらわすパラメーターを追加
4. 一般化線形混合モデルの最尤推定
   個体差 r_i を積分して消す尤度方程式
5. 現実のデータ解析には GLMM が必要
   個体差・場所差を考えないといけないから

今日の内容と「統計モデリング入門」との対応

今日はおもに「第 7 章 一般化線形混合モデル (GLMM)」の内容を説明します.

- 著者: 久保拓弥
- 出版社: 岩波書店
- 2012–05–18 刊行

http://goo.gl/Ufq2



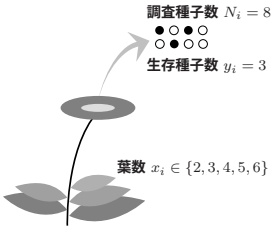
1. GLM では説明できない種子データ

overdispersion data
「ばらつき」が大きすぎる!

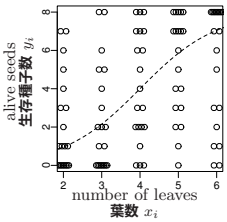
過分散 (overdispersion) とは何か?

example seed survivorship again, but ...
今日の例題: 種子の生存確率……前回と同じ?!

(A) 個体 i で観測されたデータ



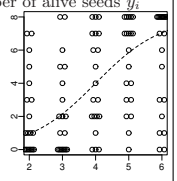
(B) 全 100 個体の x_i と y_i



logistic regression as usual?

“N 個中の y 個” というデータ → ロジスティック回帰?

number of alive seeds y_i



number of leaves x_i

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

linear predictor

- 線形予測子: $\beta_1 + \beta_2 x_i$

link function

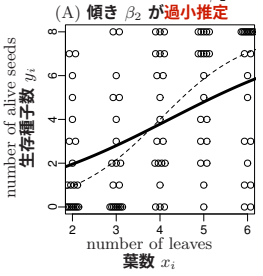
- リンク関数: logit リンク関数

GLM doesn't work!

GLM では説明できないばらつき!

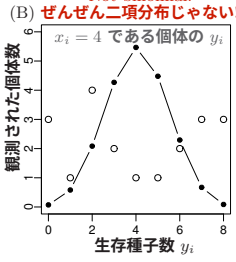
underestimated β_2

(A) 傾き β_2 が過小推定



Not binomial!

(B) ぜんぜん二項分布じゃない!



○が観測されたデータの図示

overdispersion caused by individual differences

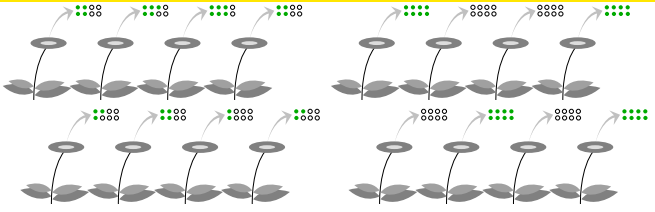
2. 過分散と個体差

観測されていない個体差がもたらす過分散

unobservable differences

観測されてない個体差って?

過分散 (overdispersion) とは何か?

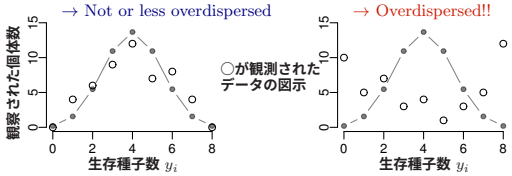


(A) 個体差のばらつきが小さい場合

→ Not or less overdispersed

(B) 個体差のばらつきが大きい場合

→ Overdispersed!!



ロジスティック回帰やポアソン回帰 といった GLM では 全サンプルの均質性を仮定している

GLM does not take into account individual differences

現実のカウントデータは ほとんど過分散

Almost all “real” data are overdispersed!

Genralized Liear Mixed Model

3. 一般化線形混合モデル

個体差をあらわすパラメーターを追加

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

an improvement of logistic regression model

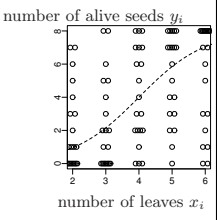
ロジスティック回帰のモデルを改良する

ロジスティック回帰のモデル

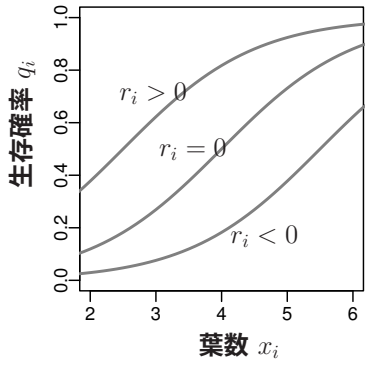
probability distribution binomial distribution
- 確率分布 : 二項分布

linear predictor
- 線形予測子: $\beta_1 + \beta_2 x_i + r_i$

link function
- リンク関数: logit リンク関数

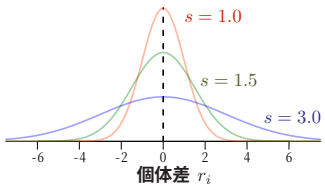


個体 i の個体差を r_i としてみよう



suppose $\{r_i\}$ follow the Gausssian distribution

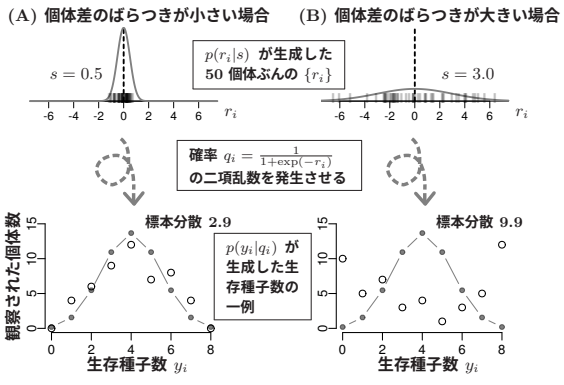
$\{r_i\}$ のばらつきは正規分布だと考えてみる



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

この確率密度 $p(r_i \mid s)$ 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」.

個体差 r_i の分布と過分散の関係



a numerical experiment using random numbers

ちょっと乱数を使った数値実験をしてみましょう

```
> # defining logistic function
> logistic <- function(z) { 1 / (1 + exp(-z)) }
> # random numbers following binomial distribution
> rbinom(100, 8, prob = logistic(0))
> # random numbers following Gausssian distribution
> rnorm(100, mu = 0, sd = 0.5)
> r <- rnorm(100, mu = 0, sd = 0.5)
> # random numbers following ... ?
> rbinom(100, 8, prob = logistic(0 + r))
```

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

Generalized Linear Mixed Model (GLMM)

で使う Mixed **な 線形予測子**: $\underset{\text{linear predictor}}{\beta_1 + \beta_2 x_i} + r_i$

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

fixed? random? **よくわからん……?**

global parameter と local parameter

Generalized Linear Mixed Model (GLMM)

で使う Mixed **な 線形予測子**: $\underset{\text{linear predictor}}{\beta_1 + \beta_2 x_i} + r_i$

- 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

4. 一般化線形混合モデルの最尤推定

個体差 r_i **を積分して消す尤度方程式**

「積分する」とは分布を混ぜること

個体差 r_i は最尤推定できない

local parameters: $\{r_1, r_2, \cdots, r_{100}\}$

全 100 個体に対して，個体ごとにいちいち r_i **の値を最尤推定すると**

飽和モデル ($\underset{\text{saturation model}}{}$) **の推定になってしまう**

```
> d <- read.csv("data.csv")
> head(d)
  N y x id
1 8 0 2  1
2 8 1 2  2
3 8 2 2  3
4 8 4 2  4
5 8 1 2  5
6 8 0 2  6
```


尤度関数の中で r_i を積分してしまえばよい

データ y_i **のばらつき** — $\underset{\text{binomial distribution}}{}$ **二項分布**

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

個体差 r_i **のばらつき** — $\underset{\text{Gaussian distribution}}{}$ **正規分布**

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

個体 i **の** $\underset{\text{likelihood}}{}$ **尤度** — $\underset{\text{to remove } r_i}{}$ r_i **を消す**

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

$\underset{\text{likedhood for all data}}{}$ **全データの尤度** — β_1, β_2, s **の関数**

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

global parameter と local parameter

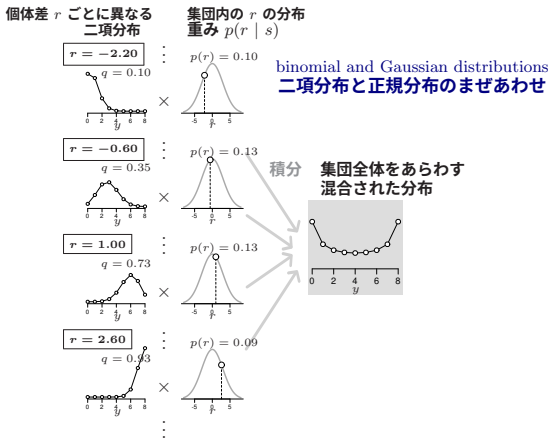
Generalized Linear Mixed Model (GLMM)

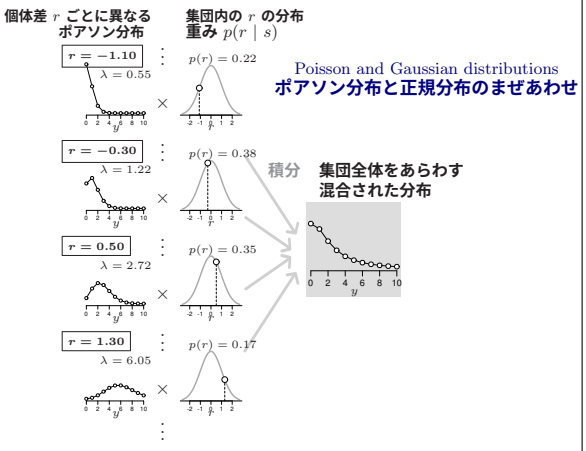
で使う Mixed **な 線形予測子**: $\underset{\text{linear predictor}}{\beta_1 + \beta_2 x_i} + r_i$

- global parameter **は最尤推定できる**
  - fixed effects: β_1, β_2
  - **全個体のばらつき**: s
- local parameter **は最尤推定できない**
  - random effects: $\{r_1, r_2, \cdots, r_{100}\}$

個体差 r_i について積分するということは二項分布と正規分布をまぜあわせること

Integral of $r_i \rightarrow$ mixture distribution of the binomial and Gaussian distributions





glmmML package を使って GLMM の推定

```
> install.packages("glmmML") # if you don't have glmmML
> library(glmmML)
> glmmML(cbind(y, N - y) ~ x, data = d, family = binomial
+ cluster = id)

> d <- read.csv("data.csv")
> head(d)
  N y x id
1 8 0 2  1
2 8 1 2  2
3 8 2 2  3
4 8 4 2  4
5 8 1 2  5
6 8 0 2  6
```


GLMM の 推定値: $\hat{\beta}_1, \hat{\beta}_2, \hat{s}$ (estimates)

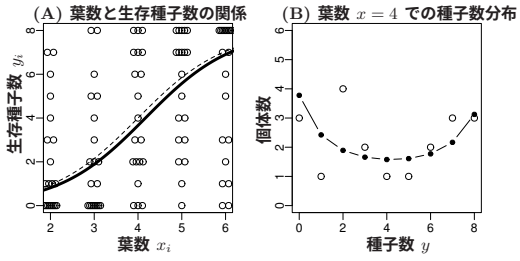
```
> glmmML(cbind(y, N - y) ~ x, data = d, family = binomial,
+ cluster = id)
...(snip)...
             coef se(coef)     z Pr(>|z|)
(Intercept) -4.13    0.906 -4.56  5.1e-06
x            0.99    0.214  4.62  3.8e-06

Scale parameter in mixing distribution:  2.49 gaussian
Std. Error:                              0.309

Residual deviance: 264 on 97 degrees of freedom  AIC: 270
```

$$\hat{\beta}_1 = -4.13, \hat{\beta}_2 = 0.99, \hat{s} = 2.49$$

推定された GLMM を使った 予測 (prediction)



5. 現実のデータ解析には GLMM が必要

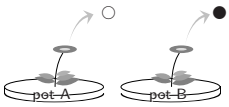
個体差・場所差を考えないといけないから

反復・擬似反復に注意しよう

differences both in plants and pots

個体差 + 場所差の GLMM I

(A) 個体・植木鉢が反復

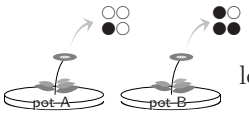


個体差も植木鉢差も推定できない

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

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

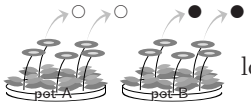
$$\mathrm{logit} q_i = \beta_1 + \beta_2 x_i + r_i$$

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

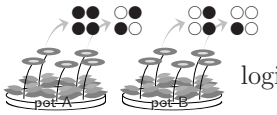
(C) 個体は反復，植木鉢は擬似反復



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

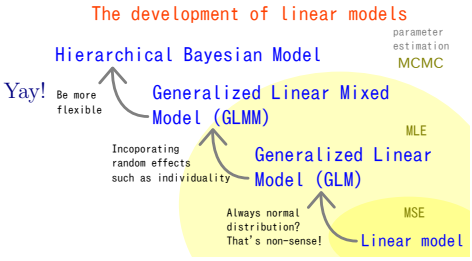
summary

GLMM まとめ

- 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
- 実際のところは — 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) パラメーター推定がたいへんになる — ということで ……

次回予告

The next topic



階層ベイズモデル

Hierarchical Bayesiam Model (HBM)