

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

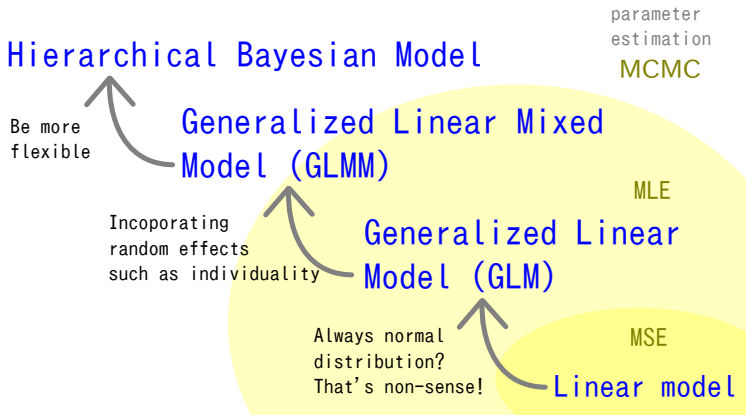
北大環境科学院の講義 <http://goo.gl/76c4i>

2015-07-27

ファイル更新時刻: 2015-07-28 19:35

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

The development of linear models



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!

overdispersion caused by individual differences

② 過分散と個体差

観測されていない個体差をもたらす過分散

Generalized Linear 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

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

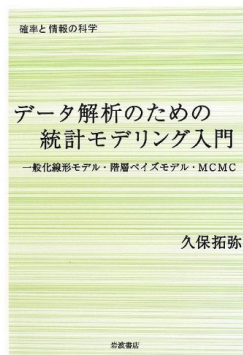
個体差・場所差を考えないといけないから

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>



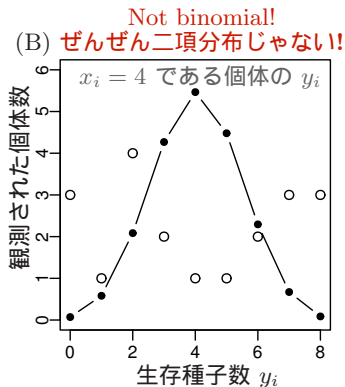
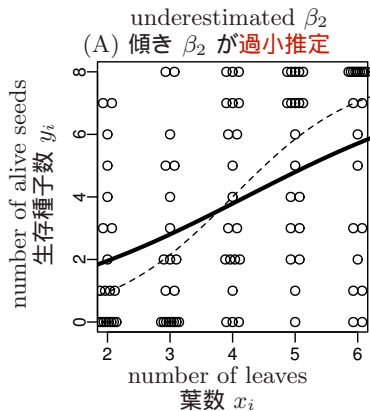
1. GLM では説明できない種子データ

overdispersion data
「ばらつき」が大きすぎる!

過分散 (overdispersion) とは何か?

GLM doesn't work!

GLM では説明できないばらつき!



が観測されたデータの図示

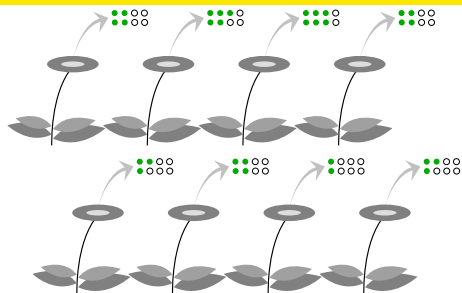
overdispersion caused by individual differences

2. 過分散と個体差

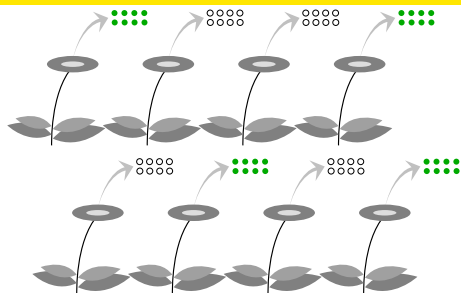
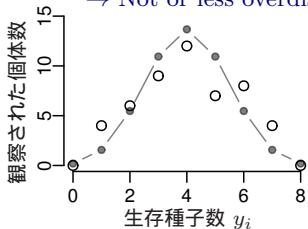
観測されていない個体差がもたらす過分散

unobservable differences
観測されていない個体差って？

過分散 (overdispersion) とは何か?

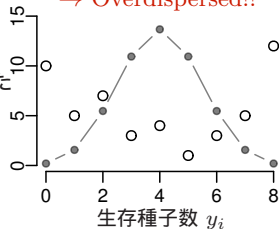


(A) 個体差のばらつきが小さい場合
→ Not or less overdispersed



(B) 個体差のばらつきが大きい場合
→ Overdispersed!!

が観測された
データの図示



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!

Generalized Linear 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

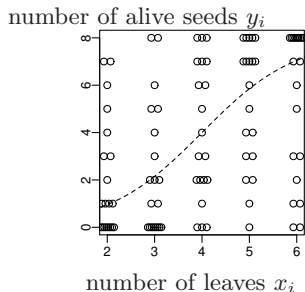
- 確率分布 : 二項分布

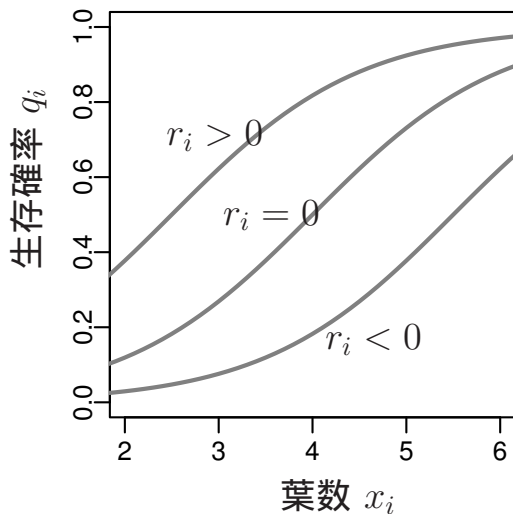
linear predictor

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

link function

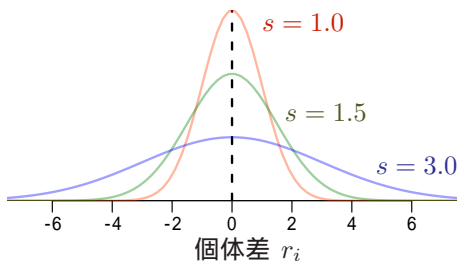
- リンク関数: logit リンク関数



個体 i の個体差を r_i としてみよう

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

$\{r_i\}$ のばらつきは正規分布だと考えてみる

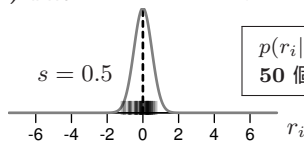


$$p(r_i | s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{r_i^2}{2s^2}\right)$$

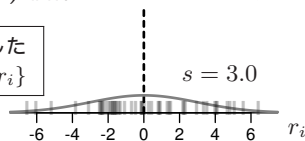
この確率密度 $p(r_i | s)$ は r_i の「出現しやすさ」をあらわしていると解釈すればよいでしょう。 r_i がゼロに近い個体はわりと「ありがち」で、 r_i の絶対値が大きな個体は相対的に「あまりいない」。

個体差 r_i の分布と過分散の関係

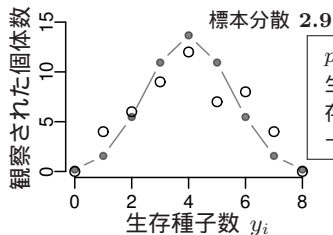
(A) 個体差のばらつきが小さい場合 (B) 個体差のばらつきが大きい場合



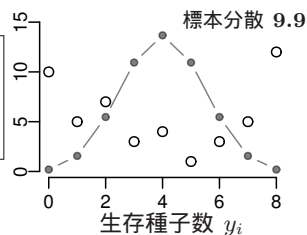
$p(r_i|s)$ が生成した
50 個体ぶんの $\{r_i\}$



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



$p(y_i|q_i)$ が
生成した生
存種子数の
一例



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 Gaussssian 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

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.....?

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

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: $\{r_1, r_2, \dots, r_{100}\}$

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
saturation model

飽和モデル 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 — binomial distribution
二項分布

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

個体差 r_i のばらつき — Gaussian distribution
正規分布

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

個体 i の 尤度 — likelihood to remove r_i
 r_i を消す

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

likelihood for all data
全データの尤度 — 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)

linear predictor

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

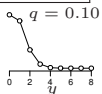
- global parameter は最尤推定できる
 - fixed effects: β_1, β_2
 - 全個体のばらつき: s
- local parameter は最尤推定できない
 - random effects: $\{r_1, r_2, \dots, r_{100}\}$

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せ
あわせること

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

個体差 r ごとに異なる
二項分布

$$r = -2.20$$



⋮

×

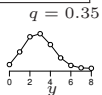
集団内の r の分布
重み $p(r | s)$

$$p(r) = 0.10$$



binomial and Gaussian distributions
二項分布と正規分布のまぜあわせ

$$r = -0.60$$



⋮

×

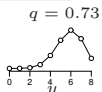
$$p(r) = 0.13$$



積分

集団全体をあらわす
混合された分布

$$r = 1.00$$



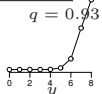
⋮

×

$$p(r) = 0.13$$



$$r = 2.60$$



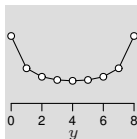
⋮

×

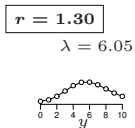
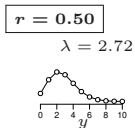
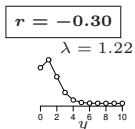
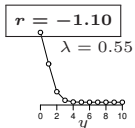
$$p(r) = 0.09$$



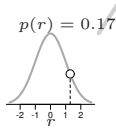
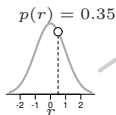
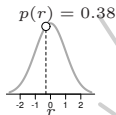
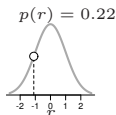
⋮



個体差 r ごとに異なる
ポアソン分布

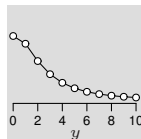


集団内の r の分布
重み $p(r | s)$



Poisson and Gaussian distributions
ポアソン分布と正規分布のまぜあわせ

積分
集団全体をあらわす
混合された分布



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
```

estimates

GLMM の推定値: $\hat{\beta}_1, \hat{\beta}_2, \hat{s}$

```
> 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

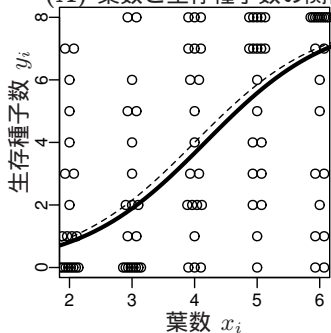
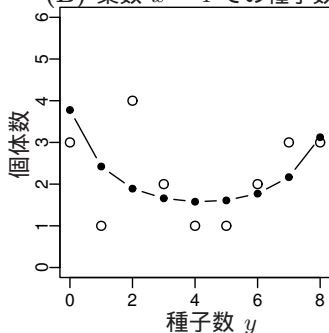
Residual deviance: 264 on 97 degrees of freedom AIC: 270

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

prediction

推定された GLMM を使った 予測

(A) 葉数と生存種子数の関係

(B) 葉数 $x = 4$ での種子数分布

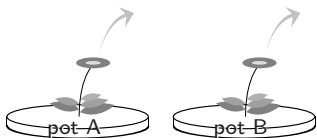
5. 現実のデータ解析には GLMM が必要

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

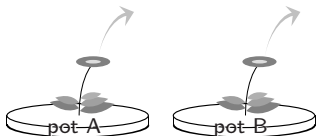


個体差も植木鉢差も
推定できない

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

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

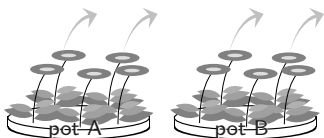
$$\text{logit}q_i = \beta_1 + \beta_2 x_i + r_i$$

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

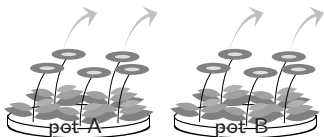
(C) 個体は反復，植木鉢は擬似反復



個体差は推定できない
植木鉢差は推定できる

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

(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

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

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

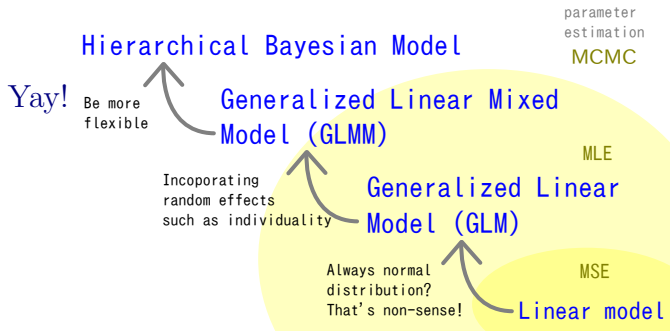
GLMM まとめ

- 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

The development of linear models



階層ベイズモデル

Hierarchical Bayesian Model (HBM)