

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がもたらす過分散
- ③ Generalized Linear Mixed Model
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 v_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

<http://goo.gl/Ufq2>

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

GLM では説明できない種子データ 「ばらつき」が大きすぎる!

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

調査種子数 $N_i = 8$

生存種子数 $y_i = 3$

葉数 $x_i \in \{2, 3, 4, 5, 6\}$

(B) 全 100 個体の x_i と y_i

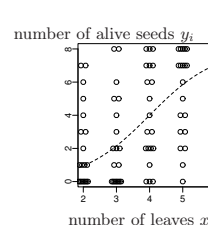
GLM では説明できない種子データ 「ばらつき」が大きすぎる!

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?

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



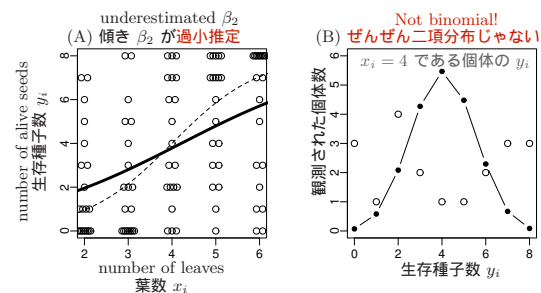
number of alive seeds y_i

number of leaves x_i

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 7 / 35

GLM では説明できない種子データ 「ばらつき」が大きすぎる!

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!
ぜんぜん二項分布じゃない!
 $x_i = 4$ である個体の y_i

number of alive seeds 生存種子数 y_i

number of leaves 葉数 x_i

観測されたデータの図示

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 8 / 35

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

overdispersion caused by individual differences

2. 過分散と個体差

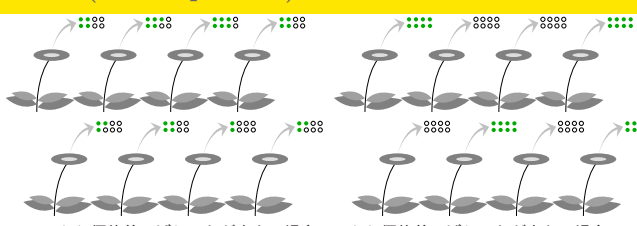
観測されていない個体差がもたらす過分散

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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 9 / 35

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

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!!

観測されたデータの図示

number of alive seeds 生存種子数 y_i

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 10 / 35

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

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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 11 / 35

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

現実のカウントデータは ほとんど過分散

Almost all “real” data are overdispersed!

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 12 / 35

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

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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 13 / 35

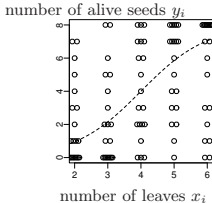
一般化線形混合モデル 個体差をあらわすパラメーターを追加

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



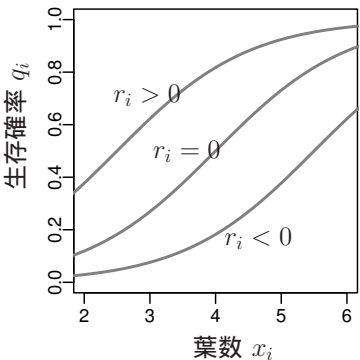
number of alive seeds y_i

number of leaves x_i

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 14 / 35

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

個体 i の個体差を r_i としてみよう



生存確率 q_i

葉数 x_i

$r_i > 0$

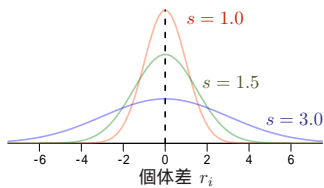
$r_i = 0$

$r_i < 0$

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 15 / 35

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

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



個体差 r_i

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

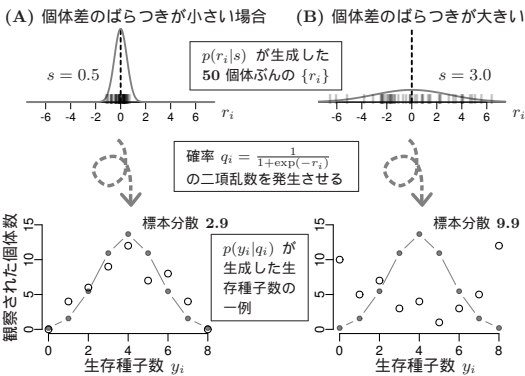
この確率密度 $p(r_i | s)$ は r_i の「出現しやすさ」をあらわしていると解釈すればよいでしょう。 r_i がゼロに近い個体はわりと「ありがち」で、 r_i の絶対値が大きな個体は相対的に「あまりいない」。

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 16 / 35

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

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



$s = 0.5$ $s = 3.0$

$p(r_i | s)$ が生成した 50 個体ぶん $\{r_i\}$

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

生存種子数 y_i

標本分散 2.9

標本分散 9.9

$p(y_i | q_i)$ が生成した生存種子数の一例

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 17 / 35

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

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 Gaussian 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))
    
```

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 18 / 35

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

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.....?

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 19 / 35

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

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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 20 / 35

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

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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 21 / 35

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

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
```

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 22 / 35

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

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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 23 / 35

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

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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 24 / 35

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

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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 25 / 35

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

個体差 r ごとに異なる二項分布 \times 集団内の r の分布 重み $p(r | s)$

binomial and Gaussian distributions
二項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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 26 / 35

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

個体差 r ごとに異なるポアソン分布 \times 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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 27 / 35

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

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
```

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 28 / 35

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

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

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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 29 / 35

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

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

生存種子数 y_i 葉数 x_i 種子数 y 個体数

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 30 / 35

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

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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 31 / 35

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

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

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

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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 32 / 35

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

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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 33 / 35

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

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

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 34 / 35

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

次回予告

The next topic

The development of linear models

Hierarchical Bayesian Model

Yay! Be more flexible

Generalized Linear Mixed Model (GLMM)

Incorporating random effects such as individuality

Generalized Linear Model (GLM)

Always normal distribution? That's non-sense!

Linear model

parameter estimation MCMC

MLE

MSE

階層ベイズモデル

Hierarchical Bayesian Model (HBM)

kubostat2015f (http://goo.gl/76c4i) 統計モデリング入門 2015 (f) 2015-07-27 35 / 35