第７章 一般化線形混合モデル（ＧＬＭＭ）
―個体差のモデリング―

新納研究室

12T4069L 佐鳥 恭太郎

はじめに

- 一般化線形モデル（GLM）
  - 第 3 章例題にて，説明変数が全て同じならば，どの個体の種子数も平均 λ のポアソン分布に従うはず。
  - しかし現実には，説明変数以外の要因によって平均 λ は変化する。
- 一般化線形混合モデル（generalize linear mixed model， GLMM）
  - データのばらつきは二項分布・ポアソン分布，固体のばらつきは正規分布で表す。
  - 複数の確率分布を部品とする統計モデル

例題：ＧＬＭでは説明できないカウントデータ

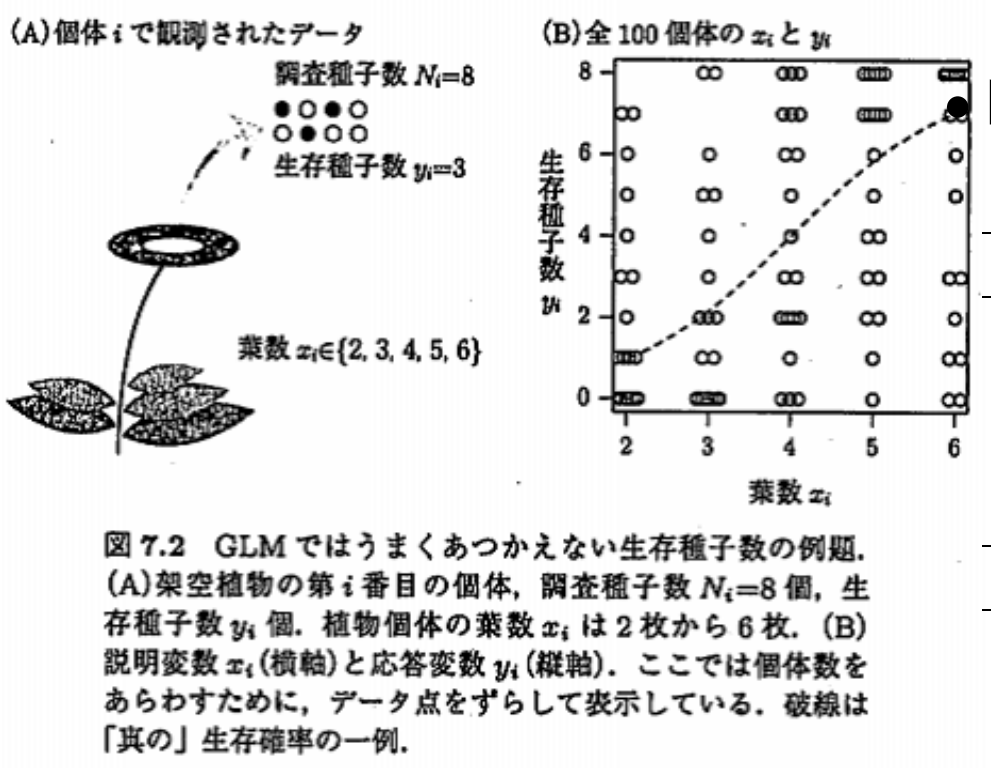

図 7.2　GLM ではうまくあつかえない生存種子数の例題．(A)架空植物の第 i 番目の個体，調査種子数 N_i=8 個，生存種子数 y_i 個．植物個体の葉数 x_i は 2 枚から 6 枚．(B) 説明変数 x_i(横軸)と応答変数 y_i(縦軸)．ここでは個体数をあらわすために，データ点をずらして表示している．破線は「真の」生存確率の一例．

問題設定：

- 個体数：100
- 個体 i に対して
  - ＊調査種子数 $N_i = 8$ 個
  - ＊そのうち生存していた種子数 y_i
  - ＊葉数 $x_i \in \{2, 3, 4, 5, 6\}$
- 葉数ごとの調査個体数は 20
- 種子の生存確率 q_i が，個体ごとに異なる葉数 x_i に依存する

ＧＬＭで試す

- ロジットリンク関数：

$$logit(q_i) = \beta_1 + \beta_2 x_i$$

- 観測された生存種子数がy_iである確率が二項分布に従うとすると，

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

- 全体の対数尤度は$\log L = \sum_i \log p(y_i|\,\beta_1, \beta_2)$となる
  - 最尤推定値：切片 $\hat{\beta}_1 = -2.15$で傾き $\hat{\beta}_2 = 0.51$

ＧＬＭで試した結果

- 図（Ａ）より，真の傾きと比べて，推定された傾きは小さい
- 図（Ｂ）より，推定されたモデルでは，
  - $x_i = 4$のとき，生存確率が$logistic(-2.15 + 0.51 * 4) = 0.47$である二項分布となるはず
  - しかし，データは二項分布に従っているようには見えない

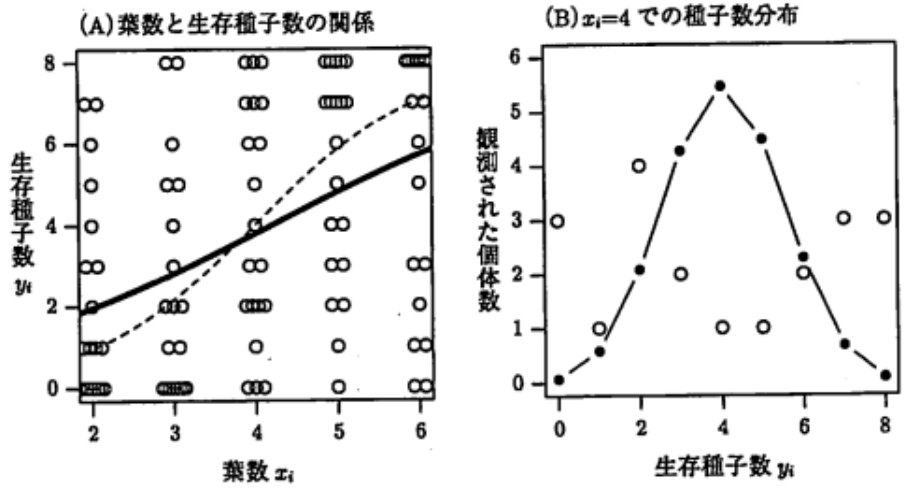

図 7.3　ロジスティック回帰がうまくいかない例題．(A)図 7.2(B)のデータの上に GLM の予測結果(実線)を重ねたもの．真の葉数依存性(破線)より小さい傾きが推定されている．(B)葉数 x_i=4 における種子数分布(白丸)と，推定された GLM から予測される二項分布(黒丸と実線)．

過分散

- 過分散
  - 二項分布で説明できないデータを二項分布と仮定したので，正しい推定値が得られなかった
  - 二項分布で期待されるよりも大きなばらつきのことを過分散という

個体差

- 観測されていない個体差がもたらす過分散
  - 極端な過分散の例：ある葉数について
  - 半分の個体は生存種子数がゼロ，
    もう半分の個体は全種子が生存
  - 種子の生存確率=０．５ -> 平均生存種子数=４
    ：生存種子数４の個体はない

  - 8個体の標本分散は$8 * 4^2/8 = 16$
  - 二項分布から期待される分散は
    $Nq(1-q) = 8 * 0.5 * 0.5 = 2$
  - よって，この例は過分散といえる

一般化線形混合モデル

- 個体差を表すパラメータの追加
  - 架空植物の種子の生存確率q_iを表す式に，個体iの個体差を表すパラメータ$r_i(-\infty \leq r_i \leq +\infty)$を追加する

$$logit(q_i) = \beta_1 + \beta_2 x_i + r_i$$

- 個体差の生存確率に与える影響を図に示す
  - 個体差が高いほど生存確率も高くなる
  - 個体差が低いほど生存確率も低くなる

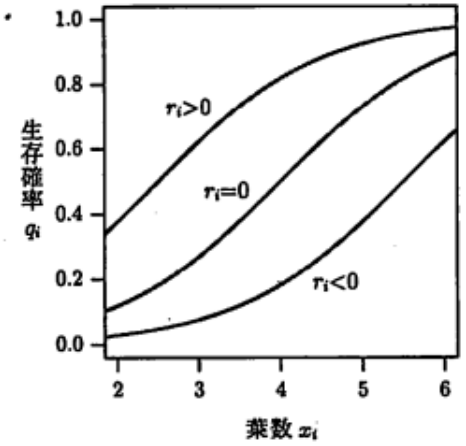

図 7.5　個体差 r_i と生存確率 q_i.

個体差のばらつきを表す確率分布

- ＧＬＭＭの特徴：r_iが何かの確率分布に従っていると仮定する
  - r_iはある確率分布に従う -> データからパラメータ推定できる
  - ただし，切片β_1や傾きβ_2といったパラメータは，とくに何か確率分布に従うと考えているわけではない
- 仮定：個体差r_iは平均ゼロで標準偏差sの正規分布に従う
  - 正規分布の理由：このような統計モデリングに便利だから
  - 単純化のため，各個体のr_iは個体間で相互に独立した確率変数である
  - 確率密度関数$p(r_i|s)$

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

  - r_iの絶対値が小さいほど「ありがち」，大きいほど「あまりいない」ことを表現する
  - 標準偏差sは「集団内の個体差r_iのばらつき」を表している

線形予測子の構成要素：固定効果とランダム効果

- 一般化線形混合モデルの「混合」の意味
  - 統計モデルに線形予測子が含まれている場合，
    線形予測子の要素は
    固定効果(fixed effects)とランダム効果(random effects)
    に分類される
- つまり，ＧＬＭＭは線形予測子に固定効果とランダム効果の項を持っているので，
  混合(mixed)モデルまたは混合効果（mixed effects）モデル
  と呼ばれる
- このモデルの線形予測子$logit(q_i) = \beta_1 + \beta_2 x_i + r_i$の項はそれぞれ以下に相当する
  - 切片β_1と葉数の影響$\beta_2 x_i$は固定効果
  - 個体差r_iはランダム効果

一般化線形混合モデルの最尤推定

- ＧＬＭＭに含まれる個体差r_iは最尤推定できない
  - 100個体ぶんの生存数データy_iを説明するために
    100個のパラメータ$\{\hat{r}_1, \hat{r}_2, \cdots \hat{r}_{100}\}$の値を最尤推定するのは
    フルモデルになってしまうから
- 対処法：個体ごとの尤度L_iの式の中で，r_iを積分する

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

  - $p(y_i|\,\beta_1, \beta_2, r_i)$：二項分布，$p(r_i|s)$：正規分布
  - いろいろなr_iの値における尤度を評価し，その期待値を計算する
  - 期待値は$p(r_i|s)$によって重み付けされる

Rを使ってGLMMのパラメータを推定

- glmmML packageを使用する
  - ダウンロードサイト「http://cran.r-project.org/」から
    インストールする
  - Rのコマンドラインで
    `> library(glmmML)`
    と入力し，パッケージを読み込む
- glmmML()関数による推定
  - r_iが「個体ごとに異なる独立なパラメータ」であることを
    clusterオプションを使って指定する
  - これはデータフレイムdのid列に格納されている個体番号を用
    いればよい

推定結果

```
> glmmML(cbind(y, N - y) ~ x, data = d, family = binomial,
+ cluster = id)
```

するとこのような結果が得られます.

```
             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
```

- 推定結果の読み方
  - coef(係数)：パラメータの最尤推定値,
    $\hat{\beta}_1 = -4.13$（真の値は-4），$\hat{\beta}_2 = 0.99$（真の値は1）
  - Scale parameter... ：「個体差r_iのばらつき」ことsの最尤推定値
  - Std. Error：sの推定値のばらつき（標準誤差）
  - 100このデータにたいして$\{\beta_1, \beta_2, s\}$の3パラメータを使っているので残りの自由度は100-3=97
  - そのときのResidual devianceは264でAICは270

現実のデータ解析にはＧＬＭＭが必要

- 過分散の有無を調べてGLMMを採用する

- GLMMのような考え方が必要かの判断ポイント
  - 同じ個体・場所などから何度もサンプリングしているか
  - 個体差や場所差が識別できてしまうようなデータのとりかたをしているか

反復・擬似反復と統計モデルの関係

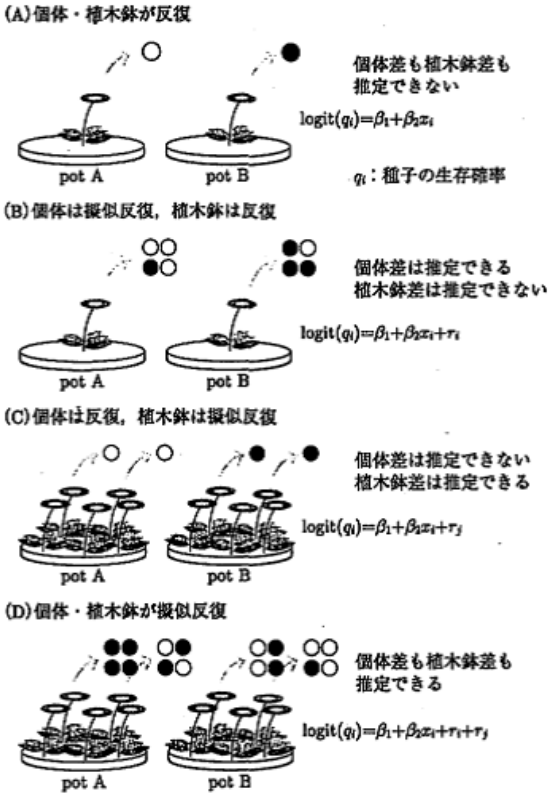

図 7.11　反復・擬似反復と個体差・植木鉢差の推定が可能かどうかの関係．(A)(C)では１個体から１個の種子を取り，その生死を調べているのに対して，(B)(D)では N 個の種子を調べている(白丸が死亡種子，黒丸が生存種子)．また(C)(D)ではひとつの植木鉢で複数の個体を育てている．

- 個体差・場所差をどのように統計モデルに組み込むかの方法は，
  データをどのようにとったに依存する
- 反復：個体差と植木鉢差の区別がつかないデータのとりかた
  - 目的：データからランダム効果の項をなくして，統計モデルを簡単にする
  - GLM で推定可能
- 擬似反復：１個体から何度もデータをとる
  - 個体差が推定可能，ランダム効果の項を追加する
  - GLMM で推定

いろいろな分布のＧＬＭＭ

- ポアソン分布と正規分布を混ぜ合わせる統計モデル
  - 平均より分散がずっと大きな値になる過分散なデータの解析に使う
  - 二項分布ＧＬＭＭと同じようにglmmML()関数が使える
- 応答変数のばらつきが負の二項分布であると仮定するＧＬＭＭ
  - MASS packageのglm.nb()関数などがある
- データのばらつきが正規分布やガンマ分布であるとき
  - lme4 packageのglmer()関数などが使える
- 分布は正規分布で恒等リンク関数とする統計モデル（線形混合モデル）
  - 上記のglmer()関数その他で使える