[学術情報記事]「個体差」の統計モデリング

久保拓弥 * ・粕谷英一 **

* 北海道大学大学院地球環境科学研究院

** 九州大学理学研究院

An introduction to mixed models in ecology

(要旨) 生態学のデータ解析で一般化線形モデル (generalized linear model; GLM) が普及していくにつれ「GLM だけでは説明がむずかしい現象」にも注目が集まりつつある。たとえば「過分散」 (overdispersion) はわれわれがあつかう観測データによくあらわれるパターンであり、これは「あり・なし」データやカウントデータのばらつきが GLM で解析できなくなるほど大きくなることだ。この過分散の原因のひとつは個体差・ブロック差といった「直接は観測されてないがばらつきを増大させる効果」(random effects) である。この解説記事ではこの random effects も組みこんだ一般化線形混合モデル (generalized linear mixed model; GLMM) で架空データを解析しながら個体差・ブロック差を考慮したモデリングについて説明する。

キーワード: overdispersion、R、random effects、一般化線形混合モデル、データ解析

[1]

生態学につきまとう個体差・環境による差

データ解析では意識しようとしまいと何らかの統計モデルを使っている。単純そうに見える解析でもやはり使っていることに変わりはない。この解説記事では、観察された現象を説明するためにつくられた数理モデル、その中でもとくに何か確率分布とそのパラメーターで構成されているものを統計モデルとよび、観測データから統計モデルのパラメーターを決めることを推定とよぶ。統計モデルは現実の近似的な表現だから、統計モデルがあらわしている状態と現実の間のくいちがいはしばしば、まちがったデータ解析の原因となる。統計学の目的のひとつは「統計モデルと現実のくいちがい」を減らすことにあり、そのためにはより現象に合致した新しいモデル・手法の開発が有効であった。昔から知られていたさまざまなくいちがい、たとえば分散がひとしくない状況に対しては不等分散統計モデルの最尤推定法を工夫し、また正規分布ではうまく表現できない現象に適合したロジスティック回帰やポアソン回帰などは一般化線形モデル (generalized linear model, GLM) としてまとめられ、ひろく使われるようになってきた (Dobson 2001, Crawley 2005, Faraway 2006 など)。

生態学であつかう観測データには、測定しきれない環境のちがいや測っていない要因による個体のちがいが「個体差」として観測される。現実のほうに個体差という個体間のばらつきがあるので、統計モデルの側にもばらつきを表現させてやらないと、くいちがいが生じる。このくいちがいはデータ解析の結論をひっくり返してしまうことも少なくない。個体差を扱う、わりと一般的に使えそうな道具として、一般化線形混合モデル (generalized linear mixed model, GLMM) があり、これについて検討してみたい。

この解説記事は 2006 年 3 月の生態学会新潟大会の自由集会「データ解析で出会う統計的問題 — 『個体差』のモデリング」における GLMM に関する話題提供と議論にもとづいている。この自由集会ならびに GLMM 関連の有用な文献・ソフトウェアなどへの最新のリンクは `http://hosho.ees.hokudai.ac.jp/~kubo/ce/` 以下にまとめている。

架空植物データの統計モデリング

この解説記事では架空植物の集団から得られた (当然ながら架空の) 観測データを解析しながら「一般化線形混合モデルとは何か? どう使えばいいのか?」に

[1] 出典: 久保, 粕谷. 2006. 日本生態学会誌 56: 181-190

ついて考えていこう。

図 1 架空植物
この解説記事の例題に登場する植物。8 個の胚珠をもち、そのうち 0-8 個が種子になる (結実する)。結実した胚珠は黒丸、しなかった胚珠は白丸であらわしている。2-6 枚の葉をもち、葉数が多いほど結実数が多くなる確率が高い。ある個体 i においてある各胚珠は独立に確率 p_i で結実する。

説明を簡単にするために、かなり単純化された (そしてあれこれと都合のよろしい) 架空植物を考える (図 1)。この架空植物は葉をもち種子を作ることができる。どの個体も胚珠 (種子になる器官) を 8 個もっている。しかしどの胚珠も種子になるとは限らない。そこで「ある胚珠が種子になる (結実する) 確率」を結実確率 p とする。この結実確率は一個体内の 8 個の胚珠で共通であり、ある個体 i での結実確率を p_i としよう。またこの植物は 2 枚から 6 枚の葉っぱをもち、個体 i ごとに異なる葉数 x_i をもつとする。この架空世界を見物しているわれわれは架空植物の挙動について「真の」情報を全部知っている (ことにしよう)。たとえば、ある個体の結実確率 p_i は葉数 x_i が大きいほど高い、という性質がある (詳細は後述)。

さて、この架空植物を研究している架空生態学者 Q 氏もまた日ごろの観察から「吾輩のみるところ葉っぱの多い個体は結実が良いように思う。理由はおそらく葉 (光合成器官) がたくさんあるので種子を作るのに必要な資源が得やすいためだろう」といった「真の」性質の一端に賢明にも気づき、現象論的な生態学者らしく「原因はともかく、まず葉数 (x_i) が増えるにつれ結実確率 (p_i) がどう増えるかという関係を明らかにしよう」なる問題に興味をもったとしよう。彼はこれをあきらかにするべく自分の調査地内をうろうろして葉数 $x = \{2, 3, 4, 5, 6\}$ 枚それぞれ 20 個体づつの結実確率を調べ (合計 100 個体)、この観測データに統計モデルをあてはめてパラメーター推定し、結実確率 p_i が葉数 x_i によってどう変わるのか知ろうとしている。Q 氏は葉数こそが重要だと確信しており、他は何も観測していない。図 2 は Q 氏が得た全ての観測データを図にまとめたものである。

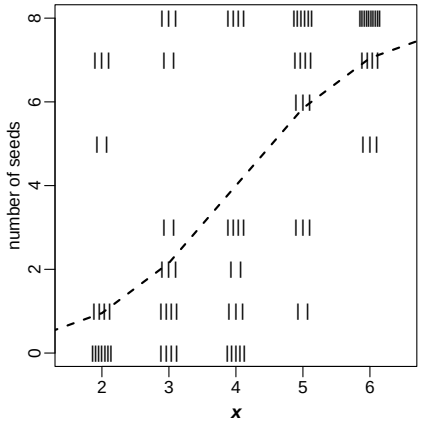

図 2 観測データ
この例題でおもに使う観測データ。横軸は個体の葉数、縦軸はその種子数である (0-8 個)。— が一個体の葉数・種子数をあらわす。破線で示しているのは「真の」(つまり観測者は知らない) 結実曲線である。

図 2 にはこの架空植物の「真の」性質のひとつである「葉数 x の関数である結実確率 p に胚珠数 (8 個) をかけたもの (ある個体での期待結実数)」も曲線 (仮に結実曲線とよぶことにする) で図示されている。後述するようにこれは「結実曲線の集団平均」であり、「個体差」は図示していない。しかしながら、しばらくは「図 2 の曲線のように葉数とともに結実数の期待値が上昇しているらしい」ぐらいのつもりで見てもらえばよい。観測者 Q 氏の研究目的は、図 2 に縦棒で示されている 100 個体ぶんのデータから、この「真の」結実曲線を推定してみせることである、といえる。

結実確率の統計モデリング

架空生態学者 Q 氏は図 2 に示されている観測データに示されている現象「個体 i の結実確率 p_i がその葉数 x_i に依存している」を説明する統計モデリングにとりくんでいる。図 2 の観測データ (縦棒) をながめつつ、彼をこれを二項分布のロジスティックモデルの推定計算 (ロジスティック回帰) を使って解明する

ことにした。

Q 氏のロジスティックモデルは個体 i の中の各胚珠の結実確率 p_i が

$$p_i = p(x_i, a, b) = \frac{1}{1 + \exp(-(a + bx_i))}$$

となる葉数 x_i の関数になると仮定した。各胚珠は独立に結実する・しないが決まるので、ある個体 i の結実数は $8p_i$ の二項分布にしたがう。ここではパラメーター a を「切片」、b を「傾き」とよぶことにしよう。パラメーター a, b とロジスティック曲線の関係を図 3 に示している。

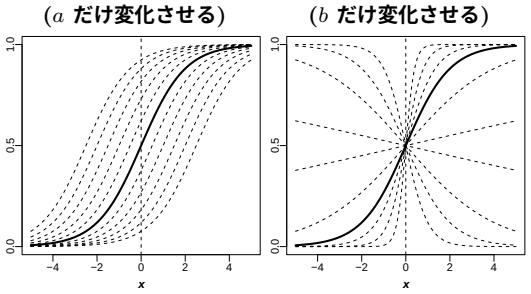

図 3 ロジスティック曲線であらわす確率
ロジスティック曲線の二つのパラメーターへの依存性を示している。

ロジスティック回帰ではパラメーター a と b の値を観測データから最尤推定法によってもとめる。Q 氏の結実確率ロジスティックモデルは

- 推定すべきこの二つのパラメーター a と b が線形に結合している
- 「各個体において 8 個の胚珠のうち y_i 個が結実した」という現象を二項分布で表現する統計モデルを使っている (そして二項分布は指数関数族とよばれる確率分布族の一部である)

といった条件をみたしている。このため、Q 氏の結実確率モデルすなわち二項分布のロジスティックモデルは一般化線形モデル (GLM) に含まれる統計モデルのひとつになっている。(図 4)。

「個体差」と過分散

ところで図 2 の架空観測データを作りだした「真の」結実曲線は (都合よろしいことに) Q 氏が仮定した線形ロジスティック曲線と同じ関数型である。パラメーターである「切片」 a と「傾き」 b は次のような性質をもつ:

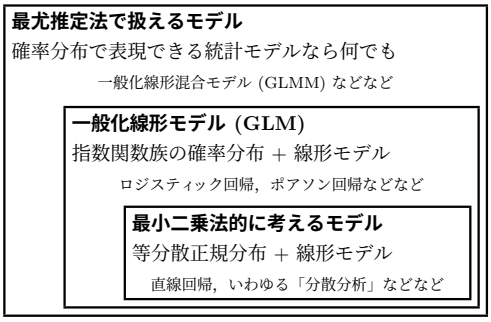

図 4 統計モデルの含有関係
最尤法で検討できるモデルの一部として一般化線形モデル (GLM) と呼ばれる種類のモデルがあり、GLM の中の特殊なものとして等分散正規分布を仮定する最小二乗法であつかうモデルがある。

- 「切片」 a: 「個体差」があり個体によって異なる; 「切片」の集団平均は -4 である
- 「傾き」 b: 「個体差」なくどの個体でも $b = 1$

図 2 に示されているように葉数が同じであっても、個体によって結実数がかなりばらついているように見えるのは a にばらつきがあるためである。このデータを解析する Q 氏は葉数の係数である b を正確に知ることに興味をもっている。このパラメーター b には「個体差」はない。こういった「個体差」ないパラメーターならば、「個体差」に注意しなくても問題なく推定できるのだろうか? この解説記事では「じつは問題がある」ことを順をおってあきらかにしていく。

さて、その前にこの解説記事でもちいている「個体差」という用語が限定された意味である (「個体差」とは「切片」 a が個体によって異なること) と注意したい。葉数が個体ごとに異なることは「個体差」とは呼ばない。この区別はどのようになされているのか? 統計モデルの中でのあつかいをみると、葉数は観測者 Q 氏によって定量化されている「説明変数」であり、いっぽうで a のばらつきは直接には観測されていない (できない) ので「個体差」と呼ぶ、としておこう。[2]

[2] あとでこの「個体差」が random effects の一部である、と説明する。

「切片」 a に「個体差」によるばらつきがあり、「傾き」 b は全個体に共通している。この状態であっても (説明変数である) 葉数 x_i の個体たちが平均的にはどれぐらいの結実確率であるかはわかる。しかしながら各個体での結実確率 p_i が決まらない。このように説明変数以外の何かによって結実確率などに個体間の相違が生じてしまう原因はいろいろ考えられる: 人間が測定していないさまざまな要因、たとえば土壌や微気象といった環境要因、遺伝的要因、あるいは訪花昆虫がどれだけ花粉を運んできたか、個体の成長履歴などなどさまざまなものがありえるだろう。この解説記事の主題は、人間には測定されていない (できない)「個体差」を含む観測データに対して統計モデルを構築する方法の検討である。

ところで、「個体差」は人間には観測されていない (できない) と述べたけれど、「個体差」が大きいときに観測データには特徴的なパターンがあらわれることがある。これは過分散 (overdispersion) と呼ばれている。そこでこの過分散について簡単に説明してみよう。

まず a に関する「個体差」がないと考えてみる。するとどの個体でも $a = -4$ で $b = 1$ となり、ある個体 i の葉数 $x_i = 4$ の場合、結実確率 p_i は 0.5 となり、8 個の種子の結実する・しないは確率 0.5 の二項分布になる。これはオモテ・ウラがでる確率が等しいコイン 8 個を同時に投げたときのオモテの枚数の分布と同じである。このときに個体ごとの結実数の平均は $8 \times 0.5 = 4$ で分散は $8 \times 0.5 \times 0.5 = 2$ となる。

さて、つぎに図 2 のように「個体差」が大きい場合を考えてみる。「葉数 4 のときの結実確率の集団平均は 0.5」がなりたっているときに、「個体差」が極端に大きい場合には図 5 のような観測データが得られるだろう。

標本個体の半数で結実数がゼロ、残り半数で全胚珠が結実となっていて、たしかに結実確率の集団平均 (= 集団内の結実数 / 集団内の全胚珠数) は 0.5 で個体ごとの平均結実数は 4 になっている (実際には結実数 4 の個体はひとつもないのだが)。

もし個体ごとの結実数が二項分布で表現できるならば (つまり「個体差」がなければ) 分散は先述したように 2 であるのに対して、「個体差」が大きい場合には個体ごとの結実数の分散は 2 より大きくなる

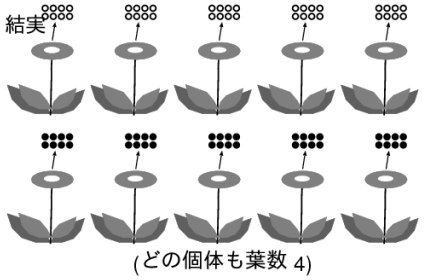

図 5 過分散の例
「結実確率の集団平均は 0.5」という条件のもとで実現している (たいへん極端な) 過分散の例。標本個体数 8 のうち半数では結実数がゼロであり ($p_i = 0$)、残り半分では全胚珠が結実している ($p_i = 1$)。

(図 5 の例だと標本分散はおよそ 18.3)。このように (単純なモデルで) 期待されるより標本分散が大きくなるのが過分散 (overdispersion) である。この定義からわかるように、「過分散がある」と言うためには、現象を説明できる何か単純な統計モデルによって分散が決められている必要がある。生態学まわりでよく使う確率分布でいうと、Q 氏が使っている二項分布にくわえてポアソン分布・幾何分布・指数分布などが「平均が決まると自動的に分散が決まる」つまり 1 パラメーターモデルで、これで現象が説明されるときにだけ過分散の有無を言える。正規分布や負の二項分布といった平均とは独立に分散も自由に指定できる確率分布を使った統計モデルを想定している場合には過分散について議論できない。[3]

「個体差」考えない GLM による推定

架空生態学者 Q 氏のデータ解析の話にもどろう。彼は自分のとったデータ (図 2) が過分散かどうかといったことに関心がなく、「個体差」を考慮していない (「切片」にばらつきのない) 統計モデルである GLM (の一部であるロジスティックモデル) によって結実確率 $p_i = p(x_i, a, b)$ が何の問題もなくパラメーターの推定ができると考えている。この GLM 推定計算のためには統計ソフトウェアが必要なので、Q 氏は (なぜかこの架空世界でも使われている) R というソフトウェアを使うことにした。

この R について簡単に説明しておこう。(もちろん現

[3] この解説記事では紹介しないが、観測データのばらつきが正規分布で表現される統計モデリングでも「個体差」が考慮された統計モデルは以前からよく使われている。これらは線形混合モデルと呼ばれることもある。

実の世界にも存在している) R は統計ソフトウェアというより「データ解析環境」である (R Development Core Team 2006)。R はデータ処理・統計解析・作図などについて高度な機能を提供してくれる free software である (この解説記事のグラフもすべて R で作ったものである)。数年前からさまざまな分野で広く普及しつつあり、日本の生態学研究者 (とくに大学院生たち) の中で利用者が増えている。誰でも無料で入手できる、いろいろな OS で同じように動作する、プログラムソースコードが完全に公開されている (オープンソース的な手法で開発が進められている) ので開発者が企業など特定の集団に属していない、さまざまな機能追加 package が日々更新されている、といった特徴から現代の自然科学研究の道具としてもっとも重要なもののひとつとなりつつある。

「個体差」の問題はいったんおいて、Q 氏が R で GLM のパラメーター推定する手順と結果を見てみよう。Q 氏のデータファイルは `d.csv` という名前の下のようなコンマ区切り値 (CSV) テキストファイルになっていて、[4]

```
  id,x,n.seed, comments
f001,2,     0, # ここから個体 f001
f001,2,     0,
f001,2,     0,
f001,2,     0,
f001,2,     0,
f001,2,     1,
f001,2,     0,
f001,2,     0,
f002,2,     1, # ここから個体 f002
f002,2,     1,
...(以下略)...
```

これを R に読みこんで結果を出力するには三種類の命令 (読みこみ、推定、結果表示) を与えればいいだけである。

```
d <- read.csv("d.csv") # データ読みこみ
fit <- glm( # ロジスティック回帰
    n.seed ~ x,
    family = binomial(link = "logit"),
    data = d
)
print(summary(fit)) # 推定結果の出力
```

ここでは最後の結果出力は省略するが、推定された「傾き」の推定値 $\hat{b}$ は 0.56 であった。推定された結実曲線を観測データと重ねて表示すると図 6 のようになる。Q 氏が R の `glm()` で推定した結実確率の「傾き」は「真の」値に比べてかなり小さいようだ。

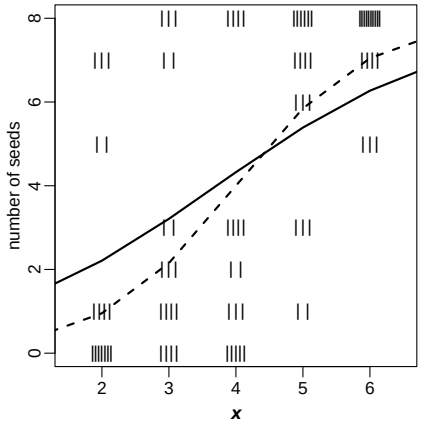

図 6 GLM がうまくいかない例
図 2 の観測データに対して「個体差」を考慮していない R の関数 `glm()` で推定された結実曲線 (実線)。推定された「傾き」は $\hat{b} = 0.56$ だった。破線の「真の」結実曲線は $b = 1$ である。

一般化線形混合モデル (GLMM)

架空生態学者 Q 氏が GLM 使って計算した解析結果は、「切片」 a が「個体差」によってばらついてるときには、「傾き」 b の推定値もその影響をうけて不正確になる、と示しているのかもしれない。

そもそも「切片」 a が個体ごとにことなっている、とはどういった状況なのだろうか。図 2 の観測データを作りだした「真の」結実曲線、つまり個体ごとに「切片」 a が異なる結実曲線群は図 7 のようになる。

図 7 のどの曲線も「傾き」 b は共通である ($b = 1$)。Q 氏の統計モデリングは a のばらつきは無視して `glm()` で「結実曲線の集団平均」(図 2 や図 7 の破線) の推定をねらったものであった。しかしながら、図 7 を見ていると「個体差」を無視して個体間で共通する「傾き」 b を正しく推定するのは難しいような気がする。このように個体ごとに「ずれ」ている結実曲線は統計モデルとしてどう表現すればよいのであろうか。

まずは (GLM の一部である) 二項分布のロジスティックモデルの基本にたちかえって、ある個体 i で観測された結実数が y_i である確率から定式化してみよ

[4] Q 氏の観測データは公開されていて `http://hosho.ees.hokudai.ac.jp/ce/glmm2006/` からダウンロードできる。

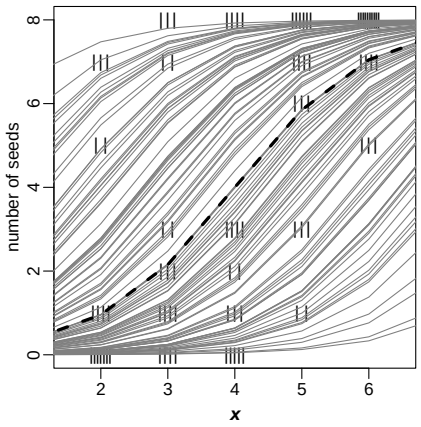

図 7 「切片」 a の「個体差」とは
うすい実線群が 100 個体それぞれのの葉数→種子数関係をあらわす。どの曲線も「傾き」 $b = 1$ である。しかし「切片」 a が個体ごとに異なる (「個体差」)。図 2 の観測データはこれら個体ごとの曲線から二項乱数によって生成された。観測者は「真の」曲線を知らない。

う。これは (各胚珠独立の) 二項分布にしたがうと仮定しているのだから

$$f(y_i|a,b) = \frac{8!}{y_i!(8-y_i)!} p_i^{y_i}(1-p_i)^{8-y_i}$$

と書ける。これを全個体について積をとったものが尤度関数となり、これを最大化するような推定値 $\{\hat{a},\hat{b}\}$ をもとめるのが最尤推定法である。この場合は「個体差」考えないロジスティックモデルとなり、推定結果はすでに図 6 に示されている。

ある個体 i 内の 8 個の種子それぞれの結実確率 p_i に関しては、Q 氏のロジスティックモデルを少しだけ改良して、「個体差」あらわす r_i というパラメーターをくみこむことにする。

$$p_i = p(x_i,a,b,r_i) = \frac{1}{1+\exp(-(a+r_i+bx_i))}$$

新しく組みこんだ r_i の「平均」はゼロとすると、図 7 で示したような「個体差」に影響される結実数の曲線群を生成できる。このように変更すると、新しく「切片」に該当する部分は $a+r_i$ となり、全個体に共通する a と個体ごとに異なる r_i に分割される。

この例にそって fixed または random effects という統計学用語の定義をしてみたい。いまや分割された「切片」の中の a や「傾き」 b は全個体に共通するパラメーターであり、結実確率の集団平均を上下している。これらの全個体共通パラメーターには “fixed effects” (日本語では母数効果あるいは固定効果) がある、と定義される。これに対して「個体 i の切片のずれ」部分を表現しているパラメーター r_i の集団平均はゼロと定義されているので、結実確率の集団平均に無関係であり、ただ個体間の結実確率のばらつきにのみ影響を与えている。このように個体ごとに異なる効果は “random effects” (変量効果またはランダム効果) とよばれている。

Q 氏のロジスティックモデルに r_i を追加して改良したモデルは fixed と random effects の両方を含まれているので混合モデル (mixed model) と呼ばれ、「個体差」を明示的にあつかえる統計モデルである。なかでもロジスティックモデルなど GLM に属するモデルを混合化した場合には、一般化線形混合モデル (GLMM) と呼ばれる。

さて、この「個体差」あるいは random effects をあらわす r_i は観測データをどのように対応づければよいのだろうか? パラメーター a や b のように r_i の値も最尤推定してやることも不可能ではない。しかしながら個体数 100 に対応させるべく r_i も $\{\hat{r}_1,\hat{r}_2,\cdots,\hat{r}_{100}\}$ と 100 個の推定値を決めてやることになり、モデルの自由度はいっきに 100 も少なくなる。このように莫大な個数のパラメーターを導入して自由度を減らす、つまり「次元の呪い」(サルツブルグ 2006) に呪われた統計モデルを構築するのはさけたいところだ。

そこで混合モデルでは自由度を減らさずに random effects r_i をあつかう。その方法は r_i が平均ゼロの確率分布の何か、たとえば正規分布で表現できると仮定する。すると「個体差」 r_i の確率分布 $g(r_i|s)$ は

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

となり (図 8)、新しく導入したパラメーター s は「集団内の r_i のばらつき」をあらわす標準偏差である。[5]

このように random effects あらわす r_i の確率分布を導入することによって、個体 i で観測された結実種子数が y_i であるときの尤度 L_i は「すべての可能な r_i」に関する期待値が計算でき、

$$L_i(a,b,s|x_i,y_i) = \int_{-\infty}^{\infty} f(y_i|a,b,r_i)\ g(r_i|s)dr_i$$

[5] 毎度のことながら、今回もまた都合のよろしいことに、ここであつかっている架空植物の「個体差」も平均ゼロで $s = 2$ の正規分布からのランダムサンプルとして生成されたものである。

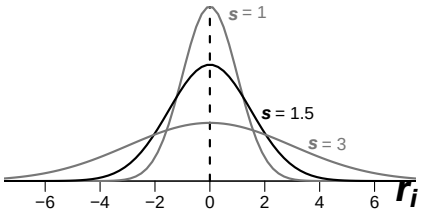

図 8 「個体差」の確率分布
標準偏差 s の正規分布で「個体差」 r_i の確率分布を表現している。

このように個体ごとの尤度は「実際の r_i」を知らなくても計算できるようになる。集団全体の尤度は L_i を全個体ぶんの積として表現されるので、

$$L(a,b,s|\{x_i\},\{y_i\}) = \prod_{i\in\{\text{ 全個体 }\}} L_i(a,b,s|x_i,y_i)$$

となり、これを最大化する最尤推定値、「切片」 $\hat{a}$ と「傾き」 $\hat{b}$ そして「集団内の r_i のばらつき」 $\hat{s}$ を同時に求めればよい。これが (GLMM の一部である) 混合ロジスティックモデルの最尤推定である。

R でやってみる GLMM 推定

さて、R でこの混合ロジスティックモデル (GLMM) を推定してみよう。ここでは `glmmML()` という関数を使った最尤推定を試みる。この推定関数 `glmmML()` は標準ではない R package なので CRAN サイト[6] からダウンロードしてインストールし、計算前に `library()` 関数を使って読みこむ必要がある。この点に注意すればあとの手順は `glm()` による推定とほとんど同じだ。ただし r_i が個体ごとに与えられる量であることを `cluster` オプションで指定する必要がある。

```
library(glmmML) # glmmML package を使用
d <- read.csv("d.csv") # データ読みこみ
fit <- glmmML( # 混合ロジスティック回帰
    n.seed ~ x,
    cluster = d$id, # 個体 ID ごとに
    family = binomial,
    data = d
)
print(summary(fit)) # 推定結果の出力
```

[6] R の機能拡張用の追加 pacakge は The Comprehensive R Archive Network (CRAN) `http://cran.r-project.org/` で配布されていて誰でも自由に入手できる。R 本体もここからダウンロードできる。

推定された結果を図示すると、「個体差」がきわめて大きい観測データであるにもかかわらず、「真の」 b に近い推定値が得られたことがわかる (図 9)。

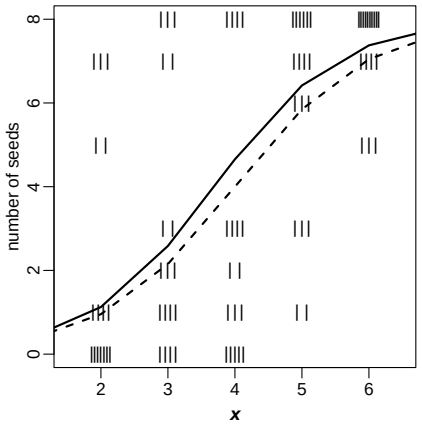

図 9 GLMM 推定結果
図 2 の Q 氏のデータに対して、R の `glmmML()` 関数を使って結実曲線 (実線) を推定した。「真の」曲線 (破線) にかなり近づいている。「傾き」の推定値は $\hat{b} = 1.07$ となった。

「個体差」さえなければ ……

「個体差」を考慮していない GLM では推定値が過小推定になり (図 6 の Q 氏の推定結果)、「個体差」をくみこんだ GLMM では「真の」値に近い値が得られたように見える (図 9)。しかしながら、そのことを指摘してみても、「個体差」とか重視していない Q 氏はひょっとしたら「標本数が少なかったからだ」とか「たまたま観測データが悪かった」などと反論してくるかもしれない。

パラメーター b を正確に推定するにあたって、100 個体 $\times$ 8 胚珠という標本数は少なすぎたのだろうか。たとえばもし「切片」 a に「個体差」がなくどの個体も集団平均値である $a = 4$ となっていたと仮定しよう (先ほどと同じく「傾き」 $b = 1$ とする)。そのような架空植物集団から Q 氏と同じく 100 個体 $\times$ 8 胚珠のデータが得られたなら `glm()` のロジスティック回帰による推定で「真の」 b に近い推定値 $\hat{b}$ が得られる (図 10)。

つまり「個体差」さえなければ、この標本数であっても GLM のパラメーターを正確に推定できそうだ。Q 氏の推定結果 (図 6) がうまくいってないのは観測

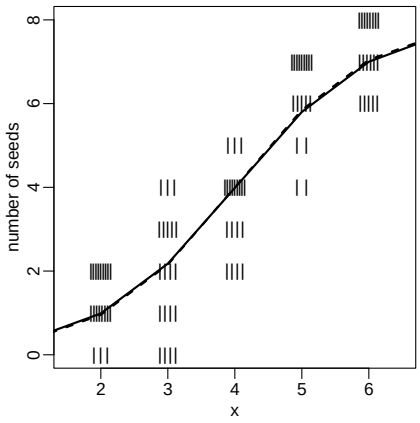

図 10 GLM がうまくいく例
もし「切片」 a に個体差がない植物集団で観測データを得ていたならば、このようにロジスティック回帰で「真の」値に近い推定値が得られる。「傾き」推定値 $\hat{b} = 0.97$(「真の」 $b = 1$)。

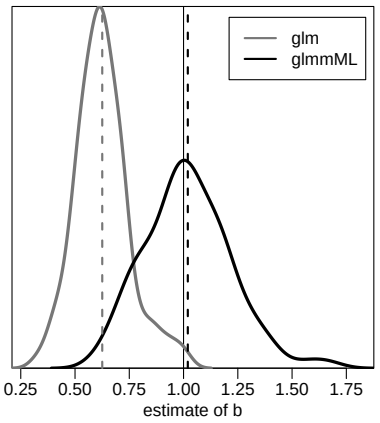

図 11 「傾き」 b の推定値の分布の比較
「個体差」ある架空植物集団から独立に 100 回生成して、それぞれについてロジスティック回帰したときに得られた「傾き」の推定値 $\hat{b}$ の分布。灰色の線が GLM の推定結果であり、黒の線が GLMM によるものである。縦の破線は得られた推定値の分布の平均である。

データ (図 2) が「たまたま悪かった」せいだろうか。この問題についてさらに詳しく調べてみよう。

GLMM は推定を改善しているのか?

図 2 で示されている観測データに対して、GLM では「傾き」 b に関してあまりよい推定結果が得られず (図 6)、GLMM を使うと改善されたという例 (図 9) を示した。このように「個体差」ある観測データに対して、GLM と GLMM はそれぞれどれぐらい「良い」統計モデルなのだろうか。ここでは「切片」が個体ごとにばらついている集団から得られた標本 100 個体から得られる「傾き」 b の推定値の分布について調べてみよう。

この分布を作るためには「サンプリングと GLM と GLMM による推定」という手順をくりかえすシミュレイション、つまり

(1) 同じ個体群から 100 個体ランダムサンプリングしてきて結実ぐあいを調べる

(2) `glm()` と `glmmML()` それぞれで b の推定値を得る

といった観測・推定を 100 回 繰りかえした。

このようにして「個体差」ある集団から 100 個体の標本から得られる「傾き」 b の推定値の分布を図 11 に示している。これをみると、「個体差」は考慮してない `glm()` の推定結果では、何回サンプリングと推定をやりなおしても「傾き」 b の推定値は過小推定になる、とわかる。つまり GLM の推定量は偏り (bias) がある。Q 氏が GLM を使って得た b の過小推定値は「たまたまヘンな観測データだった」からではなく、「個体差」を無視した統計モデリングに原因があるといえる。

これに対して、「個体差」を考慮している GLMM の最尤推定つまり `glmmML()` で得られた推定値 $\hat{b}$ の分布は偏りがほとんどない。ただし $\hat{b}$ のばらつきは小さくない。図では示していないが、「個体差」である「切片」 a の個体間のちがいが大きくなると、この「傾き」 b (こちらは個体間で共通の値) の推定値のばらつきも増大する。すなわち、$\hat{b}$ のばらつきの大きさがあらわしているのは、「個体差」が大きい観測データからパラメーターの値を正確に推定するのは難しい、ということなのである。

過分散のみつけかた

「個体差」は過分散の原因のひとつであり、過分散は統計モデルの中のパラメーターの推定値を偏らせる原因となっていた。それならば、ある観測データが「過分散である」かどうかを判定するにはどうしたらよいのだろうか。最後にもう一度この問題を検討したい。というのも、そもそも Q 氏は図 2 の自分のデータが過分散であるとは知らないので、モデルをどう改良すればよいのかわからないからだ。

Crawley (2005) をはじめとしていくつかの教科書

では GLM 推定時の residual deviance と自由度を比較すればある観測データが過分散かどうか判断できる (residual deviance のほうが大なら過分散)、といった方法が紹介されている。Q 氏の観測データ (図 2) で比較すると、residual deviance が自由度の 5 倍も大きいのでこの方法は有効なのかもしれない。しかしながら、この判定方法は標本数もかなり多いときにのみ有効になるものである。

ここで取りあつかっているような単純な構造のデータの場合には、むしろまた二項分布モデルの基本にたちかえり、もっと素朴に過分散の有無を判定する方法を考えてみよう。

すでに説明したように過分散とは「平均が決まれば自動的にばらつきも決まる」単純な統計モデルからの逸脱を意味する。この例題での単純モデルは二項分布をつかった種子結実の統計モデルであり、過分散をみつけるためには、まず「個体差」がないと仮定した二項分布モデルによる推定結果が必要となる。

つまり Q 氏が図 6 で示している `glm(..., family = binomial)` による (どちらかといえば不適切な) 推定結果は「この観測データは過分散かどうか」を判定するときにたいへん役にたつのである! Q 氏は葉数と結実確率の関係を `glm()` を推定している、つまり 8 個の胚珠のうち y 個が結実する確率が二項分布によって具体的に表現されていることになる。さて、ここでこの Q 氏の二項分布モデルが正しいと仮定したときに「標本のうち 99% はどの範囲にちらばるか」を図上に示すことができる (図 12)。もし Q 氏の推定したモデル (二項分布) から 100 個体が選ばれたのであれば、ほとんどの点はこの 99% 区間内に存在するはずである。しかしながら図 12 をみればわかるように、99% 区間からはみだしている個体の数はひとつやふたつではない。このことから、図 2 の観測データが過分散であり、「個体差」を考慮したデータ解析が必要であることがわかる。[7]

[7] ここで示した方法は残差を標準偏差で規準化した Pearson residual の解析に似ている。Pearson residual もまた overdispersion を探すときに使う手法である。またこの例の場合、説明変数 x も離散値なので、葉数ごとに結実確率の部分集団平均をとって 99% 区間を計算してみるのも overdispersion 発見に有用である。

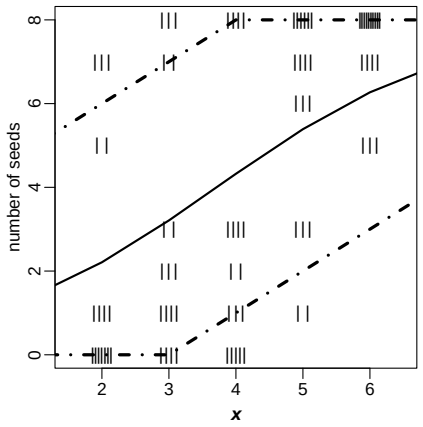

図 12 二項分布 99% 範囲
図 6 で示されている GLM 推定結果が正しいと仮定した場合に期待される標本のばらつく範囲。横軸の葉数 x が与えられたときに、上下の一点破線に囲まれた領域の外で標本が得られる確率は 1% である。これは R の `qbinom()` 関数を利用すると簡単に計算できる。

これまでとこれから

過去 20 年間ぐらいの日本人研究者による生態学での統計モデルの使われかたを乱暴に要約してみると、まず「世の中なんでも等分散正規分布」と仮定する最小二乗法あるいはいわゆる「分散分析」ばかりが使われていた時代があり、次に順位和検定などノンパラメトリック検定が万能の統計学的手法と一部に錯覚されていた一時期 (この手法の適用にもさまざまな注意が必要なことは Kasuya 2001 など参照) を経て、ここ 4-5 年のあいだに「あり・なし」データやカウントデータの解析にはロジスティック回帰・ポアソン回帰などを統合した一般化線形モデル (GLM) が使われるようになり、最尤推定法の考えかたも普及してきた。その GLM に「個体差」こと random effects を「混合」した一般化線形混合モデル (GLMM) も少しずつ使われるようになってきたので、この解説記事ではその基本となる考えかたについて説明した。

混合モデルの考えかたを発展させていくと (この解説記事では紹介できなかった) いろいろと興味ぶかい現象の統計モデリングに応用できるだろう。たとえば

- 一個体から何度もサンプリングをくりかえす縦断的データ (longitudinal data) の解析、あるいは擬似反復 (pseudo replication) 問題への対策

- 実験ブロック差がある中での個体差といった階層性のある (あるいはネストされた) random effects のモデリング
- 距離が近いほど二個体は挙動がより似ている、といった空間相関の問題

など、生態学データ解析で「よくある状況」に対処できる統計モデルを考えるときに、ここでとりあつかった最も簡単な GLMM はその出発点となる。

また混合モデルは「経験ベイズ法」とよばれることもあり (石黒ほか 2004 など)、ベイズ統計学と密接な関係がある。ベイズ統計学 (伊庭 2003 など) は 1990 年以降の計算機集約的な MCMC 計算の発展に伴い、さまざまな分野のデータ解析の統計モデリングにおいてこれまでの難問を解決できる強力な手法として応用範囲が拡大しつつある。生態学周辺においても事情は同様で、(さきほど「よくある状況」として列挙したような) 複雑な構造をもつデータをあつかわねばならぬ生態学研究者はみなベイズ推定する Bayesian になっていくだろう、と先鋭的な生態学データ解析者の一人である Clark は指摘している (Clark 2005)。こういったさまざま新しいデータ解析手法の理解するときにも、この解説記事で説明しているような GLMM の基礎の勉強、そしてその応用方法の試行錯誤はたいへん役にたつことになるだろう。

参考文献

Clark JS (2005) Why environmental scientists are becoming Bayesians. Ecology Letters 8:2-14

Crawley MJ (2005) Statistics: an introduction using R. John Wiley & Sons, West Sussex

Dobson AJ (2001) An introduction to generalized linear models (2nd ed.). CRC Press, Boca Raton

Faraway JJ (2006) Extending the linear lodel with R: generalized linear, mixed effects and nonparametric regression models. CRC Press, Boca Raton

伊庭幸人 (2003) ベイズ統計と統計物理. 岩波書店, 東京

石黒真木夫, 松本隆, 乾敏郎, 田邉國士 (2004) 階層ベイズモデルとその周辺. 岩波書店, 東京

Kasuya E (2001) Mann-Whitney U test when variances are unequal. Animal Behaviour 61: 1247-1249

R Development Core Team (2006) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/

サルツブルグ D (2006) 統計学を拓いた異才たち (竹内惠行・熊谷悦生訳). 日本経済新聞社, 東京