

Generalized Nonlinear Models の実践的な活用の要領

大西 修平^{*1}・松村 哲^{*2}・赤川 泉^{*1}

Tips on software operation of gnm used for constructing Generalized Nonlinear Models

Shuhei OHNISHI^{*1}, Satoru MATSUMURA^{*2}, Izumi AKAGAWA^{*1}

Abstract

Generalized Linear Models (GLM) are currently widely applied for biological data analysis, although limited to linear models. In particular, increasing emphasis on the programming language R has contributed greatly to the spread of GLM usage. In contrast, nonlinear models can be constructed with 'gnm', a registered application within R-Packages. This study describes in detail the use of gnm, based on a case study on habitat preference of freshwater fish.

Keywords: GLM, R-Packages, Generalized Nonlinear Models

はじめに

資源生物研究に利用される回帰分析の種類には、初歩的な直線回帰や多項式回帰に始まり、複数の独立変数 (independent variables) による (線形) 重回帰モデルなどが知られる。また独立変数の個数に関わらず、モデルの関数に制約を設けない一般的なツールでは、非線形回帰モデルがある。“回帰モデル”の範疇には、広範囲の統計処理系が含まれる。

回帰モデルのもう一方の要素；従属変数 (dependent variable) は誤差を伴う。計算の簡便さから、正規分布 (Gaussian distribution) を仮定することが多いが、この条件は保障されるわけではない。データ (従属変数) 発生の仕組みと観測の特性に合わせて、誤差が正規分布に従わない回帰モデルについても、準備が必要である。非正規型 (非ガウス型: non-Gaussian) のモデルの場合、モデル推定は最小二乗法に代えて、最尤推定法 (Maximum likelihood estimation: MLE) に依らなければならない。総じて MLE の演算は複雑なので、統計のエンドユーザにとっての MLE は「ハードルの高いツール」になりがちである。

回帰モデルを非ガウス型に拡張するアイデアは, Nelder

and Wedderburn (1972) において、一般化線形モデル (Generalized Linear Model: GLM) として示された。GLM により、従属変数の確率構造を二項分布やポアソン分布などに拡張できる。GLM は MLE を解法とするので、この意味においては GLM も依然「ハードル」が高い。さて、計算機技術の向上と、ソフトウェアの充実が後押しする形で、GLM は多くの統計アプリケーションに実装されるようになった。MLE のアルゴリズムが処理系に組み込まれ、数値計算が自動化されることで、MLE を意識せず GLM を利用できる環境が整備された。その結果、GLM は「決してハードルの高くないツール」に定着しつつある。

GLM により非ガウス型の回帰分析が幅広く理解されるようになった。これには、統計分析アプリケーション；R 言語 (R Core Team, 2014) が GLM を標準組み込み関数に採用していることや、R 言語のシェア拡大が大きく貢献した。R 言語での GLM については、出版物やウェブサイトを通じて、日常的に話題が共有される。このように飛躍的に普及した GLM であるが、回帰式の定義は線形モデルであり、非線形モデルに向けた一般化には至っていない。

以上のような計算環境のもと、R 言語のプラグイン

2014年12月10日受付 2015年1月13日受理

*1 東海大学海洋学部, 海洋生物学科 (Department of Marine Biology, The School of Marine Science and Technology Tokai University, Shizuoka Japan 424-8610)

*2 東京都葛西臨海水族園 (Tokyo Sea Life Park, Tokyo Japan 134-8587)

(R-packages) に、非線形・非ガウス型の回帰分析を実行できるアプリケーション gnm (R では Generalized Nonlinear Models と表記) がエントリされている (Turner and Firth, 2007a). gnm は事実上 GLM を一般化したツールであり (Turner and Firth, 2007b), これにより非線形・非ガウス型の回帰分析が、MLE を意識することなく行える。gnm は未だ十分に知られておらず、解説書も出版されていない。また具体的な用法を分かりやすく読み取れる研究事例も知られていない。そこで本報告は、資源生物系の身近な例題に沿って、gnm の用法のポイントを解説する。

GLM と gnm

gnm を GLM と対比させて簡単に説明する。ただし GLM については、既に詳しい解説が多数あるので、ここでは多くふれない。なお gnm は R-packages のひとつで、現時点では R 言語の開発上のコードネームである。そのためオリジナルの表記に従い、小文字のまま“gnm”とする。

まず GLM は、 N 組の標本について、以下のような一般形で定義される。

$$g(E[y_i]) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} \quad (i=1, \cdots, N) \quad (1)$$

ここで x , y はそれぞれ独立変数と従属変数、 p は変数の個数である。 β は線形モデルの係数を、 β_0 は定数項を意味する。(1)の $E[\cdot]$ は期待値である。 $g(\cdot)$ は変数変換を行う関数で、リンク関数 (link function) と呼ばれる。

gnm は、(1)の右辺を任意の関数 f で、

$$g(E[y]) = f(\mathbf{x}|\boldsymbol{\beta}) \quad (2)$$

と一般化したモデルである。太字はベクトルを意味する。

(1)から(2)への拡張によって、gnm は非線形・非ガウス型の回帰モデルとなる。

gnm が求められる場面

具体的なデータに沿って gnm の使い方を説明する。Fig.1は、淡水魚の河川環境に対する選好性を調べるために、環境特性と生息数の関係を示したものである。横軸には河川の環境特性として、 x : 平均底層流速 (m/sec) を、縦軸には条件 x での y : 生息個体数を表示している。ただし x では最小値と最大値に約70倍の開きがあるため、常用対数で表示している。本データは、2013年7月から9月、静岡県富士宮市・富士川水系・芝川支流大倉川での、アマゴ *Oncorhynchus masou ishikawae* の生息調査によるものである。

(i) 非ガウス型の必要性

Fig.1では、まず縦軸 y (生息個体数) の値が 0, 1,

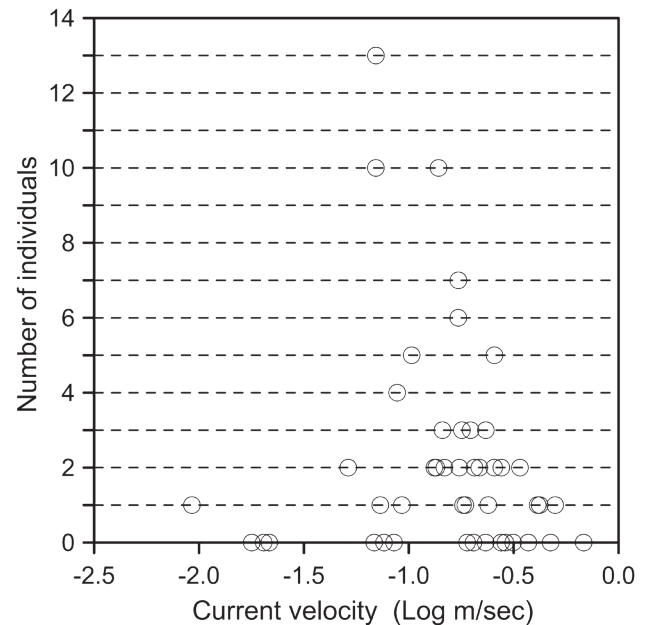


Fig.1 Relationship between current velocity as an environmental index and the number of observed individuals of red-spotted masu salmon, *Oncorhynchus masou ishikawae*, in Okura-gawa stream, Fuji River system, Japan.

2, ... の離散値であることに注意が必要である。つまり y の誤差に、離散型確率分布を仮定しなければならないという意味である。このことは実用的に言えば、Fig.1の縦軸上では正規分布は無効であり、よって最小二乗法による回帰分析は、この場合使えない“統計処理の誤用に注意”という解釈になる。

(ii) 非線形モデルの必要性

(i) で挙げた課題 (非ガウス型の必要性) をクリアするだけなら GLM で十分である。しかしながら、生物が対象の場合、GLM が不適合のケースは特に多いと思われる。Fig.1のように選好性を調べる場合、「最適性」も欠かせない。 x と y 間の単調増加・減少とともに、 y にとって最適な x がある (かも知れない) という仮説の導入である。

Fig.1について y の極大値を表すために、ここでは Lorentz 型関数;

$$y = \frac{\beta_1}{1 + \left(\frac{x - \beta_2}{\beta_3} \right)^2} \quad (3)$$

を使う。(3)は未知パラメータ3個を含む非線形モデルになっている。

R での gnm スクリプトコーディング

(3)を Fig.1のデータにフィットさせるための gnm スクリプト例を Fig.2に示し、ポイントを説明する。ここではコ

```

01: library(gnm)
02: F<-function(x) {
03:   list(predictors=list(a=1,b=1,c=1),
04:   variables=list(substitute(x)),
05:   term=function(predictors,variables){
06:     pred<-paste(predictors[1],"/(1.0+(",
07:       variables,"-",predictors[2],")/",
08:       predictors[3],")^2)",sep="")})}
09: class(F)<-"nonlin"
10: res<-gnm(y~0+F(x),family=poisson,
11:   start=c(a=5.0,b=0.0,c=0.5),
12:   verbose=FALSE)
13: exitInfo(res)

```

Fig.2 The R script for a Poisson regression analysis by gnm package.

ンピュータにR言語がインストール済みとする。

準備したデータは、先に述べた x と y (流速と個体数) である。gnm スクリプト中の x と y (02, 04, 10各行) がそのまま対応する。01行は gnm を呼び出す命令である。gnm は R に標準装備されていないので、R-packages からダウンロードし、R 立ち上げの際に1度だけ呼び出せばよい。02-08行は、回帰させる関数の記述と、使用データやパラメータの宣言である。03行の a , b , c はパラメータの任意の命名であり、(3)の β_1 , β_2 , β_3 に対応する。またこれらを predictors として配列にまとめている。 a , b , c すべてに1を代入しており、それぞれのパラメータが唯一の推定値をもつように制約を与えている (Turner and Firth, 2012)。通常、パラメータ推定値はひとつなので、03行の値は1のままでよい。05-08行は(3)を具体的にコーディングしている。gnm では R のコマンド paste() を使い、文字列の連結で関数を記述する。09行は、扱う問題が非線形の関数を含むことを、形式的に宣言しているにすぎない。09行もこのままでよい。

10-12行は、gnm の計算オプション設定と実行命令である。この部分は問題ごとに注意深く記述する。 $y \sim 0 + F(x)$ は、02-08行で定義したデータ・パラメータ・関数を一括して計算に取り込む命令である。0 は定数項を使わない指示である。Fig.1の縦軸 y が離散値であることは既に説明した。そのため、離散型確率分布を選択しなければならない。この場合、最も適切な確率分布はポアソン分布である。10行の family=poisson は、 y にポアソン分布を仮定して回帰分析を行う宣言である。本報告の問題は「ポアソン回帰分析」となる。

(1)と(2)には、リンク関数 $g(\cdot)$ という演算が含まれていた。GLM のポアソン回帰分析の場合、リンク関数には対数関数 (log) を使う。この取り決めは、ポアソン分布の期待値が正値 (> 0) の条件を満たすための便宜的なものである (Dobson, 2008)。 $g(\cdot)$ の逆関数 $g(\cdot)^{-1}$ が指数関数 (exp) となるので、(1)と(2)の期待値は正値に維持で

```

01: > res$converged
02: [1] TRUE
03: >
04: >
05: > res$iter
06: [1] 45
07: >
08: >
09: > res$coefficients[1:3]
10:           a           b           c
11:  1.4608239 -1.0078131  0.3214983
12: >
13: >
14: > res$aic[1]
15: [1] 210.7771

```

Fig.3 R operations for obtaining stored statistics.

きる。ポアソン分布以外の場合、リンク関数の選択肢は多い。リンク関数の選び方については、別途解説を利用されたい。gnm も GLM と同様、リンク関数に対数関数を使うが、family=poisson さえ宣言すれば、対数関数の利用はデフォルト設定になる。gnm のポアソン回帰では、リンク関数の宣言は省略してよい。

11行は、MLE (最尤推定法) の数値的最適化でのパラメータ初期値を与える。12行 verbose は、数値的最適化の繰り返し計算の経過を出力する / しないの選択である。ここでは verbose=FALSE (経過出力なし) としている。gnm の処理の結果は、全て res に代入している (10行)。計算終了後 res を呼び出せば、結果を取り出すことができる。13行の exitInfo() は、数値解の収束を調べるコマンドで、計算の正常終了を判別する。exitInfo() は、スクリプトの末尾に必ず付けておくことを薦める。

特に詳しく説明しなかった箇所は、スクリプトの形式上の記述なので、そのままでよい。Fig.2の構造を参考にし、スクリプトが完成したら、R のコマンドラインに打ち込めばよい。gnm が直ちに演算を開始し、結果は res に保存される。

処理結果を取り出す方法

res に保存された処理結果は、随時出力が可能である。一連の操作 (Fig.3) のように、コマンドを入力するだけでよい。ここでは主要な操作だけを示した。なお入力はプロンプト (>) に続く文字列である。

01-02行は数値解の収束を確認している。02行の TRUE は正常終了を意味し、後に示すパラメータ推定値 (解) が、信頼できるという解釈を与える。05-06行は繰り返し計算の回数である。09-11行はパラメータ推定値の表示である。パラメータ推定値は、期待値 $E[y_i]$ の計算に必要である。たとえば a だけ取り出すには、09行を res\$coefficients[1] と変更すればよい。14-15行は

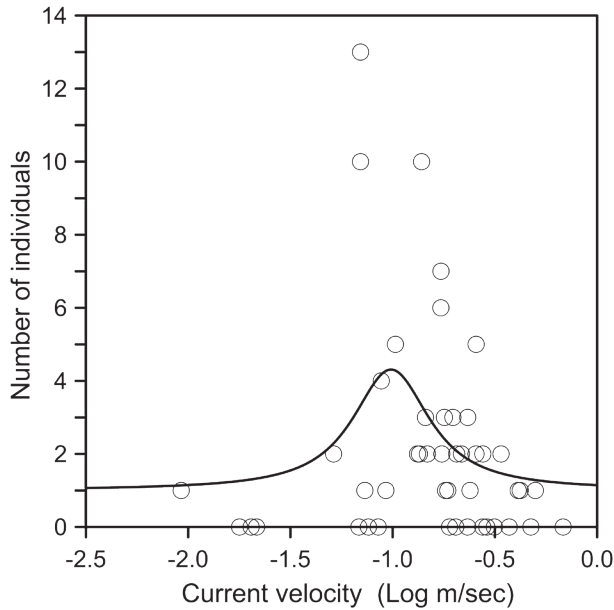


Fig.4 The solution of curve fitting for the Lorentzian function.

情報量規準 AIC を出力している。

09行の操作で取り出したパラメータ推定値を再利用し、(3)の y の期待値 $E[y_i]$ を計算し、Fig.1に重ねて描いた (Fig.4)。Fig.4はデータに対して最も良い当てはまりを示した曲線である。しかしながら、点（データ）の間を縫う曲線（期待値）になっておらず、Fig.4は当てはまりの良さを強く印象づけるものではない。この理由は、回帰モデルの y に、ポアソン分布を仮定しているからである。最小二乗法の場合、 y の周囲の誤差は正規分布で定義する。正規分布は左右対称形の確率分布であるので、期待値がデータの間を縫うような、いわゆるカーブフィットが得られる。ポアソン分布の確率密度関数は、左右非対称の裾を引いた山型をとる。Fig.4は、縦軸方向でゼロ寄りの領域に曲線が内挿されている。データと期待値（点と曲線）の位置関係 (Fig.4) は、上記のポアソン分布の持つ形状を反

映したものとなっている。非ガウス型回帰モデルを利用する際は、普段見慣れた最小二乗法の結果と、当てはまりの視覚的パターンが異なるという特性についても、理解が必要である。

まとめ

本報告では、gnm の（非線形）ポアソン回帰分析への利用に限定して説明したが、これにとどまらず、二項分布を使う場合はロジット回帰を非線形モデルに拡張できる。フィールドデータでは実験室内のデータと違い、誤差を押さえ込むための効果的な手段はない。野外で収集したデータの“振れと暴れ”に対して、分析方法（モデル）の側で工夫するという考えが現実的である。その際には、統計モデルは自由度の高いものが望ましい。フィールドでは、計画的な実験や追試は難しいという事情を考慮すれば、高い自由度を備えた非線形・非ガウス型の回帰モデルとして、gnm は頼もしいツールである。研究と教育、両方の活動において、gnm の活躍の機会は非常に多いと考えられる。

参考文献

- Dobson, A.J. (2008) : 一般化線形モデル入門 原著第2版 (田中豊・森川敏彦・山中竹春・富田誠訳). 共立出版. 東京. 264pp.
- Nelder, J.A. and Wedderburn, R.W.M. (1972): Generalized Linear Models. *J.R.Statist.Soc.A*, 135(3), 370-384.
- R Core Team (2014): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- Turner, H. and Firth, D. (2007a): Generalized Nonlinear Models in R. *Statist.Comput.Graph.*, 18(1), 11-16.
- Turner, H. and Firth, D. (2007b): gnm: A Package for Generalized Nonlinear Models. *R News*, 7(2), 8-12.
- Turner, H. and Firth, D. (2012): Generalized nonlinear models in R: An overview of the gnm package. (R package version 1.0-7). <http://CRAN.R-project.org/package=gnm>.

要 旨

一般化線形モデル (Generalized Linear Model: GLM) は生物データの分析において、幅広く活用されている。特に R 言語のシェア拡大が GLM の普及に大きく貢献してきた。GLM は汎用的な統計処理法であるが、分析は線形モデルに限定される。これに対して、R-packages に登録されたアプリケーション gnm により、非線形モデルの構築が可能となる。本報告では、生物系の例題に基づいて、gnm の初歩的な使い方を解説する。

キーワード : 一般化線形モデル, R-Packages, Generalized Nonlinear Models