線形モデルから一般化線形モデル（GLM） へ

誌名	雑草研究
ISSN	0372798X
著者名	大東,健太郎
発行元	日本雑草防除研究会
巻/号	55巻4号
掲載ページ	p. 268-274
発行年月	2010年12月

農林水産省 農林水産技術会議事務局筑波産学連携支援センター
Tsukuba Business-Academia Cooperation Support Center, Agriculture, Forestry and Fisheries Research Council Secretariat

雑草研究　Vol. 55(4) 268 ～ 274 (2010)
J. Weed Sci. Tech.
総　説

特集　統計解析（再？）入門
線形モデルから一般化線形モデル（GLM）へ

大東健太郎*

キーワード：統計モデル，一般線形モデル，一般化線形モデル，一般化線形混合モデル，逸脱度
Keywords：statistical modeling, general linear model, generalized linear model, generalized linear mixed model, deviance

はじめに

R. Fisher 以来，統計学はたゆまぬ発展を遂げて来たが，実際の研究上では，比較的古いアプローチである分散分析と回帰分析に代表される，誤差の正規性を要求する解析手法が長きに渡って利用されている。かつては定性的な要因（カテゴリカルデータ）を扱う場合には分散分析，定量的な要因（連続データ）を扱う場合には回帰分析を行うことが主流であった。しかし，これら二つの解析手法は，観測データを「要因効果と誤差の和」として考えている点で，同一の手法であると言える。このように，分散分析と回帰分析をあわせて扱う場合に「線形モデル」あるいは「一般線形モデル」と呼ぶ（general linear models）。以下本稿では線形モデルとする。

今 p 個のパラメータで，目的変数 Y を説明する回帰モデルを予測子 X の和の式で表すと以下の（1）式のように記述される。

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_{p-1}X_{p-1}+E \qquad (1)$$

線形モデルとは，（1）式のように全て足し算の関係（相加的）で表されるモデルで，この時の X_i を線形予測子と言う。例えば，農薬散布量や積算温度の 2 つのパラメータで，ある時点での雑草重がどの程度になるかを説明する回帰モデルを考えたとすると Y が雑草重で，農薬散布量や積算温度が予測子 X にあたる。ここで，Y は n 個のデータ，$\{y_1, y_2, \cdots, y_n\}$ からなる縦のベクトル，β は回帰係数 $\{\beta_0, \beta_1, \cdots\beta_{p-1}\}$ のベクトル，E を正規分布する誤差，$\{\varepsilon_1, \varepsilon_2,$
なる縦のベクトルとし，

$$X=\begin{bmatrix}1, X_{11}, \cdots X_{1p-1}\\ 1, X_{21}, \cdots X_{2p-1}\\ \vdots\\ 1, X_{n1}, \cdots X_{np-1}\end{bmatrix}$$ と置くことで（2）式のよ

$$\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}=\begin{bmatrix}1, X_{11}, \cdots X_{1p-1}\\ 1, X_{21}, \cdots X_{2p-1}\\ \vdots\\ 1, X_{n1}, \cdots X_{np-1}\end{bmatrix}\begin{bmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_{p-1}\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{bmatrix}\Leftrightarrow Y$$

これを変形するといわゆる残差，

$$E=Y-XB$$

が得られる。

線形モデルでは，ε の母集団が等分散正るとの仮定から，この二乗，即ち残差平方和するβを推定している。詳しくは記述しない分析の場合でも同様の行列式として（2），表すことができ，誤差構造が等分散正規分定している。本稿の主題である一般化線（generalized linear model）は，線形モデル拡張であり，「誤差構造が正規分布する」を取り払い，誤差構造を任意の指数型分布分布（例えばポワソン分布や二項分布）ととができる。雑草学，生態学などで扱われ正規分布していないことも多い。その場合に法のひとつが一般化線形モデルである。

変数変換と一般化線形モデル

正規分布していないデータに対して我々のプローチは主に 2 つある。

(1)　変数変換をして正規分布に近づける

変数変換は等分散正規分布していない任意の変換を行って，等分散正規分布に近らパラメータを推定する方法である。最もよく

*（独）農業環境技術研究所
〒 305-8604　つくば市観音台 3-1-3
kennchin@affrc.go.jp
Kentaro Ohigashi : An introductory guide to the statistical analysis – from general linear model to the generalized linear model –
（2010 年 10 月 19 日受理）

のは対数変換（log 変換）であろう。これは要因間の関係が相乗的であったとしても全て加法的な関係に整理することができ，線形モデルにおいて非常に都合が良いためである。相加的な関係になれば，独立な変量の和が正規分布に近づく，という中心極限定理によって，正規性も確保されやすくなる。またよく使われる変換に Box-Cox 変換があるが，この場合は変換結果の等分散性が最も高くなるように調整されるため，変換の適切性に悩む必要がない。

(2) 誤差構造に任意の確率分布を仮定して一般化線形モデルを使う

一般化線形モデルは誤差構造に任意の指数分布族型の確率分布を仮定してデータを変換せずに扱うため，直接パラメータを推定することができる。しかし，一般化線形モデルにおいても，相加的関係にするための変換は行われる。ただしそれは誤差構造として任意の分布を仮定してパラメータを推定した後である，という点が変数変換とは異なる。このパラメータ推定後に相加的関係を作る変換関数のことをリンク関数と呼ぶ。

変数変換において，変換方法が完全な正規分布への近似を与えないのと全く同様に，一般化線形モデルにおいても，選択した分布と誤差構造は完全には一致しないので，常に一般化線形モデルのほうがよい推定を行えるというものではない。

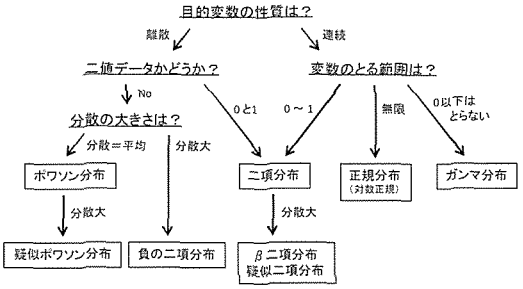
一般化線形モデルの利点

では変数変換を用いた方法と一般化線形モデルとでは何が違うのだろうか。端的に言えば解釈のしやすさである。変数変換した後にパラメータを推定した場合，変換前の期待値は逆変換で簡単に求まるが，パラメータの係数推定値の解釈が直観的にはわかりづらい。一般化線形モデルは，リンク関数による変換前の状態でモデル化されているため，パラメータの係数推定値の意味はそのまま解釈することができる。

誤差構造の選択

変数変換する場合に，どのような変換をするかという目安があるように，一般化線形モデルにおいても，どの誤差構造を選ぶかという目安はある。これは絶対視できるものではないが，誤差構造にどの確率分布を選択するかという問題にある程度の方向性は与えてくれる。それを簡単なフローチャートに示したのが第 1 図である。

まず最も基本的な分岐として，目的とする変量が連続変量なのか離散変量なのか，という違いがある。目的変量が気温や乾物重などの連続型の分布としては正規分布，ガンマ分布などがよく使われる。正規分布は取りうる範囲に制限は無く左右対称である。ガンマ分布はゼロ以下の値は取らない分布で左側が膨らんだ形をしている。



第 1 図　誤差構造の選択

第 1 表　誤差構造の選択

分布	主な変数の例
ポワソン分布	個体数，種子数，
二項分布	発芽率，生存率，植被率
正規分布	気温，千粒重，乾物重
ガンマ分布	種子重，草丈，（乾物重）

目的変量が生存率のような 0 から 1 の割合データの場合，それ自体は連続変量だが，実際には試行回数と成功率によって決定されていると考えられるので，離散型分布である二項分布が適している。調査地に対象種がいたかいなかったかのような，0 か 1 の二値データだった場合も二項分布を使う。二項分布の成功率 p が β 分布に従うと仮定し，より大きな分散（過分散，これについては後述）に対応した分布が β 二項分布である。目的変量が種数などのカウントデータの場合，正の整数しか取らないので離散型の分布を取る。かつ分散が平均と等しい時は，ポワソン分布，分散が大きい場合は負の二項分布である。このような判断はあくまで目安であり，別の分布を使った方が良い場合もある（参考：第 1 表）。その他に，疑似二項分布（quasi-binomial），疑似ポワソン分布（quasi-poisson）などの後述する疑似尤度を用いる確率分布もある。

一般化線形モデルの例

実際に解析してみよう。まずサンプルデータとして Lesnoff *et al.*（2004）によるエチオピアでのコブウシの伝染性胸膜肺炎を観察したデータを用いる。このデー

第2表 コブウシの伝染性胸膜炎の調査データ

	herd	incidence	size	period	size-incidence
1	1	2	14	1	12
2	1	3	12	2	9
3	1	4	9	3	5
4	1	0	5	4	5
5	2	3	22	1	19
6	2	1	18	2	17
⋮					
56	15	0	15	4	15

タはフリーソフトウェアRのパッケージlme4に入っているcbppというデータセットである。データの詳細についてはcbpp{lme4}のヘルプを参考にして欲しい。cbppはherd(群れの番号), incidence(発症頭数), size(群れの大きさ), period(観察時期)の4つの変量で成り立っている。

ここで, 発症割合が観察時期と群れの違いによって, どのように変化するのかをモデル化することを考える。目的変量は割合データなので, 誤差構造として二項分布を仮定する。データを編集して, size-incidence(非発症頭数)という列を作る(第2表)。

このモデル式は以下のように書くことができる。

(発症頭数, 非発症頭数)=観察時期+群れの番号

このモデル式は, 発症頭数と非発症頭数(すなわち発症割合)を観察時期と群れの番号の関数としてモデル化するということを表しており, Rの場合, glm( )という関数で, 以下の様に実装できる。

result<-glm(cbind(incidence, size-incidence)~period+herd, data=cbpp, family=binomial (link=logit))

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(incidence, size - incidence)

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev
NULL                     55    154.817
period  3   40.716        52    114.102
herd   14   45.276        38     68.826
```

第2図 Rのanova{stats}関数による逸脱度分析表の出力結果

summary (result)とすると結果の要約が, anova (result)とすると分散分析表に似た逸脱度分析表が得られ, 分散分析と同様の解釈をすることができる(第2図)。

逸脱度分析

逸脱度分析は, 要因をモデルに加えたことで変化する尤度を用いて, 要因の効果を検証する方法である。尤度とはあるデータセットが得られた場合に, その仮定している前提条件(仮説)の尤もらしさを表す値である。簡単に言えば, 仮説のもとにそのデータセットが得られる確率を示すものである。逸脱度(deviance)とは$D=2\phi[(L_L-L_s)]$で表される値で, モデルによる期待値と実際のデータの差を表している。L_Lは全てのデータをパラメータと見なした, データとモデルが完全に一致しているモデルの対数尤度, L_Sは自分が調べた要因により構成されるモデルの対数尤度である。逸脱度は線形モデルにおける残差平方和に相当するものである。つまり逸脱度Dが低いほどモデルがデータによく当てはまっているといえる。ϕはDispersion Parameterと呼ばれるバラツキに関するパラメータで, 分布によって値が決められており, 正規分布の場合には分散(σ^2)そのものであり, 二項分布, ポワソン分布では1である。

第2図が示す逸脱度分析の結果の意味は, もともと55の自由度があったものが, periodの自由度3とherdの自由度14, 合計17の自由度(パラメータ)を消費して, 帰無仮説(切片のみのモデル)の逸脱度: Null Deviance=154.8が, 対立仮説(要因を加えたモデル)の逸脱度: Residual Deviance=68.8まで減った, ということを意味する。Residual Devianceのことを残差逸脱度という。

この逸脱度の差を用いて, 要因の効果の検定を行う事もできる(第3図)。検定はD/ϕがカイ二乗分布に近似できることからカイ二乗検定が用いられることが

```
Anova Table (Type II tests)


Response: cbind(incidence, size - incidence)
       LR Chisq Df Pr(>Chisq)
period   20.521  3  0.0001324 ***
herd     45.276 14  3.680e-05 ***
---
Signif. codes:  0'***'0.001'**'0.01'*'0.05'.'0.1' '1
```

第3図 RのAnova{car}による逸脱度分析結果の出力結果

多い。また分散分析と同様に，要因の除去順序によって，Type I，Type II，Type III の検定が行える（市川・大橋 1987）。R では car パッケージの Anova 関数などで実行できる。コブウシの調査の場合，観察時期と群れの違いの両方が有意な効果を持っていたと言える（第 3 図）。

要因を加えたモデルによってデータの持っているバラツキをどの程度説明できているかを調べるには $1-\frac{\text{Residual Deviance}}{\text{Null Deviance}}$ で計算される explained deviance を調べるとよい。値が大きいほど，帰無仮説に比べて対立仮説モデルは説明力が高いといえる。ここでの例では，1-(68.826/154.817) = 0.5554…となり，群れの違いと観察時期の違いで，全バラツキの 55.5% を説明できることを意味する。

過 分 散

仮定した誤差構造が，データの母集団が持つ誤差を正確に表現できているかどうかを検定するために，Dispersion Parameter が規程の値と一致しているかを調べる必要がある。データ数が十分に多い場合は，残差自由度で残差逸脱度を割った値 $\frac{\text{Residual Deviance}}{\text{Residual degree of freedom}}$，または $\frac{\textit{Pearson の } X^2}{\text{Residual degree of freedom}}$ で Dispersion Parameter を推定できる。Pearson のカイ二乗とは観測値 O とモデルによる推定値 E の差の二乗を E で割ったものの総和である。式で表せば，$X^2=\sum\frac{(O-E)^2}{E}$ で表せる。前者は R の結果から簡単に計算できるため，目安としては便利だが，安定性に問題があり，後者を使ったほうが正確であるとする意見も多いのでここでは後者で統一する（McCullagh and Nelder 1989，Faraway 2006）。R で計算する場合には

```
dp<- sum(residuals(result, type="pearson")^2)/result$df.res
```

とするとよい。

コブウシのサンプルデータにおける一般化線形モデルの結果について，実際に Dispersion Parameter を計算すると 1.56 程度（前者の残差逸脱度を用いた場合では 1.81）である。今回のモデルで誤差構造として仮定した二項分布の場合，Dispersion Parameter = 1 と仮定されているため，コブウシのデータは二項分布からはややズレているということになる。

このように仮定された Dispersion Parameter よりも実際のデータから計算された Dispersion Parameter が大きい状態を過分散（over dispersion）という。過分散になっている場合は信頼区間が狭くなり，本当は有意ではないパラメータが有意であると判断されてしまう場合が多い。つまり第一種の過誤を犯しやすくなるため，過分散の状態でモデルの結果を解釈するのは避けた方が良い。Dispersion Parameter がいくつ以上までなら許容できるかという明確な基準はないが，基本的には常に過分散を疑うべきである（McClluagh and Nelder 1989）。現実的には 1.5 程度であれば過分散ではない，と判断しても構わないかもしれない。

過分散が何故生じるかということの最も考えうる可能性としては，データが下部構造（サブクラスタ）を持っているという可能性が考えられる。例えば目的変量が雑草重であったとしよう。単に雑草重といっても，複数の種の重量の合計値になっているはずである。この場合，各種の重量がサブクラスタに相当する。我々が見ている目的変量は複数の異なるパラメータを持った確率分布が混合した状態であると考えられる。

逆に，Dispersion Parameter が 1 よりも小さくなっている場合を過小分散（underdispersion）と言う。この場合，データが期待されるバラツキよりもずっと小さくなっていることを示し，一般的な解決方法はない。起こることがまれな現象を調べている場合のデータ不足などが疑われる。解決する為にはより多くのデータを取り直す，などの根本的な対策が必要となる。

疑 似 尤 度

過分散による第一種の過誤を回避する最も簡単な方法として，サブクラスタが混合された状態の分散を調整することで，無理矢理 1 つの分散とみなしてしまうやり方がある。計算上は，Dispersion Parameter に ϕ の推定値である $\hat{\phi}$ を掛けてデータのバラツキを実際より大きく評価する方法である（McCullagh and Nelder 1989）。残念ながらこの方法を使った場合，誤差構造の確率分布が仮定できなくなるので尤度が計算できない。逸脱度の計算には尤度が必要なので，尤度が計算できないと逸脱度が求められないなどのデメリットがある。このような場合，疑似逸脱度（quasi-deviance）を計算することが必要になるが，疑似尤度と尤度の関係は明確でなく，あまり適切な方法とは言えないかもしれない（Penheiro and Chao 2006）。これらの方法は

疑似二項分布（quasi-binomial）や疑似ポワソン分布（quasi-poisson）といった分布を仮定した方法と等価である。

Dispersion Parameter による調整を行っても変わるのは推定値の標準誤差だけであり，推定値そのものは変わらない。Dispersion Parameter を用いた調整の最も大きな問題はデータセットがただひとつのϕを持っている，という仮定の妥当性である。先程の雑草重の話で言えば，全ての種の重量の確率分布が共通のパラメータを持つ，という仮定はあまり現実的なものとは言えないだろう。

一般化線形混合モデル (Generalized linear mixed model)

過分散を回避するための，より現実に即した方法は一般化線形混合モデルを使うことである。一般化線形混合モデルは，目的変数の期待値そのものがなんらかの要因によって変動していることを想定した方法である。過分散はサブクラスタの存在によって生じると先ほど述べたが，一般化線形混合モデルでは，サブクラスタ毎に期待値が異なると仮定し，各サブクラスタにもう一つの誤差（ランダム誤差）を与える。これは，分散分析における変量効果と同じ方法で，正規分布以外の誤差構造を仮定した場合に一般化線形混合モデルとなる。

実際のデータで処理区や，実施日，個体などがサブクラスタとして存在すると考えた場合に，それぞれの処理区や，実施日，個体それぞれにランダムな誤差を与えることになる。例えば処理区が 1 から 10 まであったとすると，処理区 1 のデータには γ_1，処理区 2 のデータには $\gamma_2 \cdots$，処理区 10 のデータには γ_{10} といった具合になる。R の場合には，これらのランダム誤差は線形化されたモデルの切片の推定値に対して与えられることが多く，γ_1 はデータから推定された等分散正規分布に従ったランダムな値を取る。このようなサブクラスタをランダム変量と言い，その効果をランダム効果という。これを簡単な式で表すと以下のようになる。

$$Y = X\beta + Z\gamma + E \qquad (4)$$

前述の（2）式に $Z\gamma$ としてランダム変量を追加している。Z はサブクラスタの水準と繰り返しを表す行列で，γ は等分散正規分布に従うランダムな値のベクトルである。一般化線形モデルは線形モデルの単純な拡張であると述べたとおり，分散分析においても同様の記述がなされる。

実際に R での解析例を見てみよう。一般化線形混合モデルを実装するパッケージとしては glmmML や lme4 などがあるが，ここでは先ほども登場した lme4 を使う。一般化線形モデルで使ったコブウシの伝染性胸膜肺炎のデータのモデル化の場合では時間経過と群れの番号を固定効果として扱った。ここで調査の目的が，コブウシが発症する割合が時間経過とともにどのように変化するか，であったとする。このような目的の調査でも，群れごとに発症割合が異なる可能性が想定される場合は，複数の群れからデータをとることになるだろう。ここで重要なのは，複数の群れからデータを取った目的である。すなわち，発症割合を変動させる要因として群れの違いを考慮するためにデータを取ったのであり，群れの違いが発症割合に与える効果を推定するためにデータを取ったわけではない。このような場合に，目的変量を変動させる要因（ここでは群れの違い）の効果=ランダム効果を考慮して，効果を知りたい要因（ここでは時間経過）の影響=固定効果を解析しようというのが混合モデルの意図である。

雑草学においても，種内変異の影響を考慮するために，複数の地域から採集した材料で実験を行うことはよくあるのではないだろうか。このような場面では各地域をランダム変量として扱った混合モデルを使うことが相応しい。

R による解析は，glmer（　）という関数を用いて以下のように行う。

```
result<-glmer(cbind(incidence, size - incidence)~
period +(1 | herd),
family = binomial, data = cbpp))
```

lme4 の場合，（1 | herd）が切片のランダム効果として herd を扱うという宣言になっている。ランダム効果の指定方法はパッケージによって異なるのでそれぞれの help を参照されたい。

glm（　）と同様に summary（result）で結果の要約が得られる（第 4 図）。

summary（result）の中身をみると，まず *AIC*, *BIC*, 対数尤度，Deviance が計算されている。*AIC* は，赤池の情報量基準の意味で，$AIC = -2\log(L) + 2df$（L は最大尤度，df はモデルに使用する自由度）で表される指標である。この値が小さいほうがよいモデルであるとされる。使用する自由度が大きくなると *AIC* が大きくなるので，一種の重みとなっている。少ないパラメータでよりよい説明力を持つモデルを選ぶ，モデル選択と

```
Generalized linear mixed model fit by the Laplace approximation
Formula: cbind(incidence, size - incidence) ~ period + (1 |
herd)
   Data: cbpp
   AIC   BIC logLik deviance
 110.1 120.2 -50.05    100.1
Random effects:
 Groups Name        Variance Std.Dev.
 herd   (Intercept) 0.4125   0.64226
Number of obs: 56, groups: herd, 15

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3985     0.2279  -6.137 8.42e-10 ***
period2      -0.9923     0.3054  -3.249 0.001156 **
period3      -1.1287     0.3260  -3.462 0.000537 ***
period4      -1.5804     0.4288  -3.686 0.000228 ***
---
Signif. codes:  0'***'0.001'**'0.01'*'0.05'.'0.1' '1

Correlation of Fixed Effects:
        (Intr) perid2 perid3
period2 -0.351
period3 -0.329  0.267
period4 -0.249  0.202  0.186
```

第4図 glmer{lme4}による一般化線形混合モデルの推定結果

呼ばれる作業に用いるものである。*BIC*はベイズ情報量基準と呼ばれるもので，*AIC*と同じような指標であり，自由度への重みのつけ方が異なる。これらの指標を用いてモデル選択を行うと，必要のないパラメータを省いて，モデルの予測力を最大化することができる。モデル選択やそれらの指標についての詳細は下平（2004）を参照されたい。

第4図の解釈として注意すべきは，Devianceの値である。Devianceとだけ書かれているが，意味するところはglmで言うところのResidual Devianceであり（第4図），前述の一般化線形モデルを使った場合のperiodとherdを固定効果として扱ったDeviance，68.826より大きくなっている（第2図）。

第4図の「Random effects：」以降にランダム変量の分散（Variance）と標準偏差（Std.Dev.）が推定されている。このモデルの場合，群れを切片のランダム効果として指定したので，ランダム効果の平均値は固定効果の切片（第4図のFixed effectsのIntercept）に相当する。これらは（4）式のγが従う正規分布のパラメータのことであり，今回の結果で言えばランダム変量として指定した群れによる変動が，γは平均-1.3985，分散0.4125の正規分布に従っていると推定されたことになる。今回，ランダム変量に指定した群れは15水準と十分な数があったが，一般にランダム変量の水準が3以下の場合は，その標準偏差の推定が難しくなるため，ランダム変量として指定しない方が良い（Bolker *et al.* 2009）。

次に第4図の「Fixed effects：」以降には固定効果の検定結果が示されている。名義変量の場合は表示されていない水準（period1）に対して，表示されている水準（period2, 3, 4）間の差が示されている。ここでのPr値（Pr（>|z|））は，対照となる水準との差を示す偏回帰係数（Estimate）が0と異なるかどうかで求められており，係数が負である事からperiod1よりも他の水準でのコブウシの胸膜炎の発症割合が有意に低いといえる。

第4図の「Correlation of Fixed Effects：」以降では固定効果間の相関が示されている。これは，period1とperiod2, 3, 4それぞれの間の偏回帰係数を求める際に利用される。以上の結果を受けて，さらに水準間のどこに差があるかを検定したい場合は，Rの場合はmultcompというパッケージのglht関数を利用して多重比較を行うことができる。

今回のコブウシのデータの場合，ランダム効果の自由度を2とすると，固定効果のperiodの自由度3と併せて，消費する自由度は5となり，残差自由度は50となる。それを用いて混合モデルのDispersion Parameterを計算すると1.26となり，群れの番号を固定効果として扱った一般化線形モデルの時の1.56と比較すると過分散傾向は改善されていると言える。

ランダム効果を導入することの実際的な利点は，自由度を節約すること，つまり推定するパラメータの数を減らせることである。自由度は逸脱度分析の尤度比検定やモデル選択などで使われ，結果の解釈に大いに影響している。先ほどの例で言えば，herdに使われていた14の水準間の差を示す偏回帰係数を求める必要があったのが，herd間のバラツキを示す正規分布の平均と標準偏差を推定するだけでよくなる。しかしランダム効果を導入した場合にどの程度自由度を節約できるのかは，データから推定されることになる。サブクラスタのバラツキが一つの正規分布で近似できそうなのであれば，自由度を大きく節約することができるが，一つの正規分布で近似できそうにないようであれば，自由度の節約を小さくするべきである。そうすることで，結果の有意性はより安全なものとなるからだ。一般にサン

プル数が多ければ正規分布で近似しやすくなる。推定される自由度 df は，最も節約できたときに 1，全く節約できなかった時にはランダム変量の水準 N−1 となり，その間のいずれかの値（1≦df≦N-1）を取ることになる。今のところ，混合モデルの場合の自由度については推定方法が複数提案されており，それぞれ相応しい場面があると言われているが，議論が行われている段階である。残念ながら現段階の R ではランダム効果の自由度を最も小さく取る方法しか選択することはできない。サンプル数が大きい時（残差自由度が 25 以上ある時）は，推定方法による差は小さいとされている。もしサンプルサイズが小さく，p 値が微妙な値であり，それが気になるようであれば，SAS を使用して自由度の推定方法を Kenward-Roger（KR）に指定するとよいかもしれない（Bolker *et al.* 2009）。

ま と め

本稿では分散分析や回帰分析を包括した概念として，線形モデルがあり，その拡張として一般化線形モデルが使える，ということを述べてきた。しかし本稿で紹介したコブウシの例で見たように，二項分布やポワソン分布を仮定した一般化線形モデルは過分散が解消されないことも多く，第一種の過誤を犯してしまう危険性がある。そのような事態を避けるために，筆者が薦めたいのは一般化線形混合モデルである。一般化線形混合モデルは実際の実験現場でよく体験される，個体差や，調査者による偏りなどを表現することに非常に適している。一般化線形混合モデルは我々が，本当に興味がある変量の効果を，よりよく推定する上での強力なツールとなるであろう。

謝　　辞

本特集は第 49 回雑草学会講演会の「若手の会」の勉強会での内容について取りまとめたものです。若手の会開催にあたって，ご協力頂いた大会関係者の皆様に感謝いたします。

引用文献

Bolker, B.M., M.E. Brooks, C.J. Clark, S.W. Geange, J.R. Pulsen, M.H.H. Stevens and J-S.S White 2009. Generalized linear mixed models : a practival guide for ecology and evolution. Trends in Ecol. & Evol. 24, 127-135.

Faraway, J.J. 2006. Extending the Linear Model with R : Generalized linear, mixed effects and nonparametric regression models. Chapman & Hall CRC, Boca Raton.

Lesnoff, M., G. Laval, P. Bonnet, S. Abdicho, A. Workalemahu, D. Kifle, A. Peyraud, R. Lancelot and F. Thiaucourt 2004. Within-herd spread of contagious bovine pleuropneumonia in Ethiopian highlands. Prev. Vet. Med. 64, 27-40.

Pinheiro, J.C. and E.C. Chao 2006. Efficient laplacian and adaptive gaussian quadrature algorithm for multilevel generalized linear mixed models. J. Comput. Graph. Statist. 15, 58-81.

McCllagh, P. and J.A. Nelder 1989. Generalized linear models. second edition Chapman & Hall London.

市川伸一・大橋靖雄・竹内啓 1987. SAS によるデータ解析入門. 東京大学出版会，東京.

下平英寿 2004. 情報量規準によるモデル選択とその信頼性評価　モデル選択－予測・検定・推定の交差点. 統計科学のフロンティア 3，岩波書店，東京.