統計数理(2013)
第 61 巻 第 2 号 271–287

特集「環境リスクと統計解析—データ基盤構築と解析—」
［原著論文］

ゼロの多いデータの解析：負の2項回帰モデルによる傾向の過大推定

南 美穂子[1] ・Cleridy E. Lennert-Cody[2]

（受付 2013 年 3 月 3 日；改訂 6 月 18 日；採択 7 月 30 日）

要　　旨

動物の生態研究や生物資源評価のために観測される個体数などの計数データには，多くのゼロが含まれることがよくある．また，観測された環境や時間などの違い，または，観測されることのない潜在的な条件の違いなどの影響により，ほとんどの場合，個体数のばらつきも平均に比べてはるかに大きい．負の 2 項回帰モデルは，分散が平均よりも大きい過分散の計数データに対する回帰モデルとして生物個体数の解析に多くの研究で用いられている．しかし，Minami et al.(2007)はゼロが多いデータに対して負の 2 項回帰モデルを用いて，他の変数の影響を調整した年ごとのサメの平均混獲数(混獲とは，漁業の際に対象魚種とは異なる種を意図せずに捕獲してしまうことをいう)を推定すると，混獲数の減少傾向を過大に推定することを示した．これは，例えば，研究の目的が生息数の増減の推測にある場合には，実際以上に急激な減少という誤った結論を導くことにつながり危険である．本論文では，ゼロの多いデータに対して負の 2 項回帰モデルをあてはめた場合に傾向の過大推定が起こることを実データで示すとともにシミュレーションでも再現し，どのような状況のときに過大推定が起こるか，それはどのような現象が起きた結果なのか，また，理論的にどのように説明できるのかを考察する．

キーワード： 過分散，影響関数，サイズパラメータ，zero-inflated 負の 2 項回帰モデル，クックの距離，テコ比．

1. はじめに

観察された計数データがゼロの値を多く含むことはしばしばある．例えば，製品の製造過程における不良品の個数(Lambert, 1992)，生産年齢にある人が病気のために主たる活動を休んだ日数(Frankenberg and Thomas, 2000; Lam et al., 2006)，あるいは，生態学や環境調査で収集された単位面積あたりの動物数（Welsh et al., 1996)など，多くの研究においてゼロの値を多く含む計数データが観察され，その解析方法について議論がされている．そのようなデータに対して，どのようにゼロが多いのかを考慮し，適切なモデル，例えば，zero-inflated ポアソンモデル(Lambert, 1992)，zero-inflated 負の 2 項回帰モデル(Greene, 1994)，ハードルモデル(Mullahy, 1986)などを選択することは重要なことである．しかし，ゼロが多いことに注意を払う必要性がまだ十分に認識されておらず，ゼロの割合が高いデータに対して，そのことを十分

[1] 慶應義塾大学 理工学部：〒223–8522 神奈川県横浜市港北区 3–14–1
[2] Inter-American Tropical Tuna Commission, 8606 La Jolla Shores Dr. La Jolla, CA 92037–1508, U.S.A.

に考慮せずに解析している場合がまだある．本論文は，ゼロが多いデータをゼロが多いことに注意を払わずに解析することによって誤った推測をしてしまう危険性を示し，ゼロが多いことを考慮した適切なモデルを用いる必要性の認識を高めることを意図するものである．

計数データを反応変量とする回帰モデルの代表的なものにポアソン回帰モデルがある．しかし，現実のデータでは，分散が期待値より大きい「過分散」を示し，ポアソン回帰モデルが適切であるとは思われない場合も多い．負の 2 項回帰モデル（Lawless, 1987; Hilbe, 2007）は，そのようなデータに対してよく用いられるモデルで，反応変量が期待値 μ の負の 2 項分布に従い，期待値 μ と説明変数ベクトル $\boldsymbol{x}$ との関係がリンク関数 $g(\cdot)$ を用いて $g(\mu)=\boldsymbol{x}^T\boldsymbol{\beta}$ と表せるとする回帰モデルである．負の 2 項分布の確率関数は以下で与えられる．

$$f_{NB}(y|\mu,\theta)=\frac{\Gamma(\theta+y)}{\Gamma(\theta)\Gamma(y+1)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{y} \tag{1.1}$$

ここで μ は平均，$\theta(>0)$ はサイズパラメータである．負の 2 項分布は θ が正の整数であるとき，コイン投げで表が θ 回出るまでに裏が出た回数の分布と考えられるが，パラメータ θ の範囲は正の実数に拡張できる．リンク関数としてよく用いられるのは対数リンクであり，ここでも対数リンク

$$\log(\mu)=\boldsymbol{x}^T\boldsymbol{\beta} \tag{1.2}$$

を用いたモデルについて考える．

サイズパラメータが既知のとき負の 2 項分布は指数型分布族に属するので，負の 2 項回帰モデルは一般化線形モデルであるが，サイズパラメータが未知で推定しなければならない場合は，一般化線形モデルではない．負の 2 項分布の分散は

$$\mathrm{Var}(Y)=\mu+\frac{1}{\theta}\mu^2$$

で，θ を $+\infty$ としたとき負の 2 項分布はポアソン分布に帰着する．

図 1 はポアソン分布（負の 2 項分布で $\theta=\infty$）と $\theta=2$ および $\theta=0.5$ のときの負の 2 項分布の確率関数を示したものである．平均がある程度大きいとき，サイズパラメータの値によって分布はほぼ対称であったり，右裾の重い歪んだ形を取ったりすることがわかる．負の 2 項分布は，ポアソン分布の平均がガンマ分布に従う確率変動を持つときの周辺分布と考えられる．よって対数リンクを用いた負の 2 項回帰モデルは，ポアソン分布に従う反応変量の平均のばらつきの一部は説明変数で表すことができるが，説明変数では説明できない変動もある場合の回帰モデルと理解できる．つまり，確率変数 R が，形状パラメータもレートパラメータも θ のガンマ分布に従い，$R=r$ が与えられたときの Y は平均 m が

$$\log(m)=\boldsymbol{\beta}^T\boldsymbol{x}+\log r$$

であるようなポアソン分布に従うときに，Y の周辺確率関数 $f(y;\boldsymbol{x},\boldsymbol{\beta},\theta)$ は

$$\begin{aligned}f(y;\boldsymbol{x},\boldsymbol{\beta},\theta)&=\int_0^{\infty}\exp(-re^{\beta^T\boldsymbol{x}})\frac{r^y e^{(\beta^T\boldsymbol{x})y}}{y!}\frac{\theta^{\theta}}{\Gamma(\theta)}r^{\theta-1}e^{-\theta r}dr\\&=\frac{\Gamma(\theta+y)}{\Gamma(\theta)\Gamma(y+1)}\left(\frac{\theta}{\theta+e^{\beta^T\boldsymbol{x}}}\right)^{\theta}\left(\frac{e^{\beta^T\boldsymbol{x}}}{\theta+e^{\beta^T\boldsymbol{x}}}\right)^{y}\end{aligned}$$

となり，負の 2 項回帰モデルに対するものとなる．

負の 2 項分布は平均が大きくてもサイズパラメータの値によってゼロを取る確率が高くなるため，ゼロが多く含まれるデータに対しても，そのことに特に注意が払われずに負の 2 項回帰

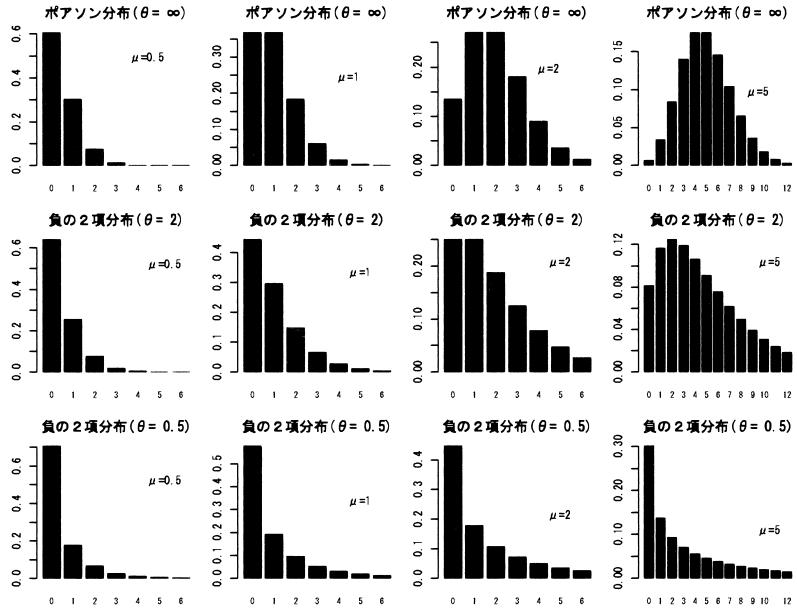

図 1.　ポアソン分布（上段）と負の 2 項分布（中段：サイズパラメータ $\theta = 2$，下段：サイズパラメータ $\theta = 0.5$）の確率関数．平均は左の列から，0.5, 1, 2, 5.

モデルが用いられることがある．しかし，Minami et al.(2007) はサメの混獲数に対して複数の回帰モデルをあてはめたところ，負の 2 項回帰モデルは他の回帰モデルに比べてサメの生息数の変動傾向を過大に推定するということを示した．

本論文では，まず，2 節でゼロの多いデータに負の 2 項回帰モデルを用いた場合に傾向を過大推定する例としてサメの混獲数データの解析結果を示し，シミュレーションデータでも現象が再現できることを示す．3 節ではどのような状況の場合に傾向の過大推定が起こるのかを考察し，4 節でそれはどのような現象が起きた結果なのか，またどのような理由で起きるのかのメカニズムを解明し，5 節でまとめを述べる．

2.　負の 2 項回帰モデルによる傾向の過大推定

ゼロの多いデータに負の 2 項回帰モデルを当てはめた場合に傾向を過大に推定する例として，まず，サメの混獲データにおける解析例を紹介し，次に，ゼロの多いデータを生成して解析し，同様な現象が起こることを確認する．

2.1　サメの混獲データに対する傾向の過大推定

負の 2 項分布の確率関数は，サイズパラメータの値が小さいとき単調減少でゼロを取る確率が他の値を取る確率よりも高くなる．そのため，ゼロの多い計数データにもゼロが多いことに特に注意を払わずに負の 2 項回帰モデルを用いて解析されることがよくある．しかし，Minami et al.(2007) はマグロ巻き網漁におけるサメの混獲数データを負の 2 項回帰モデルで解析したところ，混獲数の減少を過度に表すことを示した．このデータはマグロ巻き網漁の操業ごとに対象魚種の漁獲量，混獲種の混獲数，気象要因，環境要因などを観測したものである．クロトガリザメの混獲数は 30,000 以上の操業のうち 50% 以上の操業でゼロであった．Minami らは，

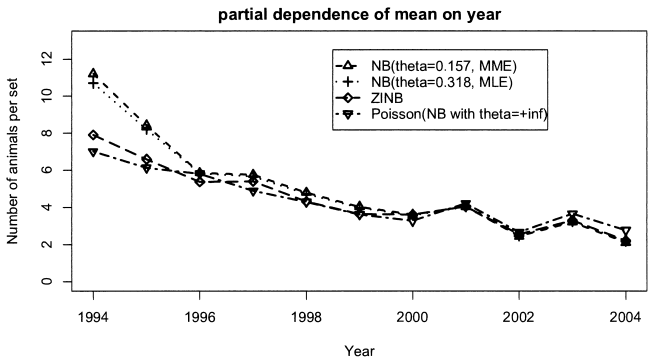

図 2. クロトガリザメの標準化平均混獲数の年次変動推定値：負の 2 項回帰モデル(モーメント法推定値 $\hat{\theta}_{MM}=0.157$)，負の 2 項回帰モデル(最尤推定値 $\hat{\theta}_{ML}=0.318$)，zero-inflated 負の 2 回帰モデル(サイズパラメータは最尤推定値)，ポアソン回帰モデルの推定結果を用いた年に関する partial dependence を標準化年次平均とした．

操業には混獲の起こらない完全状態(ゼロ状態)と負の 2 項回帰モデルに従う不完全状態の 2 つの状態があり，どちらの状態を取るかがロジスティック回帰モデルに従うとする zero-inflated 負の 2 項回帰モデルを用いてこのデータを解析し，負の 2 項回帰モデルやポアソン回帰モデルによる結果と比較した．図 2 はその結果を示すもので，負の 2 項回帰モデル(サイズパラメータはモーメント法推定値 $\hat{\theta}_{MM}=0.157$)，負の 2 項回帰モデル(サイズパラメータは最尤推定値 $\hat{\theta}_{ML}=0.318$)，zero-inflated 負の 2 項回帰モデル(サイズパラメータは最尤推定値)，ポアソン回帰モデルの推定結果を用いて年に関する partial dependence (Hastie et al., 2009，付録 A)を計算し，標準化年次平均として示したものである．Partial dependence は他の変量の影響を除くために他の変量に関する期待値を取ったものであり，「標準化」とは環境要因や操業条件の影響を取り除いて評価したことを意味する(Maunder and Punt, 2004)．

この期間にこの水域では，操業方法や混獲に対する対処法などの大きな変化は無かったので，混獲数の推移は，生息数の推移を反映していると考えられる．図 2 の結果は，他のモデルでは，クロトガリザメの生息数は 2004 年には 1994 年の 40% に減少したということであるが，負の 2 項回帰モデルでは，20% にまで減少したと推測することになる．なぜ，これほどまでにモデルによって違いが出るのであろうか．

2.2 シミュレーションデータによる結果の再現

ここでは，負の 2 項回帰モデルに従うデータの一部をゼロに置き換えてゼロの多いデータを作成し，それに負の 2 項回帰モデルをあてはめて前節で示した現象を再現する．

2.2.1 ゼロの多いデータの生成

まず，2 つの水準 A と B からなる因子変量と連続変量 x を説明変量とする負の 2 項回帰モデルに従う確率変数 W の値 $w_i (i=1,\ldots,2N)$ を以下の手順で生成し，その一部をゼロに置き換えて $y_i (i=1,\ldots,2N)$ を生成する．この生成モデルについては 2.2.3 節で説明する．N は各水準の標本数で $N=10000$ である．

(1) 区間 $[0,1]$ の一様分布に従う値を N 個生成し，$u_{i'}$ $(i'=1,\ldots,N)$ とおく．
(2) $x_{i'}=5.5u_{i'}^2-2.5$ $(i'=1,\ldots,N)$ とする．
(3) 以下の平均 $\mu_{i'}^A, \mu_{i'}^B$ $(i'=1,\ldots,N)$ を持ち，サイズパラメータ $\theta=0.6$ の負の 2 項分布に

従う値 $w_{i'}^A, w_{i'}^B$ $(i'=1,\ldots,N)$ を生成する．

$$\log(\mu_{i'}^A) = x_{i'} + 3$$
$$\log(\mu_{i'}^B) = x_{i'} + 1.5$$

$w_{i'}^A$ と $w_{i'}^B$ $(i'=1,\ldots,N)$ を合わせて w_i $(i=1,\ldots,2N)$ と記し，同様にその平均も $\mu_{i'}^A$ と $\mu_{i'}^B$ $(i'=1,\ldots,N)$ を合わせて μ_i と記す．$I_B(i)$ を標本 i が水準 B であれば 1，そうでなければ 0 であるような関数とすると平均 μ_i は

$$\log(\mu_i) = x_i + 3 - 1.5\ I_B(i)$$

と表せる．

(4)　確率 p_i を

$$\log\left(\frac{p_i}{1-p_i}\right) = -3x_i - 5 + 2.5 I_B(i)$$

で定義し，w_i を確率 p_i で 0 に置き換えて y_i と置く．つまり，

$$y_i = \begin{cases} 0 & \text{with probability} \quad p_i \\ w_i & \text{with probability} \quad 1-p_i \end{cases}$$

とする．

y_i は w_i に比べてゼロが多く，また，その生成方法により，

$$y_i \le w_i \quad (i=1,\ldots,2N)$$

であることに注意されたい．

2.2.2　ポアソン回帰モデルと負の 2 項回帰モデルの当てはめ

負の 2 項回帰データ $\boldsymbol{w}$ とゼロの多いデータ $\boldsymbol{y}$ に，平均構造に対数線形モデルを仮定したポアソン回帰モデルと負の 2 項回帰モデルをあてはめた．統計ソフトウェアは R を用いた．負の 2 項回帰モデルのサイズパラメータの推定には，最尤推定法（MASS パッケージの関数 `glm.nb`）と，モーメント法（mgcv パッケージ Ver. 1.3-31 の関数 `gam`）の 2 つの方法を用いた．その他のパラメータの推定はいずれも最尤推定法による．図 3 は 2 つの回帰モデルによる当てはめ値の水準ごとの平均のプロットで，表 1 に 2 つの回帰モデルのパラメータ推定値と当てはめ値の水

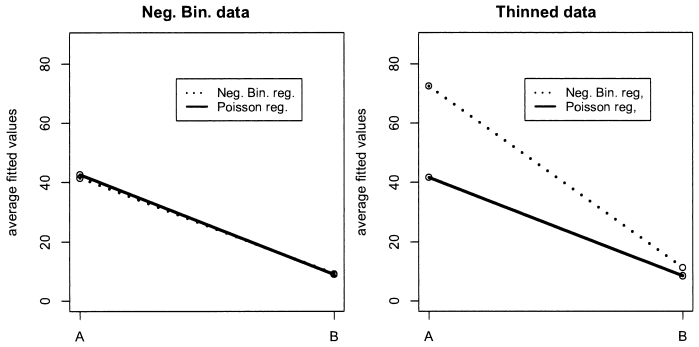

図 3.　当てはめ値の水準ごとの平均値．$\boldsymbol{w}$（左）と $\boldsymbol{y}$（右）に負の 2 項回帰モデル（点線）とポアソン回帰モデル（実線）を当てはめたときの，水準 A と B の当てはめ値の平均値．

表 1. 負の 2 項回帰モデルとポアソン回帰モデルによるパラメータ推定値.

データ		負の 2 項回帰データ $\boldsymbol{w}$		ゼロの多いデータ $\boldsymbol{y}$		
パラメータ (( ) 内は真の値)		ポアソン回帰モデル	負の 2 項回帰モデル (MLE)	ポアソン回帰モデル	負の 2 項回帰モデル (MLE)	負の 2 項回帰モデル (MME)
β_0	(3.0)	2.969	2.982	2.782	2.670	2.703
β_B	(-1.5)	-1.560	-1.502	-1.594	-1.858	-1.893
β_x	(1.0)	1.026	1.003	1.111	1.414	1.441
θ	(0.6)	—	0.605	—	0.422	0.305
当てはめ値平均 : A		42.67	41.49	41.75	72.53	77.40
当てはめ値平均 : B		8.96	9.24	8.48	11.31	11.66

準ごとの平均を示した．負の 2 項回帰データ ($\boldsymbol{w}$, 左図) に対しては 2 つの回帰モデルで平均当てはめ値にあまり違いがないが，ゼロの多いデータ ($\boldsymbol{y}$, 右図) に対しては，負の 2 項回帰モデルによる水準 A の平均当てはめ値は，ポアソン回帰モデルによる平均当てはめ値よりもはるかに大きい値となった．

2 つの回帰モデルで仮定した平均構造は

$$\log(\mu_i) = \beta_0 + \beta_x x_i + \beta_B I_B(i)$$

である．負の 2 項回帰データ $\boldsymbol{w}$ に対しては，この平均構造はデータを生成したモデルなので，どちらの回帰モデルの最尤推定量も一致性を持つ．パラメータの推定値も 2 つの回帰モデルで大きく異なることはなく，データ生成に用いたパラメータ値 ($\beta_0 = 3$, $\beta_B = -1.5$, $\beta_x = 1$ and $\theta = 0.6$) と近い値であった．負の 2 項回帰モデルはデータを生成したモデルなので，パラメータの最尤推定量は有効推定量であり，推定値はデータ生成に用いたパラメータ値により近いものとなっている．

一方，ゼロの多いデータ $\boldsymbol{y}$ に対しては，2 つの回帰モデルで仮定した平均構造は正しいものではない．しかし，対数リンクはポアソン回帰モデルの自然リンク関数であるので，各水準の当てはめ値の平均は観測値の平均と等しくなる (付録 B)．一方，負の 2 項回帰モデルによる水準 A の当てはめ値の平均は 72.53 ($\hat{\theta}_{MLE}$,0.422) と 77.40 ($\hat{\theta}_{MME}$,0.305) で，観測値の平均 41.75 (=ポアソン回帰モデルによる当てはめ値の平均) よりもはるかに大きい値となった．

負の 2 項回帰データ $\boldsymbol{w}$ とゼロの多いデータ $\boldsymbol{y}$ は，すべての i に対し $w_i \geq y_i$ であるから，ゼロの多いデータ y_i の水準 A の平均 41.75 は負の 2 項回帰データ w_i の平均 42.67 より小さい．しかし，負の 2 項回帰モデルによる水準 A の当てはめ値は，サイズパラメータに最尤推定値を用いた場合で比較すると，ゼロの多いデータの当てはめ値 $\hat{y}_i$ の平均 72.53 は 負の 2 項回帰データの当てはめ値 $\hat{w}_i$ の平均 41.49 よりはるかに大きい値となった．

β_x と β_B の最尤推定値は，いずれも符号は変わらないが，絶対値はデータを生成した値よりも大きくなっている．サイズパラメータの推定値は共にデータを生成した値 $\theta = 0.6$ よりも小さいが，モーメント法推定値 (0.305) の方が最尤推定値 (0.422) より小さい．

2.2.3 ゼロの多いデータの生成モデル

ゼロの多いデータ $\boldsymbol{y}$ は，負の 2 項回帰モデルに従う値 $\boldsymbol{w}$ をロジスティック回帰モデルに従う確率変数の値によってゼロに置き換えたデータであった．この $\boldsymbol{y}$ の従うモデルは zero-inflated 負の 2 項回帰モデル (Greene, 1994) と呼ばれる．これは平均 μ の負の 2 項分布に従う確率変数の値を確率 p でゼロに置き換えたときの分布である zero-inflated 負の 2 項分布の，パラメータ μ および p と説明変数を対数リンク関数，ロジスティックリンク関数で結びつけた回帰モデルである．

負の 2 項分布に従う確率変数の値を確率 p でゼロに置き換える，ということは言い換えれば，確率 p でゼロの値をとり，確率 $1-p$ で負の 2 項分布に従うということなので，zero-inflated 負の 2 項回帰モデルの確率関数は

$$f(y|\boldsymbol{b},\boldsymbol{g},\boldsymbol{\beta},\boldsymbol{\gamma},\theta)=\begin{cases} p+(1-p)f_{NB}(0|\boldsymbol{b},\boldsymbol{\beta},\theta) & \text{for} \quad y=0 \\ (1-p)f_{NB}(y|\boldsymbol{b},\boldsymbol{\beta},\theta) & \text{for} \quad y=1,2,\ldots \end{cases} \tag{2.1}$$

で与えられる．ここで，

$$\begin{aligned} f_{NB}(y|\boldsymbol{b},\boldsymbol{\beta},\theta) &= \frac{\Gamma(\theta+y)}{\Gamma(\theta)\Gamma(y+1)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{y} \\ \log(\mu) &= \beta_0+b_1\beta_1+\cdots+b_{k_\beta}\beta_{k_\beta}=\boldsymbol{b}^T\boldsymbol{\beta} \\ \mathrm{logit}(p) &= \gamma_0+g_1\gamma_1+\cdots+g_{k_\gamma}\gamma_{k_\gamma}=\boldsymbol{g}^T\boldsymbol{\gamma} \end{aligned}$$

であり，$\boldsymbol{b}$ と $\boldsymbol{g}$ はロジスティック回帰モデル部分，および，負の 2 項回帰部分の説明変数ベクトルで，$\boldsymbol{\beta}$ と $\boldsymbol{\gamma}$ は対応する係数ベクトルである．

ゼロの値しかとらない状態をゼロ状態（または，完全状態）と呼び，負の 2 項回帰モデルに従う状態を不完全状態と呼ぶ．不完全状態のときも負の 2 項回帰モデルによる 0 の値を取ることがあることに注意されたい．よって，y が正の整数値のときは不完全状態であったとわかるが，y が 0 のときはゼロ状態であったのか，不完全状態だが 0 の値を取ったのかはわからない．

3. 傾向の過大推定が起こる状況

なぜ傾向の過大推定が起こるのかを考えるために，まず，混獲データ，およびその解析結果にどのような特徴があるのかを調べ，どのような状況の場合に傾向の過大推定が起こるのかを考えよう．

3.1 サメの混獲データの場合

サメの混獲データは 1994 年から 2004 年の間に東部太平洋で操業された大型船によるマグロ巻き網漁で混獲されたクロトガリザメの個体数を乗船した全米熱帯マグロ類委員会の監視員が操業ごとに記録したものである（Román-Verdesoto and Orozco-Zöller, 2005; Minami et al., 2007）．サメの混獲数は多くのゼロを含むが，その比率は年ごとにばらつきがある．図 4 は 1994 年と 2004 年の操業ごとの混獲数の頻度分布を示したものである．1994 年にはゼロの比率は 37.4% で

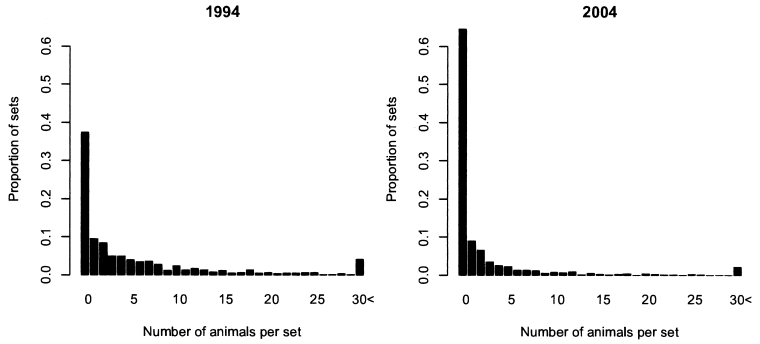

図 4.　1994 年と 2004 年の漁あたりのクロトガリザメ混獲数の頻度表．

表 2. サメの混獲データに対する係数推定値．1995 年から 2004 年の係数は，各年を水準とする因子変量として 1994 年との対比．変量 Z_1, Z_2, Z_3, Z_4 は環境要因や操業条件を表すものである(Minami et al., 2007)．$\bar{Y}_{1994}$ は 1994 年の partial dependence (Hastie et al., 2009)．

	ZINB		ポアソン	負の 2 項	負の 2 項
	ロジスティック	負の 2 項		MLE	MME
size parameter θ		0.568	$(+\infty)$	0.318	0.157
year 1994	5.461	-3.544	-4.852	-6.593	-6.739
year 1995	-0.206	-0.199	-0.134	-0.266	-0.285
year 1996	0.781	-0.299	-0.187	-0.609	-0.648
year 1997	0.836	-0.286	-0.361	-0.635	-0.664
year 1998	0.050	-0.602	-0.490	-0.818	-0.848
year 1999	1.687	-0.513	-0.661	-0.991	-1.022
year 2000	2.039	-0.420	-0.764	-1.097	-1.132
year 2001	1.575	-0.429	-0.509	-0.979	-1.021
year 2002	1.536	-0.900	-0.967	-1.468	-1.522
year 2003	1.895	-0.556	-0.652	-1.195	-1.249
year 2004	2.447	-0.773	-0.928	-1.594	-1.669
Z_1	-0.260	0.188	0.225	0.284	0.288
Z_2	-0.138	0.263	0.283	0.301	0.303
Z_3	-0.252	0.093	0.125	0.180	0.186
Z_4	0.0029	-0.0034	-0.0044	-0.0045	-0.0046
AIC	64510		170100	65640	67530
$\bar{Y}_{1994}$	7.914		7.01	10.70	11.21

あったが 2004 年は 64.6% であった．また，ゼロの比率が増加しているだけでなく，混獲が起こった場合の混獲数も減少している．

このデータに対して，Minami et al.(2007) は zero-inflated 負の 2 項回帰モデル，ポアソン回帰モデル，サイズパラメータに最尤推定値，および，モーメント法推定値を用いた負の 2 項回帰モデルを当てはめた．Zero-inflated 負の 2 項回帰モデルのサイズパラメータ，および，すべてのモデルのその他のパラメータには最尤推定法を用いた．説明変量には各年を水準とする因子変量と操業条件および環境要因を表す変数を用いた．表 2 はこれらのモデルによる回帰係数推定値で，1995 年から 2004 年の係数は，各年を水準とする因子変量として 1994 年との対比である．

Zero-inflated 負の 2 項回帰モデルの係数推定値に関しては，ロジスティック回帰部分の係数推定値は 1995 年を除いてすべて，負の 2 項回帰部分の係数推定値と反対の符号となっている．これは，ある要因の変化が混獲数の減少と関連があるのであれば，この変化はゼロ状態の確率の増加とも関連があるということであるから，自然なことのように思われる．Zero-inflated 負の 2 項回帰モデルにおけるサイズパラメータの最尤推定値は 0.568 であった．

AIC は zero-inflated 負の 2 項回帰モデルが最も小さく，次に負の 2 項回帰モデルでポアソン回帰モデルは他の 2 つのモデルよりはるかに大きい値であった．Zero-inflated 負の 2 項回帰モデルのロジスティック回帰部分の係数推定値の多くが有意に 0 から離れており，また，AIC が最小であるので，このデータには zero-inflated 負の 2 項回帰モデルが他のモデルより適切だと思われる．また，Vuong 検定(Vuong, 1989) も負の 2 項回帰モデルより zero-inflated 負の 2 項回帰モデルを支持する結果となった．

次に，ポアソン回帰モデルと負の 2 項回帰モデル ($\hat{\theta}_{ML} = 0.318$) による係数推定値を比較しよう．すべての係数推定値が 2 つのモデルで同じ符号となっている．しかし，その絶対値は

すべての変数で負の 2 項回帰モデル ($\hat{\theta}_{ML} = 0.318$) による推定値の方が大きい．また負の 2 項回帰モデルでサイズパラメータの最尤推定値 $\hat{\theta}_{ML} = 0.318$ を用いた場合とモーメント法推定値 ($\hat{\theta}_{MM} = 0.157$) を用いた場合の係数推定値も，いずれも符号は同じだが絶対値はモーメント法推定値を用いた場合の方が大きくなった．

この他に，サイズパラメータを推定せずに様々な値を与えて負の 2 項回帰モデルを当てはめてみたが，サイズパラメータの値が小さいほど係数推定値の絶対値は大きくなった．

3.2 過大推定の起こる状況とそれによって起こりうること

サメ混獲数データの解析結果に対する考察を踏まえて，傾向の過大推定の起こる状況をまとめると

(1) 反応変量の観測値にはゼロが多くあり，zero-inflated 負の 2 項回帰モデルに従うと考えるのは妥当である．また，ゼロ状態を取る確率は説明変数の値に依存する．
(2) 正の値の観測値は散らばりが大きく，負の 2 項回帰モデル部分のサイズパラメータは小さい値である．
(3) Zero-inflated 負の 2 項回帰モデルを当てはめたとき，負の 2 項回帰部分でもロジスティック回帰部分でも有意となる説明変数があり，その係数の符号は逆になっている．

である．2 節で扱ったシミュレーションデータ w も上記が該当するものであった．w は zero-inflated 負の 2 項回帰モデルに従い，ゼロ状態を取る確率は変数 x と I_B に依存する．また，サイズパラメータの値は 0.6 と小さい．変数 x と I_B は両方の回帰部分の説明変数であり，係数は 2 つの回帰部分で符号が逆になっている．

上記が該当するデータに負の 2 項回帰モデルを当てはめると起こることは

(1) サイズパラメータの推定値は，zero-inflated 負の 2 項回帰モデルにおけるサイズパラメータの値に比べて小さい値になる．
(2) 係数推定値は，データが従う zero-inflated 負の 2 項回帰モデルの負の 2 項回帰部分の係数と符号は同じだが，絶対値は大きくなる．
(3) その結果，推定したモデルによる当てはめ値は，説明変数の値がその係数と同じ符号で絶対値が大きいとき，大きな値となる．そして，その変数に対する当てはめ値の傾向は過大に推測される．

であろうと考えられる．

4. 傾向の過大推定のメカニズムの解明

ゼロの多いデータに負の 2 項回帰モデルを当てはめると前節で述べたことがなぜ起こるのか，そのメカニズムに関しては次のように考える．

(1) サイズパラメータ推定値が小さくなる．
(2) 平均の小さいデータが推定に与える影響が大きくなる．
(3) モデルの当てはめが平均が小さいところでの局所的なものとなり，係数の絶対値が大きく推定される．

以下ではこれらについてより詳細に考察しよう．

4.1 サイズパラメータ推定値が小さくなる

Zero-inflated 負の 2 項分布に従うデータに負の 2 項分布を当てはめたときのサイズパラメー

タ推定値に関しては，以下の定理が成り立つ．

定理 1. パラメータ $p,\mu_0,\theta_0(0<p<1,0<\mu_0,0<\theta_0)$ の zero-inflated 負の 2 項分布に従うデータにパラメータを μ,θ とする負の 2 項分布を当てはめると標本数 n が $+\infty$ に近づくにつれ，

(1) モーメント法推定値 $\hat{\theta}_{MM}$ は以下で示される $\theta^*(<\theta)$ に確率収束する．

$$\theta^* = \theta_0\left(\frac{1-p_0}{1+\theta_0 p_0}\right)$$

(2) 最尤推定値 $\hat{\theta}_{ML}$ は $\theta^\dagger<\theta_0$ であるような $\theta^\dagger$ に確率収束する．

証明は付録 C を参照のこと．この定理は zero-inflated 負の 2 項分布に対してのものだが，平均が説明変数に依存するとする zero-inflated 負の 2 項回帰モデルの場合もサイズパラメータの推定には同様な現象が起こると思われる．シミュレーションデータの例では，データ生成時の zero-inflated 負の 2 項回帰モデルにおけるサイズパラメータの値は $\theta^{ZINB}=0.6$ だが，負の 2 項回帰モデルを当てはめたときのモーメント法推定値は $\hat{\theta}^{NB}_{\mathrm{MM}}=0.305$，最尤推定値は $\hat{\theta}^{NB}_{\mathrm{ML}}=0.422$ であった．サメの混獲数データでは zero-inflated 負の 2 項回帰モデルを当てはめたときのサイズパラメータの最尤推定値は $\hat{\theta}^{ZINB}=0.568$ だったが，負の 2 項回帰モデルを当てはめたときのモーメント法推定値は $\hat{\theta}^{NB}_{\mathrm{MM}}=0.157$，最尤推定値は $\hat{\theta}^{NB}_{\mathrm{ML}}=0.318$ であった．

4.2 平均の小さいデータの影響が大きくなる

影響関数（Hampel et al., 1986）は観測値が推定に与える影響を測るもので，最尤推定量に対する影響関数はスコア関数に比例する．影響関数は絶対値が大きいほど，推測に与える影響が大きいと考えられる．

負の 2 項回帰モデルの係数に対するスコア関数は

$$\frac{\partial \log f_{NB}(\boldsymbol{\beta},\theta|y,\boldsymbol{b})}{\partial \boldsymbol{\beta}} = (y-\mu)\cdot\left(\frac{\theta}{\theta+\mu}\right)\boldsymbol{b} \quad \text{ここで} \quad \mu=\exp(\boldsymbol{b}^T\boldsymbol{\beta})$$

である．ポアソン回帰モデルに対するスコア関数はサイズパラメータ θ が $+\infty$ のときのもので，

$$\frac{\partial \log f_{Poi}(\boldsymbol{\beta},\theta|y,\boldsymbol{b})}{\partial \boldsymbol{\beta}} = (y-\mu)\boldsymbol{b}$$

で与えられる．負の 2 項回帰モデルのスコア関数にかかる $\theta/(\theta+\mu)$ は μ が増加するにつれ減少する．よって推定への影響は，平均 μ が大きいほどポアソン回帰モデルに比べて小さくなる．これはサイズパラメータ θ の値が小さいほど顕著で，μ が小さい観測値が推定に与える影響はより大きく，μ が大きい観測値の与える影響はより小さくなる．

一般化線形モデルに対して観測値の推定に与える影響を測る指標にクックの距離とテコ比（Williams, 1987; Chambers and Hastie, 1992）がある．クックの距離は各観測値の係数推定値に与える影響を測り，テコ比は各観測値の当てはめに与える影響を測る．どちらも値が大きいほど影響が大きいことを示す．図 5 は 1994 年の操業あたりのクロトガリザメ混獲数に負の 2 項回帰モデル（左列）とポアソン回帰モデル（右列）を当てはめた場合のクックの距離（上段）とテコ比（下段）を示したもので，横軸にはモデルによる当てはめ値を示した．

負の 2 項回帰モデルに関しては，当てはめ値が 15 以上の観測値に対してはクックの距離が小さいが，当てはめ値が小さい観測値にはクックの距離が目立って大きいものがいくつかあり，当てはめ値が小さい区間にクックの距離が比較的大きいものがまとまってある．一方，ポアソン回帰モデルに関しては，クックの距離が大きい観測値は特に当てはめ値が小さいものに限らず，また，当てはめ値が 30 以上の観測値に対しては当てはめ値が大きくなるほどクックの距離が大きくなる傾向がある．

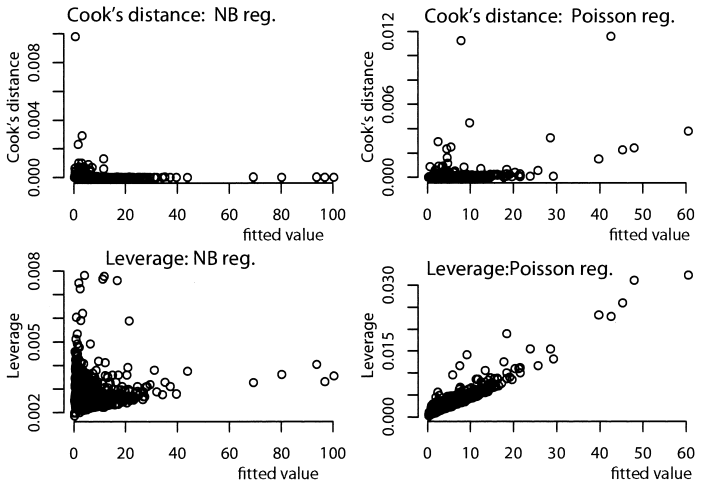

図 5. 1994 年の漁あたりのクロトガリザメ混獲数に負の 2 項回帰モデル (左列) とポアソン回帰モデル (右列) を当てはめた場合のクックの距離 (上段) とテコ比 (下段)．横軸はモデルによる当てはめ値．サイズパラメータは最尤推定値 ($\hat{\theta}_{ML} = 0.318$) を用いた．

テコ比においては，ポアソン回帰モデルと負の 2 項回帰モデルの特徴の違いがより顕著に表れた．負の 2 項回帰モデルに関しては，当てはめ値が小さいほど大きなテコ比を取ることがあり，当てはめ値が大きいものは大きなテコ比は取らない．一方，ポアソン回帰モデルでは，当てはめ値が大きいほどテコ比も大きいという相関関係がある．

図には示していないが，サイズパラメータを推定せずに大きな値を適当に与えて負の 2 項回帰モデルを当てはめるとテコ比やクックの距離はポアソン回帰モデルによる結果と似たものとなり，サイズパラメータに小さい値を与えると，当てはめ値の小さい値が推定に与える影響はより大きくなった．

4.3 局所的な当てはめと係数の過大推定

データにゼロが多く含まれており，ゼロ状態を取る確率が説明変量に依存する場合，ポアソン回帰モデルと負の 2 項回帰モデルで仮定する平均構造は真の平均構造を表していない．これは，負の 2 項回帰モデルに対数リンクを用いた場合に限ったことではなく，どのリンク関数を用いても言えることである．

データの平均構造が zero-inflated 負の 2 項回帰モデル (2.1) のもので

$$(1-p_i)\mu_i = \exp\{\boldsymbol{b}_i^T\boldsymbol{\beta}\}\left(\frac{1}{1+\exp\{\boldsymbol{g}_i^T\boldsymbol{\gamma}\}}\right) \tag{4.1}$$

$$= \exp\{\boldsymbol{b}_i^T\boldsymbol{\beta} - \boldsymbol{g}_i^T\boldsymbol{\gamma}\}\left(\frac{1}{1+\exp\{-\boldsymbol{g}_i^T\boldsymbol{\gamma}\}}\right) \tag{4.2}$$

であるとする．表現 (4.1) より，$\boldsymbol{g}_i^T\boldsymbol{\gamma}$ が負で絶対値が大きいとき，つまり，p_i が 0 に近いとき，$(1-p_i)\mu_i$ は $\exp\{\boldsymbol{b}_i^T\boldsymbol{\beta}\}$ に近くなる．また，表現 (4.2) より，$\boldsymbol{g}_i^T\boldsymbol{\gamma}$ が正で大きい値を取るとき，つまり，p_i が 1 に近いとき，$(1-p_i)\mu_i$ は $\exp\{\boldsymbol{b}_i^T\boldsymbol{\beta} - \boldsymbol{g}_i^T\boldsymbol{\gamma}\}$ に近くなる．

さて，今，簡単のために説明変数が負の 2 項回帰部分とロジスティック回帰部分で等しく $\boldsymbol{b}_i = \boldsymbol{g}_i$ であるとする．このとき，

$$\exp\{\boldsymbol{b}_i^T\boldsymbol{\beta} - \boldsymbol{g}_i^T\boldsymbol{\gamma}\} = \exp\{\boldsymbol{b}_i^T(\boldsymbol{\beta}-\boldsymbol{\gamma})\}.$$

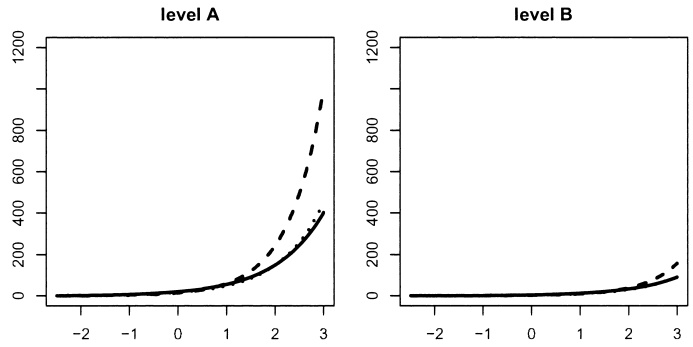

図 6. シミュレーションデータに対する回帰曲線．データを生成した真の平均(実線)，ポアソン回帰モデルによる推定回帰曲線(点線)，負の 2 項回帰モデルによる推定回帰曲線(破線)．左図は水準 A，右側は水準 B に対する回帰曲線で，横軸は説明変量 x の値を示している．

混獲データでは，ゼロ状態の確率を増加させる要因は不完全状態での混獲数を減少させ，ゼロ状態の確率を減少させる要因は不完全状態での混獲数を増加させるという推定結果を得た．このことは，$\boldsymbol{\beta}$ の成分と γ の成分が逆の符号を持つということを意味する．現実の問題において要因が同様な影響を与えると考えることは自然なことが多い．Lambert (1992) も zero-inflated ポアソン回帰モデルにおいてポアソン回帰部分とロジスティック回帰部分の係数が逆の符号を持つという制約を置いたモデルを提案している．ここで $\boldsymbol{\beta}$ と $\boldsymbol{\gamma}$ の第 j 成分 β_j と γ_j の符号が逆であるとすると絶対値の間に

$$|\beta_j - \gamma_j| \geq |\beta_j|$$

という関係が成り立つ．ゼロ状態を取る確率 p_i が 1 に近いとき，観測値の平均 $(1-p_i)\mu_i$ は $\exp\{\boldsymbol{b}_j^T(\boldsymbol{\beta}-\boldsymbol{\gamma})\}$ に近くなるから，ゼロ状態を取る確率が 1 に近い観測値が多い場合，負の 2 項回帰モデルを当てはめると係数推定値の絶対値は β_j の絶対値より大きい値を取ると考えられる．

上述のように，不完全状態(非ゼロ状態)の期待値 μ_i が小さい観測値はゼロ状態を取る確率 p_i も 1 に近い傾向にあり，つまりは，平均が小さい観測値が多くあると，係数推定値の絶対値は大きくなると考えられる．

図 6 に zero-inflated 負の 2 項回帰モデルのシミュレーションデータ y に対する回帰曲線を示した．データを生成した真の平均を実線で，ポアソン回帰モデルによる推定回帰曲線を点線で，負の 2 項回帰モデル(サイズパラメータは最尤推定値)による推定回帰曲線を破線で示している．左図は水準 A，右側は水準 B に対する回帰曲線で，横軸は説明変量 x の値を表す．水準 A に対する(推定)回帰曲線は

$$\begin{aligned}
&\text{真の回帰曲線} && \mu = \frac{1}{1+\exp\{-3x-5\}}\exp\{x+3\}\\
&\text{ポアソン回帰モデルによる推定回帰曲線} && \mu = \exp\{1.111x+2.782\}\\
&\text{負の 2 項回帰モデルによる推定回帰曲線} && \mu = \exp\{1.414x+2.670\}
\end{aligned}$$

である．負の 2 項回帰モデルは x が小さい観測値に大きな重みを与えるので x の係数推定値がポアソン回帰モデルの推定値よりも大きくなっており，x が大きいところでは，負の 2 項回帰モデルによる推定回帰曲線と真の回帰曲線との差が大きくなっている．結果として，水準 A に対する平均当てはめ値は，観測値の標本平均よりもずっと大きくなる．

5. まとめ

本論文では，ゼロの多い計数データに負の 2 項回帰モデルを当てはめると，平均の小さい観測値が推測に大きな影響を与えて回帰係数推定値の絶対値が大きくなりすぎるため，変動傾向を過大に推定してしまう，ということを実データとシミュレーションデータで示し，それがなぜ起こるのかを理論的に解明した．

このような現象は，反応変量の分布が多くのゼロと右裾が重く散らばりの大きい分布の混合分布であり，説明変量の値の増加が，非ゼロ状態のときの値の減少とゼロ状態になる確率の増加の両方に関連がある，という状況のときに起こる．例えば，反応変量が生物の個体数や行動に関連した計数データである場合には，ゼロが多く，かつ，正の値の散らばりが大きいということはよくある．また，ある要因が個体数の減少とゼロ確率の増加の両方に影響を与えるということも多くの問題において自然なことである．Lambert (1992) は zero-inflated ポアソン回帰モデルにおいて，両方の回帰部分で説明変数が等しく，その係数ベクトルの間に $\boldsymbol{\gamma} = -\tau\boldsymbol{\beta}$ という制約をおいた ZIP(τ) モデルを提案しているが，これは説明変量の反応変量に対する影響を上記のように想定し，かつ，係数の割合が一定であるというより強い条件も課すものである．

同じ平均構造を仮定しているのに，ポアソン回帰モデルと負の 2 項回帰モデルで推定結果が大きく異なるということは，仮定した平均構造が真の構造を適切に表現していないということである．そのため，推定における各観測値の重みが変わることによって推定値に違いが出ることになる．ポアソン回帰モデルを過分散の計数データに当てはめた場合，係数推定値の推定誤差を正しく推測しないため説明変量の有意性の判断を誤らせる可能性があり，また，AIC などの基準で比較すると負の 2 項回帰モデルに比べて当てはまりが大きく劣っているという結果を示すことになるが，これまでに議論したことを考えると，ゼロの多いデータに負の 2 項回帰モデルのみを当てはめて推測をすることは危険である．ポアソン回帰モデルを含めた複数のモデルで解析して結果を比較し，大きく異なっているようであれば，より適切なモデルを探索するということは重要なことと思われる．

付　録

A. Partial dependence

Partial dependence (Hastie et al., 2009) は，説明変量の反応変量に与える影響を示すための関数で，他の変量の影響を除くために他の変量に関する期待値を取ったものである．X_s を興味のある説明変量，$\boldsymbol{X}_c$ をその他の説明変量ベクトルとし，説明変量が $x_s, \boldsymbol{x}_c$ であるときの反応変量の期待値を $g(x_s, \boldsymbol{x}_c)$ で表すとする．このとき $g(X_s, \boldsymbol{X}_c)$ の説明変量 X_s に対する partial dependence は

$$g_s(x_s) = E_{\boldsymbol{X}_c}\left[g(x_s, \boldsymbol{X}_c)\right]$$

で定義され，その推定値は，他の変数のすべての観測ベクトルを $\boldsymbol{x}_{1c}, \boldsymbol{x}_{2c}, \ldots, \boldsymbol{x}_{nc}$ としたとき

$$\hat{g}_s(x_s) = \frac{1}{n}\sum_{i=1}^{n} g(x_s, \boldsymbol{x}_{ic})$$

で与えられる．

B. 自然リンクの場合の水準平均

一般化線形モデルで自然リンクを用いた場合，因子変量の各水準の当てはめ値の平均は観測

値の平均と等しくなることを示す.

指数型分布族に属する分布は確率関数/確率密度関数が

$$f(y;\eta,\phi)=\exp\{(y\eta-b(\eta))/a(\phi)+c(y,\phi)\}$$

と表せる．ここで η は自然パラメータで，平均 μ の単調関数として表すことができ，この単調関数 $g(\mu)=\eta(\mu)$ が自然リンク関数である．自然リンクを用いると，説明変量ベクトルを $\boldsymbol{x}$，係数ベクトルを $\boldsymbol{\beta}$ としたとき

$$\eta(\mu)=\boldsymbol{x}^T\boldsymbol{\beta}$$

と表せる．また，$b'(\eta)=\mu$ であるから，

$$\frac{\partial \log f(y;\eta,\phi)}{\partial \boldsymbol{\beta}}=(\boldsymbol{x}y-\boldsymbol{x}\mu)/a(\phi)$$

である.

今，観測された説明変量ベクトルと反応変量を $\boldsymbol{x}_i, y_i$ $(i=1,\ 2,\ldots,n)$ とし，係数を最尤推定法で推定したモデルによる当てはめ値を $\hat{\mu}_i (i=1,\ 2,\ldots,n)$ とすると

$$\sum_{i=1}^{n}\boldsymbol{x}_i y_i=\sum_{i=1}^{n}\boldsymbol{x}_i\hat{\mu}_i$$

が成り立つ．この式は因子変量に対しては，L_j を水準 j の標本番号の集合とすると

$$\sum_{i\in L_j} y_i=\sum_{i\in L_j}\hat{\mu}_i$$

ということであり，各水準の観測値の和は当てはめ値の和に等しいこと，つまりは，各水準の観測値の平均は当てはめ値の平均に等しいことがわかる.

C. 定理 1 の証明

(1) 標本数が大きくなるにつれモーメント法推定値は，モデルの下での平均と分散がそれぞれ真の分布の平均と分散に等しいとした式の解に収束する．負の 2 項分布 NB(μ,θ) を zero-inflated 負の 2 項分布 ZINB (p,μ_0,θ_0) に当てはめたとき，これらの式は

$$\begin{aligned}\mu &= (1-p)\mu_0\\ \mu+\frac{1}{\theta}\mu^2 &= (1-p)\mu_0+(1-p)\left(p+\frac{1}{\theta_0}\right)\mu_0^2.\end{aligned}$$

であり，これらの方程式の μ と θ に対する解 μ^* と θ^* は

$$\mu^*=(1-p)\mu_0 \quad \text{および} \quad \theta^*=\left(\frac{1-p}{1+p\theta_0}\right)\theta_0.$$

で与えられる．$(1-p)/(1+p\theta_0)<1$ であるから，$\theta^*<\theta_0$ が成り立つ.

(2) 標本数が大きくなるにつれ最尤推定値は，真の分布の下でのモデルの平均対数尤度を最大にするパラメータの値に収束する．真の分布 ZINB (p,μ_0,θ_0), $0<p<1$, $0<\mu_0$, $0<\theta$ の下でのモデル分布 NB(μ,θ) の平均対数尤度を $g(\mu,\theta)$ とし，これを最大にする μ と θ の値を $\mu^\dagger$ と $\theta^\dagger$ と記す．平均対数尤度 $g(\mu,\theta)$ は，$\theta\geq\theta_0$ に対し

$$\begin{aligned}g(\mu,\theta) &\equiv E_{ZINB(p,\mu_0,\theta_0)}[\log(f_{NB}(Y|\mu,\theta))]\\ &= (1-p)E_{NB(\mu_0,\theta_0)}[\log(f_{NB}(Y|\mu,\theta))]+p\theta[\log(\theta)-\log(\theta+\mu)]\end{aligned}$$

$$
\begin{aligned}
&= (1-p)E_{NB(\mu_0,\theta_0)}\left[\log\left(f_{NB}(Y|\mu_0,\theta)\right)\right] \\
&\quad +(1-p)E_{NB(\mu_0,\theta_0)}\left[\log\left(f_{NB}(Y|\mu,\theta)\right)-\log\left(f_{NB}(Y|\mu_0,\theta)\right)\right] \\
&\quad +p\theta\left[\log(\theta)-\log(\theta+\mu)\right]
\end{aligned}
$$

と表せる．関数 $h(\mu,\theta)$ を

$$
\begin{aligned}
h(\mu,\theta) &= (1-p)E_{NB(\mu_0,\theta_0)}\left[\log\left(f_{NB}(Y|\mu,\theta)\right)-\log\left(f_{NB}(Y|\mu_0,\theta)\right)\right] \\
&\quad +p\theta\left[\log(\theta)-\log(\theta+\mu)\right]
\end{aligned}
$$

と定義すると

$$
g(\mu,\theta) = (1-p)E_{NB(\mu_0,\theta_0)}\left[\log\left(f_{NB}(Y|\mu_0,\theta)\right)\right]+h(\mu,\theta)
$$

と表せ，

$$
\begin{aligned}
h(\mu,\theta) &= p\theta\log(\theta)+(1-p)(\theta+\mu_0)\log(\theta+\mu_0)-\theta\log(\theta+\mu) \\
&\quad -(1-p)\mu_0\log(\theta+\mu)+(1-p)\mu_0(\log\mu-\log\mu_0)
\end{aligned}
$$

となる．$g(\mu,\theta)$ の μ に関する偏微分は

$$
\frac{\partial g(\mu,\theta)}{\partial\mu}=\frac{\partial h(\mu,\theta)}{\partial\mu}=\frac{(1-p)\mu_0}{\mu}-\frac{\theta+(1-p)\mu_0}{\theta+\mu}
$$

であり，これを 0 とした方程式を解くと

$$
\mu^{\dagger}=(1-p)\mu_0
$$

が得られ，$\mu=\mu^{\dagger}$ で μ について最大化されることがわかる．次に，

$$
\theta\geq\theta_0 \text{ に対し，}\quad g(\mu^{\dagger},\theta)\leq g(\mu^{\dagger},\theta_0)\quad \text{かつ}\quad \left.\frac{\partial g(\mu^{\dagger},\theta)}{\partial\theta}\right|_{\theta=\theta_0}<0
$$

であることを示す．$h(\mu,\theta)$ に $\mu=\mu^{\dagger}$ を代入すると

$$
h(\mu^{\dagger},\theta)=p\theta\log(\theta)+(1-p)(\theta+\mu_0)\log(\theta+\mu_0)-(\theta+\mu^{\dagger})\log(\theta+\mu^{\dagger})+\mu^{\dagger}\log(1-p)
$$

と表せ，θ による偏微分は，Jensen の不等式を用いると

$$
\begin{aligned}
\frac{\partial h(\mu^{\dagger},\theta)}{\partial\theta} &= p\log(\theta)+(1-p)\log(\theta+\mu_0)-\log(\theta+\mu^{\dagger})\log(p\theta+(1-p)(\theta+\mu_0)) \\
&< \log(p\theta+(1-p)(\theta+\mu_0))-\log(\theta+\mu^{\dagger}) \\
&= \log(\theta+\mu^{\dagger})-\log(\theta+\mu^{\dagger}) \\
&= 0
\end{aligned}
$$

となる．これより $\partial h(\mu^{\dagger},\theta)/\partial\theta$ は常に負であり，$h(\mu^{\dagger},\theta)$ は狭義の単調減少関数であることがわかる．$E_{NB(\mu_0,\theta_0)}\left[\log\left(f_{NB}(Y|\mu_0,\theta)\right)\right]$ は $\theta=\theta_0$ において最大値を取るので，$\theta=\theta_0$ における偏微分は 0 となる．よって，

$$
g(\mu^{\dagger},\theta)\leq g(\mu^{\dagger},\theta_0)\quad \text{for}\quad \theta\geq\theta_0\quad \text{かつ}\quad \left.\frac{\partial g(\mu^{\dagger},\theta)}{\partial\theta}\right|_{\theta=\theta_0}=\left.\frac{\partial h(\mu^{\dagger},\theta)}{\partial\theta}\right|_{\theta=\theta_0}<0.
$$

また，$g(\mu^{\dagger},\theta)$ は $(0,\infty)$ で θ に関して連続な関数であり，$\lim_{\theta\to 0+}g(\mu^{\dagger},\theta)=-\infty$ であるから，$g(\mu^{\dagger},\theta)$ を最大にする θ の値 $\theta^{\dagger}$ は $(0,\ \theta_0)$ に存在する．

謝 辞

原稿を注意深くお読み頂き適切なコメントを頂いたことに対して，2 名の査読者の方に感謝申し上げます．本研究は JSPS 科研費 23500356 の援助を受けている．

参 考 文 献

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*, Wadsworth, Pacific Grove.

Frankenberg, E. and Thomas, D. (2000). *The Indonesian Family Life Survey: Study Design and Results from Waves 1 and 2*, RAND, Santa Monica.

Greene, W. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models, Working Paper 94–10, Department of Econometrics, Stern School of Business, New York University, New York.

Hampel, F. R., Ronchetti, E. M., Rousseuw, P. J. and Stahel, W. A. (1986). *Robust Statistics — The Approach Based on Influence Functions*, Wiley, New York.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, 2nd ed., Springer, New York.

Hilbe, J. M. (2007). *Negative Binomial Regression*, Cambridge University Press, Cambridge.

Lam, K. F., Xue, H. and Cheung, Y. B. (2006). Semiparametric analysis of zero-inflated count data, *Biometrics*, **62**, 996–1003.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, **34**, 1–14.

Lawless, J. (1987). Negative binomial and mixed Poisson regression, *The Canadian Journal of Statistics*, **15**, 209–225.

Maunder, M. and Punt, A. E. (2004). Standardizing catch and effort data: A review of recent approaches, *Fisheries Research*, **70**, 141–159.

Minami, M., Lennert-Cody, C. E., Gao, W. and Roman-Verdesoto, M. H. (2007). Modeling shark bycatch : The zero-inflated negative binomial regression model with smoothing, *Fisheries Research*, **84**, 210–221.

Mullahy, J. (1986). Specification and testing of some modified count data models, *Journal of Econometrics*, **33**, 341–365.

Román-Verdesoto, M. and Orozco-Zöller, M. (2005). Bycatches of sharks in the tuna purse-seine fishery of the eastern Pacific Ocean reported by observers of the Inter-American Tropical Tuna Commission, 1993–2004, Data Report 11, Inter-American Tropical Tuna Commission, La Jolla, California.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica*, **57**, 307–333.

Welsh, A., Cunningham, R., Donnelly, C. and Lindenmayer, D. (1996). Modelling the abundance of rare species: statistical models for counts with extra zeros, *Ecological Modelling*, **88**, 297–308.

Williams, D. A. (1987). Generalized linear model diagnostic using the deviance and single case deletions, *Applied Statistics*, **36** (2) , 181–191.

Proceedings of the Institute of Statistical Mathematics Vol. 61, No. 2, 271–287 (2013) 287

Analysis of Data with Many Zero-valued Observations: Over-estimation of Temporal Trend by Negative Binomial Regression

Mihoko Minami[1] and Cleridy E. Lennert-Cody[2]

[1]Department of Mathematics, Keio University
[2]Inter-American Tropical Tuna Commission

In ecological and environmental studies, count data such as the number of animals per unit area or unit effort often contain many zero-valued observations. Such data unfortunately may be analyzed without any special consideration given to how the zeros arose. In particular, the negative binomial regression model has been a commonly used model for count data with overdispersion. However, we found that the negative binomial regression model over-estimated temporal trends in species relative abundance. Such over-estimation could be problematic, for example, for the development of management guidelines for conservation.

In this paper, we investigate this phenomena of over-estimation. We show that when the negative binomial regression model is fitted to data with excess zeros, the estimate of the size parameter becomes too small and the observations with small fitted values have more influence. This results in estimated coefficients of predictors that are too large in absolute value, and it produces exaggerated estimates of the marginal effects.

Key words: Over-dispersion, influence function, size parameter, zero-inflated negative binomial regression model, Cook's distance, leverage.