ミニワークショップ
「CPUE 標準化における問題点の整理と解決手法の検討」

日時：2014 年 3 月 20 日（木）　13：30～17:00
会場：国際水産資源研究所　会議室富士

開催趣旨：
今年度国際水研では平成 25 年度所内シーズ研課題として、各担当者による 5 漁業 6 魚種の CPUE 標準化手法についてレビューを行ってきた。本ワークショップではこれまでに取り上げられた課題について、解決策の事例研究紹介や外部専門家を交えた討論を行い、CPUE 標準化手法の高度化を目指す。

プログラム：
13:40~14:40
CPUE 解析における統計的諸問題－特にゼロ・キャッチデータの取り扱いについて－
庄野　宏（鹿児島大学）

14:45~15:30
Random Forest Model による気仙沼の船頭ノートを用いたターゲット操業推定
金岩　稔（東京農業大学）・平岡優子（国際水産資源研究所）

15:45~16:30
はえ縄オペレーショナルデータにおける CPUE 標準化事例紹介と課題整理
井嶋　浩貴・清藤　秀理（国際水産資源研究所）

16:30~17:00
総合討論

コメンテーター：平松一彦（東京大学大気海洋研究所）・北門利英（東京海洋大学）

独立行政法人水産総合研究センター　国際水産資源研究所　平成 25 年度シーズ研究課題
ミニワークショップ「CPUE 標準化における問題点の整理と解決手法の検討」2014/3/20
講演資料（著者に無断での複製・配布はご遠慮下さい）鹿児島大学　水産学部　庄野 宏

CPUE 解析における統計的諸問題

―特にゼロ・キャッチ・データの取り扱いについて―

庄野 宏
（鹿児島大学水産学部）

はじめに

長いタイトルで恐縮だが、本稿ではゼロ・データを含む非負応答変数と説明要因間の因果推論に利用される統計モデルについて、CPUE 解析を例に取ってご紹介したい。所々に数式が出てくるが、必ずしも本質的ではないため、読み飛ばしていただいても構いません。

問題の背景

例えば、CPUE（catch per unit effort:単位努力当たり漁獲量）の時空間変動や表面水温などの海洋環境要因が CPUE に与える影響を調べる際に、以下のモデル

$$\log(\mathrm{CPUE})=(\text{切片})+(\text{年})+(\text{季節})+(\text{海区})+(\text{水温})+(\text{交互作用})+\mathrm{error},\ \mathrm{error}\sim N(0,\sigma^2) \quad (1)$$

（CPUE の自然対数に対して正規分布を当てはめる(i.e. CPUE が対数正規分布に従うと考えても良い)、すなわち観測誤差が互いに独立で同一の正規分布に従う回帰モデル）を考えることが多い。これは、説明要因が全て質的変数の場合は分散分析に、全て連続変量の場合には回帰分析であり、質的変数と連続変量を両方含む時には共分散分析となる。このとき、CPUE に 0 の値が含まれていると

$$\log(\mathrm{CPUE})=-\infty \quad (2)$$

となるため計算が出来なくなり、何らかの回避策が必要である。

例えば、まぐろはえ縄漁業におけるサメ類の混獲など非漁獲対象種ではゼロ・データの割合が非常に高く、恒常的にこの問題が生じる。

CPUE のようにゼロを含む非負応答に対する因果推論、影響を与える説明要因と応答変数の因果関係を推測する問題は、水産のみならず様々な分野で現れる。

例 1.降水量の予測：雨が降らない日の降水量は 0 となるため、雨が降るか否か、降る場合にはその量を合わせて予測したい。

例 2.銀行など金融機関の与信管理：顧客にお金を貸すべきか否か、貸しても良さそうな場合にはいくらまでにすべきか、という限度額を合わせて算出したい。

以下このようなゼロを含むデータの問題に関する対処法について見ていきたい。

Ad hoc method

全ての応答変数に正の微少量 k を足し込む方法である。

log(CPUE+k)=(切片)+(年)+(季節)+(海区)+…+(交互作用)+error, error～N(0, σ^2)(3)

この方法は扱いやすいが、点推定値・区間推定値の両方とも偏りが生じる。

計数回帰モデル（Reed, 1996）

応答変数が離散の場合に、CPUE 解析ならば漁獲尾数など離散変数に直してからゼロを含む離散確率分布（Poisson, 負の二項分布等）を当てはめる方法である。

E[Catch]=Effort*exp{(切片)+(年)+…+(交互作用)}, Catch～Po(λ) or NB(a, b)　(4)

（ただし E は期待値を、Po, NB はそれぞれ Poisson 分布, 負の二項分布を表す）

このモデルは実用上便利であるが、ゼロ・データの割合が非常に高い場合や fat tail と呼ばれる裾が広い分布の場合に当てはまりが悪いことが知られている。

注）CPUE 解析の場合には応答変数が Catch へ変更になっているが、降水量予測や与信管理の例など離散値を取る場合には、応答変数の変更を必要としない。

Delta 型 2 段階法（Lo *et al.*, 1992）

最初にゼロ・データの割合をロジスティック回帰など*により推定し、次に非ゼロ・データに式(1)の共分散分析や式(4)の計数回帰モデルを適用し、非ゼロ・データの割合と非ゼロ部分の応答変数（CPUE）の値を掛け合わせる方法であり、1st step は次のように表される。（2nd step は式(1)もしくは式(4)が適用可能である）

E[X]=p, p～Bin(θ), X=1 (if catch=0) or 0 (otherwise)

logit(p)=log(p/(1-p))=(切片)+(年)+(季節)+(海区)+(水温)+(交互作用)　(5)

（*-代わりにプロビット回帰や complementary log-log 回帰等も使用可能である）

Zero-inflated 計数回帰モデル（Lambert, 1992）

具体的な定式化は省略するが、イメージ的には Delta 型 2 段階法で 2nd step に計数回帰モデルを採用し、1st step の尤度と繋ぎ合わせて同時に推定する方法と言える（厳密に言えばそうでない部分もあるのだが、詳しい議論は割愛したい）。

Tweedie 回帰モデル（Tweedie, 1984; Jorgensen, 1997; Shono, 2008）

Tweedie 分布はゼロに mass point を持つ絶対連続な確率分布で、各イベント X が Gamma 分布に従い、イベントの起こる回数 N が Poisson 分布に従う確率過程（複合 Poisson 過程）として表される。定式化は式(6)の通りだが、平均分散関係を表す冪係数 p の存在が特徴的であり、CPUE 解析では Y が CPUE に対応する。

$$Y=\sum_{i=1}^{N}X_i,\qquad N\sim Poisson\left(\frac{\mu^{2-p}}{\sigma^2(2-p)}\right)\qquad (0<p<1\text{を除く})$$
$$X_1,...,X_n(i.i.d.)\sim Gamma\left(\sigma^2(2-p)\mu^{p-1},\frac{p-1}{2-p}\left[\sigma^2(2-p)\mu^{p-1}\right]^p\right) \tag{6}$$

ただし $E[Y]=\mu, Var[Y]=\sigma^2\mu^p$ であり、冪係数 p が 1, 2, 3 のときにそれぞれ Poisson 分布、Gamma 分布、逆正規分布を表す（*i.i.d.*は互いに独立同一分布の意味）。

私が過去に CPUE データに基づいて cross validation を行ったところ、ゼロ・データの割合が極端に高い場合（2/3 以上が目安）には Tweedie 回帰モデルや Delta 型 2 段階法が、ある程度高い場合（1/3 以上 2/3 未満が目安）には Catch を応答とした計数回帰モデル（特に負の二項分布）の性能が高くなった（Shono, 2008）。

その他の方法（正規分布の利用など）

原点に戻って考えてみると、log(CPUE)に正規分布（i.e. CPUE に対数正規分布）を当てはめようとするためにゼロ・データの問題が起こってしまった。CPUE に対して正規分布を当てはめたら良いのでは、と考えることは自然な発想である。

ただし、CPUE 解析の場合は、観測誤差が CPUE の大きさにかかわらず一定という正規分布の仮定よりは、観測誤差が CPUE の大きさに比例するという対数正規分布の条件設定の方が生物学的にマッチしていると考えられるため、伝統的に log(CPUE)に正規分布（CPUE に対数正規分布）を当てはめるモデルが広く使われている。さらに正規分布を当てはめるとマイナスの値をどうするかという問題も生じる（CPUE は非負の値しかとらない）（式(7)の切断正規分布を参照）。

なお、両者は応答変数が等しいと考えても良いため、AIC（Akaike's information Criterion:赤池情報量規準）などの情報量規準により比較することが可能である。

最後に、上記の考え方に関係する幾つかのモデルについて簡単に紹介したい。

切断正規回帰モデル

CPUE に正規分布を当てはめる回帰モデルに対して、負の値を取らないようにするためには、正規分布を切断して規準化する切断正規分布を使用すれば良い。

$$g(x)=\left\{\begin{array}{l}\dfrac{f(x)}{\int_{-\infty}^{a}f(x)dx}\,(x<a)\\ 0\ (otherwise)\end{array}\right\} \tag{7}$$

注） $f(x)$ が元の正規密度関数を表す

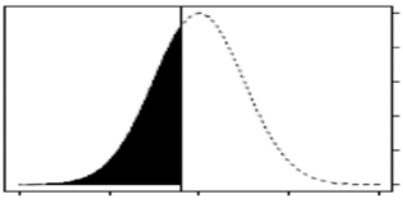

図 1. 切断正規分布（左側を削除）

Hurdle モデル

切断分布を利用しており、Zero-truncated Poisson 回帰モデルと名付けられる。

$$\Pr[Y=y]=\pi_0\begin{Bmatrix}\pi_0(y=0)\\(1-\pi_0)\dfrac{1}{(1-e^{-\lambda})}\dfrac{e^{-\lambda}\lambda^{y}}{y!}(y=1,2,...)\end{Bmatrix} \quad (8)$$

ハードルを越えると正の整数を取るため Hurdle（or Two-part）モデルと呼ばれる。

Tobit モデル

考え方は Delta 型 2 段階法と似ている。ゼロ・データの割合を Probit 回帰モデルで推定し、非ゼロ部分に切断正規回帰モデルを当てはめる方法で、計量経済学で頻出している。CPUE がゼロとなるデータはマイナスを表す左側で打ち切りの Zero-censored 回帰モデルのため、このように概念がそぐわないこともある。

おわりに

以上、ゼロ・データを含む非負応答変数と説明要因間の因果推論に用いられる統計モデルを駆け足で紹介してきたが、これらの手法は CPUE 標準化のみならず様々な問題や経済問題、自然現象の解析に適用可能である。

引用文献

Jorgensen, B. (1997) *The theory of dispersion models.* Chapman and Hall, London, 237 pp.

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

Lo, N. C. L. D., L. D. Jacobson and J. L. Squire (1992) Indices of relative abundance from fish spotter data based on Delta-Lognormal models. *Canadian Journal of Fish and Aquatic Science*. **49**, 2515-2526.

Reed, W. J. (1996) Analyzing catch-effort data allowing for randomness in the catching process. *Canadian Journal of Fish and Aquatic Science*, **43,** 174-186.

Shono, H. (2008) Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fisheries Research*, **93**(1-2), 154-162.

Tweedie M.C.K. (1984) An index which distinguishes between some important exponential families. In Statistics: Applications and New Directions. *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference.* (Eds. J. K. Ghosh and J. Roy), Calcutta: Indian Statistical Institute, 579-604.

遠洋リサーチ&トピックス Vol.7. 2010 年 3 月 p.20-21
（ISSN 1880-9103）から引用（ページレイアウトのみ変更）

統計一口メモ　第 6 回: 古くて新しい縮小推定のおはなし

本稿では、共分散分析(回帰分析・分散分析を含む)における変数選択に焦点を当てて、縮小推定と呼ばれる近年発展している包括的な変数選択手法についてご紹介したい。この方法は、従来広く用いられていた AIC などの情報量規準やF検定、カイ二乗検定に代表されるstepwise検定とは異なり、二乗誤差の式にペナルティ項を付け加えることによって、ある意味機械的に変数選択(モデル選択)を行う手法である。

$$y_i = \theta_0 + \theta_1 x_{i1} + \ldots + \theta_p x_{ip} + \varepsilon_i = \sum_{j=0}^{p} \theta_j x_{ij} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$$

where i=1,…,n, j=0,…,p, n:標本数, p:パラメータ数, $x_{i0} = 1$,

の形で表される共分散分析モデルにおいて、多重共線性および頑健性(便宜的な呼称)と呼ばれる問題が知られている。

多重共線性というのは、説明変数間に強い相関がある場合に起こる現象で、上の共分散分析モデルを行列表示すると y=Xθ+ε (y:応答変数,X:観測変数,θ:未知母数,ε:誤差)となり

$\min_\theta \|y - X\theta\|_{l_2}^2$ の解は $\hat{\theta} = (X^t X)^{-1} X^t y$　(X^{-1} : Xの逆行列, X^t : Xの転置行列)　where $\|\theta\|_{l_2} = \left(\sum_{j=1}^{p} \theta_j^2 \right)^{1/2}$, $\|\theta\|_{l_1} = \sum_{j=1}^{p} |\theta_j|$,

とexactに解けるが、X^tXが正則にならないゆえに逆行列が不定になる。そこで考案されたのが Ridge 回帰(Hoerl and Kennard, 1970)という方法である。ここではあるλ>0 に対して

$\min_\theta \{ \|y - X\theta\|_{l_2}^2 + \lambda \|\theta\|_{l_2}^2 \}$ の最適化を考えることによって、

$\hat{\theta} = (X^t X + \lambda I)^{-1} X^t y$ (I: 単位行列) における行列の退化を防いでいる(単位行列のλ倍(λI)が加わっているため)。

次に頑健性の問題である。これは外れ値に対する影響が大きいという現象にも見て取れるが、端的に言えば汎化誤差に対する予測性能が悪い、という点に尽きる。データを学習用と検証用にランダム分割した場合、学習用データにてモデルを構築し、検証用データを利用して個々の観測データとそれに対応するモデルからの推定値(予測値)の当てはまりの良さ(二乗誤差)を測定するのだが、共分散分析モデルにおいては、一般にこの汎化誤差の予測性能が良くないことが知られている。この問題は変数選択とも密接に関係しており、変数の数を増やしていけばいくほど学習用データに対する当てはまりは良くなるが(過学習)、検証用データに対する当てはまりは悪くなっていく。そこで、汎化誤差に対する予測性能向上を目的として考案されたのが、LASSO(least absolute shrinkage and selection operator: Tibshirani, 1996)と呼ばれる推定量であり、あるλ>0 に対して $\min_\theta \{ \|y - X\theta\|_{l_2}^2 + \lambda \|\theta\|_{l_1} \}$ の最適化を行う。この推定量は、形の上では Ridge 回帰における L_2 ノルムのペナルティ項を L_1 ノルムに変更しただけだが、その挙動は全く異なり、ペナルティ項の効果ゆえに有意でない多くのパラメータをゼロと推定し、スパースな解が得られる。

オリジナル論文では $\min_\theta \|y - X\theta\|_{l_2}^2$, subject to $\|\theta\|_{l_1} \le t$ と定義されているが、全てのλに対して t≧0 であるような t が必ず1つ存在し、双方とも同じ解を与える(Osbourne, *et.al*, 2000)。

ただ、このLASSOは縮小し過ぎてしまう傾向を持つことから、L_1 ノルムと L_2 ノルムの線形結合の形のペナルティを持つ Elastic Net (Zou and Hastie,2005)が新たに提案された(次式)

$$\min_\theta \{ \|y - X\theta\|_{l_2}^2 + \lambda [\alpha \|\theta\|_{l_1} + (1-\alpha) \|\theta\|_{l_2}^2] \} (\lambda > 0, 0 < \alpha < 1)$$

なお、これらの縮小推定においてチューニングパラメータ λ の値は cross-validation により決定することが一般的であるが、AIC や BIC などの情報量規準を利用することも可能である。

これまでざっと駆け足で説明してきた縮小推定法であるが、その他にもLASSOにおけるλの値をトレースするための有効なアルゴリズム(LARS: Efron, *et.al.*, 2004)など数多くの斬新な手法が考案されている(Hastie *et.al*,2009)。LASSO に代表される縮小推定手法は、DNA 解析など実社会でも多く直面する NP 問題と呼ばれる標本数よりパラメータ数が多い場合の推定手法とも密接な関わり合いを持ち、今日の数理統計学におけるホットな研究テーマの 1 つになっているが、今回は紙面の制約もあり割愛させていただきたい。個人的にはこれらブレークスルーの殆どが Stanford 大学統計科学部の研究グループによって行われたことに対し、驚きを禁じ得ない。

引用文献

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. 2004: Least angle regression (with discussion), *The Annals of Statistics*, **32**(2), 407-499

Hastie, T., Tibshirani, R. and Friedman, J. 2009: *The elements of statistical learning*, 2nd edition, Springer, 745pp.

Hoerl, A. E. and Kennard, R. 1970: Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12**: 55-67

Osbourne, M., Presnell, B. and Turlach, N. 2000: On the lasso and its dual, *J. of Computational and Graphical Statistics*, **9**: 319-337

Tibshirani, R. 1996: Regression shrinkage and selection via the Lasso, *J. of the Royal Statist. Soc. B*, **58**(1): 267-288

Zou, H. and Hastie, T. 2005: Regularization and variable selection via the elastic net, *J. of the Royal Statist. Soc. B*, **67**(2): 302-320

（遠洋水産研究所 数理解析研究室 庄野 宏）

Random Forest Model による気仙沼の船頭ノートを用いたターゲット操業推定

金岩稔（東京農業大学）・平岡優子（国際水研）

緒言：水産資源解析において、資源量の変化を相対的に表す指標として単位努力量あたり漁獲量（CPUE）はよく使われるが、CPUE をそのまま資源量指数として用いるためには、漁獲効率が一定である必要がある。しかしながら漁獲効率は時空間的、漁業的な要因で変化する。そのため、CPUE を漁獲効率が変化する要因で標準化した標準化 CPUE が資源量指数として用いられることが多い(Maunder & Punt 2004)。商業漁業においてある種を狙い（ターゲット）操業であるか非狙い操業であるかは、漁獲効率に差を生み、その結果として CPUE は変化する。そのため、狙い操業であるかを無視した CPUE の比較は資源量の変化を見誤る可能性がある（Quirijns et al. 2008）。しかしながら、多くの漁業において何を狙って操業をしているかの情報は少ない。

太平洋におけるヨシキリザメ（BSH）の漁獲は、気仙沼船団による延縄漁業が大半を占める。国際資源研究所では、2004 年 7 月から気仙沼通信協会に所属するいくつかの協力的な船に依頼して、各操業の狙い魚種を含めた詳細な操業情報の収集を行っている。この船頭ノートデータを用いて、BSH 資源量指数を推定するために使用した延縄操業データ（漁獲成績報告書データ）で狙っている魚種、特に BSH を狙っているか？を推定できれば、狙い効果を CPUE 標準化に利用できるであろう。そこで、今回船頭ノートデータを用い BSH 狙いの有無を推定するモデルを構築し、漁獲成績報告書データの BSH 狙いの有無を推定して、資源量指数標準化に利用することを試みた。

材料と方法：2004 年から 2007 年の船頭ノートの操業データ、1994 年から 2010 年の漁獲成績報告書データを用いた。Random Forest Model (RFM)はデータマイニング手法の一種であり、データと説明要因の不確実性を考慮した機械学習アルゴリズムの一種である（Breiman 2001）。第一段階として、RFM を船頭ノートデータに使用して BSH 狙いの有無を説明するモデルを構築した。RFM は、BSH 狙いの有無を考慮できる要因全てを取り入れた複雑モデルと、漁積から得られるうる要因のうち、資源量指数の標準化に使わない要因のみを取り入れた単純モデルを用いて説明し、両モデルのパフォーマンスを評価する。その後パフォーマンスに問題がなければ、第二段階として漁積データの狙い要因を推定し、その項を用いた資源量指数の標準化を行う。最後にその結果を現在 ISC で使われている標準化資源量指数と比較を行う。

結果と考察:複雑モデルも単純モデルも 9 割以上の正答率で BSH 狙いの有無を推定し、モデルの評価に問題はなかった。標準化資源量指数は ISC で現在使われているものとほぼ同様のトレンドを示した。これは、今回の手法の妥当性を示すとともに、ISC で使われているモデルでも、狙い効果がうまく標準化されていることも示唆される。しかしながら、船頭ノートの狙いの有無データには年によるばらつきが多く、その記録の妥当性に疑問が残っている。今後、その妥当性を高めた上で、実際の資源解析への応用を検討したい。

はえ縄オペレーショナルデータにおける CPUE 標準化事例紹介と課題整理

井島浩貴・清藤秀理（国際水研）

日本のはえ縄漁業の CPUE は、資源評価において最も重要なデータの一つである。しかし、CPUE の標準化の手法に関しては、ゼロキャッチやターゲッティングなど様々な問題が指摘されており、それらの問題は魚種によって様々である。このような問題点を少しでも解決するため、我々は、はえ縄のオペレーショナルデータを用いて、北西部太平洋のビンナガと、東部太平洋のメバチの CPUE 標準化を行った。ビンナガの解析では、漁獲サイズ、漁場の変動、漁具の変動、ゼロキャッチに着目した。メバチの解析では、漁具の経年変化と漁船ごとの影響に着目した。ビンナガの解析結果では、漁獲サイズごとの CPUE を算出することに成功したが、漁場や漁具の変動に関してはうまく説明することができなかった。メバチの解析では、漁船の効果を 1 隻ごとに説明変数として取り扱う場合、データ量の問題から最後まで計算することができなかった。したがって何らかのデータスクリーニングをする必要があると考えられる。例えば、メバチをターゲッティングしている船を抽出して標準化することなどが解決策として期待される。