遠洋水研報　第37号　1-8頁　平成12年3月
Bull. Nat. Res. Inst. Far Seas Fish., No. 37, 1-8, March 2000

情報量規準とステップワイズ検定の比較と水産資源解析への応用

庄野　宏*

Comparison between Information Criteria and Stepwise Test, and Their Application to the Analysis of Fishery Resources

Hiroshi SHONO*

A comparison between information criteria and stepwise test in variable selection using nested models was made from the theoretical point of view. Two statistical standards (AIC and likelihood ratio test) were checked under simple assumptions. It was found that chi-square test shows a tendency to select the simpler model with fewer parameters than AIC when the difference of the number of parameters in two corresponding models is rather small. Although the result of stepwise test may be inconsistent in case that the pass of test is not unique, it will be able to calculate the selection performance using both standards (AIC and stepwise test) by computer simulation in common cases.

Key words: stepwise test, AIC, nested model, CPUE standardization

はじめに

水産資源解析におけるモデル選択基準として，AIC(Akaike, 1973)をはじめとする情報量規準が広く使用されてきた．情報量規準AICを実際の水産資源解析における問題に適用した論文として，再生産曲線のパラメーター推定を行ったHiramatsu *et al.*(1994)の例や，イセエビの漁獲データを用いて漁具能率や環境要因等を含む様々な資源パラメーターを推定したYamakawa *et al.* (1994)らの例がある．しかし，候補となるモデルに包含関係が成り立っている場合，すなわち階層構造を持つモデルの場合には，カイ二乗検定やF検定に代表されるステップワイズ検定を用いることも可能であり，両者で結果が異なることもある．例えば，Matsumiya(1990)は標識放流におけるJolly-Seber法を用いて，AICによるモデル選択の結果が尤度比検定によるそれと異なることを示している．本論文では，包含構造が満たされているモデルを取り上げて，理論的な立場から情報量規準とステップワイズ検定の比較を行うことを目的とする．

理論的考察

1．問題設定

まず，次のような2つのモデルを仮定する．但し，qとp，nはそれぞれパラメーター数，標本数を表し，$q<p<n$とする．また，tは行列の転置を意味する．

Model-0：$\Theta_0=(\theta_1\cdots\theta_q)^t$

Model-1：$\Theta_1=(\theta_1\cdots\theta_q\cdots\theta_p)^t$

このとき$\Theta_0\subset\Theta_1$となっていることから，Model-0とModel-1の間に包含関係があり，こういったケースを便宜上Model-0⊂Model-1と表すことにする．ここではModel-0(単純)とModel-1(複雑)のモデルが2つの場合を考える．

例1．回帰分析モデル

Model-A：$Y=a+bU+e$　　$e\sim N(0,\sigma^2)$

Model-B：$Y=a+bU+cV+e$

(U，V：説明変数ベクトル，Y：応答変数ベクトル，e：誤差ベクトル，a, b, c：回帰係数)

このモデルでは，$\Theta_A=(a,b)^t$, $\Theta_B=(a,b,c)^t$と表されることから$\Theta_A\subset\Theta_B$であり，Model-A⊂Model-Bという包含関係が成り立っている．

2．ステップワイズ検定

標本ベクトルYが前節のModel-0，Model-1のいずれかに従っているとする．すなわち，$Y\sim P(\Theta_0)$ or $Y\sim P(\Theta_1)$と仮定する．そして，帰無仮説と対立仮説をそれぞれH_0, H_1とする下の検定

$$\left\{\begin{array}{l}H_0:\Theta=\Theta_0\,(\text{Model-0})\\H_1:\Theta=\Theta_1\,(\text{Model-1})\end{array}\right\}\qquad(2.1)$$

を考える．このとき，

2000年1月31日受理　遠洋水産研究所業績　第401号
Accepted on January 31, 2000. Contribution No. 401 of the National Research Institute of Far Seas Fisheries
*遠洋水産研究所(National Research Institute of Far Seas Fisheries, 5-7-1, Orido, Shimizu, Shizuoka 424-8633, Japan)

H_0が真 $\Rightarrow$ $\Delta D \approx \chi^2_{p-q}$（$\Delta D$ は漸近的に自由度 p-q のカイ二乗分布に従っている），

H_1が真 $\Rightarrow$ $\Delta D \approx$ 非心$\chi^2 (> \chi^2_{p-q})$

と結論付けられる．上式の不等号>は，非心カイ二乗分布の上側100 α %点値が自由度 p-q のカイ二乗分布のそれよりも大きくなることを表している．ただし，

$l(\Theta) = \mathrm{Log}L(\Theta)$: 対数尤度関数，$D_i$:Model-$i$ (i=0, 1)のDeviance（逸脱度），$\Delta D = D_0 - D_1$とおく．ΔD のカイ二乗分布への近似は漸近的なものであるが，誤差構造として正規分布を仮定した場合にはexactに成立する．また，

$$\Delta D = D_0 - D_1 = -2[l(\Theta_0) - l(\Theta_1)]$$

と表されることから，このカイ二乗検定では対数尤度比の(－2)倍と有意水準 α でのカイ二乗分布のパーセント点$\chi^2_{p-q}(\alpha)$との比較を行っていることになる．この検定における暗黙の了解として，H_1が正しい(i.e. $D_1 \approx \chi^2_{n-p}$)ことを仮定していることに注意する必要がある．なぜなら，この仮定の下でH_0が真(i.e.$D_0 \approx \chi^2_{n-q}$)であるときに初めて

$\Delta D = D_0 - D_1 \approx \chi^2_{p-q}$（カイ二乗分布の性質より）

が成り立つからである．逆に，H_1が正しくないとすると $D_1 \approx \chi^2_{n-p}$ とならずに非心カイ二乗分布に従ってしまうために$\Delta D \approx \chi^2_{p-q}$ は成立し得ない．この検定の論理は上の通りであるが，最終的に対立仮説H_1が正しくないとすると，間違ったモデルを選択してしまうことになる．なお，Devianceについては，McCullagh and Nelder(1989)等のGLM(generalized linear model，一般化線形モデル)のテキストで詳しく述べられている．

3．情報量規準AIC

情報量規準AICは表現のシンプルさと解釈の容易さから，水産資源分野におけるモデル選択の基準として広く用いられている．AICから派生したものも含めて様々な情報量規準が提案されているが，本論文ではAICのみを取り扱うことにする(AICの導出については付録参照のこと)．今回のモデルにおいては

$$\begin{aligned} \mathrm{AIC(Model\text{-}0)} &= -2\,l(\Theta_0) + 2q \\ \mathrm{AIC(Model\text{-}1)} &= -2\,l(\Theta_1) + 2p \qquad (2.2) \\ \Delta\mathrm{AIC} &= \mathrm{AIC(Model\text{-}0)} - \mathrm{AIC(Model\text{-}1)} \\ &= -2[\,l(\Theta_0) - l(\Theta_1)\,] - 2(p-q) \end{aligned}$$

と表されることから，情報量規準AICによるモデル選択では対数尤度比の(-2)倍とパラメーター数の差の2倍を比較していることになる．

4．AICとカイ二乗検定の比較

これまでの計算結果から，カイ二乗検定では有意水準を α とするとき，

$-2[l(\Theta_0) - l(\Theta_1)] < (\mathrm{or} >) \chi^2_{p-q}(\alpha)$ $\Rightarrow$ Model-0(or Model-1)

を選択し，AICでは

$-2[l(\Theta_0) - l(\Theta_1)] < (\mathrm{or} >) 2(p-q)$ $\Rightarrow$ Model-0(or Model-1)

を選択していることになる．この結果を図示するとFig. 1のようになる．

この図から，$p-q$(モデルにおけるパラメーター数の差)が小さい(8 ないし16未満)ときにはカイ二乗検定の方が，$p-q$が大きいときには情報量規準AICの方が，パラメーター数の少ない単純なモデルを選ぶ傾向にある．

このような理論的な考察によってモデル選択の傾向が明らかになったが，パフォーマンスの良さを調べるためには何らかのコンピューターシミュレーションが必要である．具体的には真のモデルから乱数を発生させて(真のモデルを含む)複数の候補からなるモデルの中でどのモデルが選択されるかを調べれば良い．

例2．正規誤差モデル

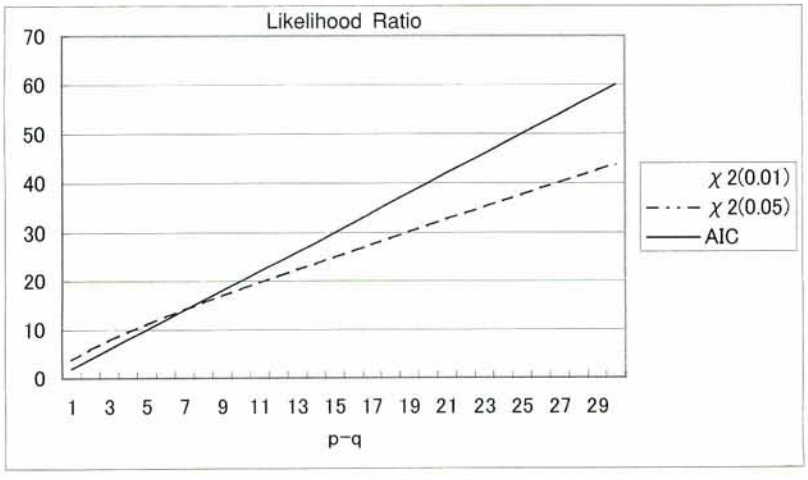

Fig. 1. Likelihood ratio plotted against the difference of the number of parameters. The graphs show the comparison with AIC and Chi-square tests in hierarchical model.

$Y_i(i=1,\cdots,n)$はそれぞれ平均μ_i，分散σ^2を持つ正規分布N(μ_i,σ^2)に従う互いに独立なn個の標本とする．このときのDevianceは残差平方和RSSを用いて

$$D=\frac{1}{\sigma^2}\mathrm{RSS} \tag{2.3}$$

と表される．ただ，このケースでのDevianceは未知母数σ^2を含むため何らかの形で推定しなければならず，通常は最尤推定量

$$\hat{\sigma}^2=\frac{1}{n}\mathrm{RSS}$$

もしくは不偏性を持つようにバイアス修正を行った最尤推定量

$$\hat{\sigma}_*^2=\frac{1}{n-p}\mathrm{RSS}$$

を用いる．しかし，下のF統計量を用いることにより局外母数σ^2の問題を解決することが出来る．何故なら，

$$F=\frac{\frac{D_0-D_1}{p-q}}{\frac{D_1}{n-p}}=\frac{\frac{\mathrm{RSS}_0-\mathrm{RSS}_1}{(p-q)\sigma^2}}{\frac{\mathrm{RSS}_1}{(n-p)\sigma^2}}=\frac{\frac{\mathrm{RSS}_0-\mathrm{RSS}_1}{p-q}}{\frac{\mathrm{RSS}_1}{n-p}} \tag{2.4}$$

となりσ^2が分子分母で打ち消しあうからである．

この統計量を用いたF検定

$$\left.\begin{aligned}H_0:\Theta=\Theta_0(\text{Model-0})\\ H_1:\Theta=\Theta_1(\text{Model-1})\end{aligned}\right\} \tag{2.5}$$

を考えると，

H_0が真 ⇒ $F\sim F_{p-q,n-p}$（Fはexactに自由度$(p-q,\ n-p)$のF分布に従っている），

H_1が真 ⇒ $F\sim$非心$F(>F_{p-q,n-p})$

となる．ただ，この検定においてもH_1が正しい（i.e. $D_1\sim\chi^2_{n-p}$）ことを暗黙に仮定していることに注意する必要がある．また，

ΔAIC=AIC(Model-0)—AIC(Model-1)

$$=n\log\left[1+\frac{p-q}{n-p}F\right]-2(p-q)$$

と表されることから，

ΔAIC=0 ⇔

$$F=\frac{p-q}{n-p}\left[\exp\left\{\frac{2}{n}(p-q)\right\}-1\right]\to 2\quad(as\ \ n\to\infty) \tag{2.6}$$

となり，AICにおけるモデル選択はF検定において$F<2$ならばModel-0を，F>2ならばModel-1を選択することに相当している．すなわち，AICによるモデル選択基準ではF統計量の値が2より大きいか小さいかで判断していることになるので，このことから漸近的な場合ではあるがAICとF検定との間の関係が認められる．

実際の計算例—CPUE標準化を例として—

本節ではステップワイズ検定におけるパス(検定の順序)について考える．水産資源解析への適用例としてCPUE標準化を取り上げる．生のCPUEデータから季節変化や漁船の能力など，資源密度以外の要因の影響を取り除く作業をCPUE標準化といい，資源の年変動の効果を取り出すことが主な目的である．また，手法としてはGLMが用いられることが多い．なお，CPUE標準化の考え方・方法論については平松(1995)に詳しく述べられている．

CPUE標準化の一例として，下のようなモデル

$\log(\mathrm{CPUE}_{ij})=$

$$\mathrm{Intercept}+\mathrm{Year}_i+\mathrm{Class}_j+(\mathrm{Year}^*\mathrm{Class})_{ij}+\mathrm{Error}_{ij} \tag{3.1}$$

Intercept: 切片

Year: 年効果

Class: 船効果

(Year*Class): 年と船の交互作用効果

Error: 誤差

を考えると，誤差項を正規分布と仮定した場合には多元配置分散分析と同一であり，(3.1)では2元配置になっている．実際には，仮定した主効果及び交互作用効果をモデルに含めるか否かを統計的な方法を用いて判断するのが一般的であるが，このモデル(3.1)は階層構造になっているために，ステップワイズ検定と情報量規準AICの両方とも使用可能である．

次に，ステップワイズ検定におけるパスが一意に決まらない例として，Hilborn and Walters(1992)によるCPUE標準化の仮想例(Table 1参照)を取り上げる．

Table 1. Virtual data for CPUE standardization (loosely based on the data. Hilborn and Walters, 1992). The values show catch rate (tons per hour) for three classes of vessel in four different years.

Year	Class-1	Class-2	Class-3
1	0.60	1.03	1.22
2	0.48	0.56	1.26
3	0.33	0.67	0.89
4	0.54	0.48	1.01

候補となる下の4つのモデル(Model-1)～(Model-4)

Model-1: log(CPUE) = Intercept + Error

Model-2: log(CPUE) = Intercept + Year + Error

Model-3: log(CPUE) = Intercept + Class + Error (3.2)

Model-4: log(CPUE) = Intercept + Year + Class + Error

（但しError～N$(0,\sigma^2)$とおく）

を想定すると，モデル間の包含関係は下のようになる．

(Model-1) ⊂ (Model-2)
∩ ∩ (3.3)
(Model-3) ⊂ (Model-4)

ステップワイズ検定においてはBackwardに変数を減らしていく方向で考えるのが普通なので，このケースでは

Path1：(Model-4) → (Model-2) → (Model-1)
Path2：(Model-4) → (Model-3) → (Model-1)

という２つのパスが考えられる．

少し詳しく見ていくと，１番目のパス(Path 1)では最初のステップ(Step 1)で

$$\left\{\begin{array}{l} H_0:\Theta=(\text{Intercept},\text{Year},\sigma^2)\,(\text{Model-2}) \\ H_1:\Theta=(\text{Intercept},\text{Year},\text{Class},\sigma^2)\,(\text{Model-4}) \end{array}\right\}$$

という検定を考え，帰無仮説H_0が棄却(i.e.対立仮説H_1が採択)されればModel-4を選択し，H_0が採択されれば次のステップ(Step 2)に進む．Step2では

$$\left\{\begin{array}{l} H_0:\Theta=(\text{Intercept},\sigma^2)\,(\text{Model-1}) \\ H_1:\Theta=(\text{Intercept},\text{Year},\sigma^2)\,(\text{Model-2}) \end{array}\right\}$$

という検定を考え，H_0が棄却(i.e.H_1が採択)されればModel-2を選択し，H_0が採択されればModel-1を選択する．

一方，２番目のパス(Path 2)では最初のステップ(Step 1)で

$$\left\{\begin{array}{l} H_0:\Theta=(\text{Intercept},\text{Class},\sigma^2)\,(\text{Model-3}) \\ H_1:\Theta=(\text{Intercept},\text{Year},\text{Class},\sigma^2)\,(\text{Model-4}) \end{array}\right\}$$

という検定を考え，帰無仮説H_0が棄却(i.e.対立仮説H_1が採択)されればModel-4を選択し，H_0が採択されれば次のステップ(Step 2)に進む．Step 2では

$$\left\{\begin{array}{l} H_0:\Theta=(\text{Intercept},\sigma^2)\,(\text{Model-1}) \\ H_1:\Theta=(\text{Intercept},\text{Class},\sigma^2)\,(\text{Model-3}) \end{array}\right\}$$

という検定を考え，H_0が棄却(i.e.H_1が採択)されればModel-3を選択し，H_0が採択されればModel-1を選択する．

このケースでは，２通りのパスによる検定において異なるモデルが選ばれることも考えられ，そのような場合の解釈は極めて難しい．それぞれの計算過程をTable 2に示すが，現にTable 1のデータから候補となる４つのモデル(Model-1)～(Model-4)における残差平方和を求めて，有意水準１%としてステップワイズF検定を適用すると，Path 1ではModel-1が選択されるのに対しPath 2ではModel-3が選択されてしまい，異なった結果を得る．また，このデータにAICを適用するとModel-4が選択される（Table 3参照）．

このようなパスの一意性がステップワイズ検定における大きな問題点になっている．しかし，AIC等の情報量規準を用いることによってこの問題は解決される．何故なら，解釈上の難解さは残るもののAICの値が一番小さくなるものをベストなモデルとして選択すれば良いからである．従って，検定のパスが一意に決まらない場合には，階層的なモデルであっても情報量規準を用いてモデル選択すべきであると考えられる．

論　議

前節の内容とこれまでに得られている知見からステップワイズ検定と情報量規準AICの問題点を列挙すると，次のようになる．

１．ステップワイズ検定の問題点

(1)パス(検定の順序)が一意に決まらない場合に矛盾が生じる可能性がある．

(2)真でないモデルを一時的に真と仮定してしまうことにより，正しいモデルが選択されないケースも起こりうる．

(1)は候補となるモデルの数が多い場合に生じる問題であり，詳しい説明と水産資源分野での適用例(CPUE標準化)については，前節で取り上げた．

(2)は検定において対立仮説を真であると暗黙に仮定していることを指しており，先に説明した通りである．

２．情報量規準AICの問題点

(1)計算量が多い上に対応していないソフトウェアも多く，実際の計算が厄介な場合が多い．

(2)平均二乗予測誤差最小の観点から導出しているため，AICを用いて選択されたモデルが真のモデルかどうか，という点において疑問が残る．

(3)AICの差が１，２程度の場合には有意でないと言わ

Table 2. Results of stepwise F tests on Table1 data. F-stat., F(0.01) and DF show the F statistics, the upper one percentile and degree of freedom of F distribution, respectively.

Pass	Step	H0	H1	F-stat.	F(0.01)	DF	Result
1	1	Model-2	Model-4	12.568	13.274	F(2,5)	Accept
1	2	Model-1	Model-2	0.475	8.451	F(3,7)	Accept
2	1	Model-3	Model-4	2.045	12.060	F(3,5)	Accept
2	2	Model-1	Model-3	9.031	8.649	F(2,8)	Reject

Table 3. Results of the model selection using AIC on Table 1 data.

Model	AIC
1	16.445
2	20.222
3	6.273
4	2.666

れているが(坂元ら, 1983), こういった場合にどのように解釈すれば良いのか？

(4)小標本あるいは大標本の場合にバイアスが生じることが多い.

(1)は技術的な問題であり本質的ではない.

(2)は導出のプロセスに関する問題である. AICとともに良く知られているBIC (Schwarz, 1978)では事後確率最大化の観点から導出しているためにダイレクトにモデルのパフォーマンスの良さを表しているのに対し, AICは平均二乗予測誤差最小の観点から導出しているために, 真のモデル, すなわち真の自由パラメーター数を推定するための規準といえるかどうか微妙である.

(3)はAIC等の情報量規準では第1種の過誤を検出することが不可能であることを示唆している. しかし, 現実問題としては1-2程度の差しかないようなケースでもAICを小さくするモデルを選ばざるを得ないのではないだろうか？AICの差の有意性を検出するためにモデル選択検定(Linhart, 1988)や多重比較の手法の利用(Shimodaira, 1998)なども提案されているが, あまりにも数学的な議論になってしまうため, ここではこれ以上詳しく触れない. また, Bootstrap法等の計算機シミュレーションを用いたAICの差の検定も, 理論的には構築可能である.

(4)はAICの導出原理に関わる問題である. AICの導出過程において漸近理論 (標本が無限にあると仮定した上で展開される理論) を用いているために, 小標本の場合に偏りが生じてしまう. また, 大標本の場合にはペナルティ項、つまり自由パラメーター数の2倍の効果が少なくなり, AICの値が最大対数尤度のそれに近づいてしまうため, パラメーター数の多い複雑なモデルを選ぶ傾向にある. 要するに, AICは一致性を持たないためにバイアスが生じてしまうことになる. これらのバイアスを修正するために様々な情報量規準が提案されており, 小標本の場合のc-AIC (Sugiura, 1978)や大標本の場合のHQ (Hannan and Quinn, 1979)などが代表的である. しかし, これらの情報量規準については別途詳しく議論する予定である.

まとめと今後の課題

階層的なモデルにおける選択基準としてステップワイズ検定と情報量規準AICのどちらを用いるべきか？という問いに対して, 検定のパスが一意に決まらない場合には矛盾を避ける意味においてもAICを用いるべき, と結論付けられる. しかし, 検定のパスが一意に決められる場合には, シミュレーションなどの結果を参考にしてケースに応じて両方の基準を使い分けるべきである, と思われる. 複数モデルを等ウェイトで扱う場合にはAICを, 想定されるモデルをベースにして変数の取捨選択を考える場合には検定を用いるべき, という考え方は一般的であるが, シミュレーションのパフォーマンスのみで判断することも可能である.

次報においては, 複数の仮定(検定のパスが一意である場合と一意でない場合等), 複数のモデル(回帰分析モデル, 分散分析モデル, 時系列モデル等)におけるコンピューターシミュレーションを行って, そのセレクションパフォーマンスを報告する. また, 検定の有意水準αは制御可能なパラメーターとも考えられるため, αを幾つに設定すべきか？という問題も生じる. 一般には有意水準として5％あるいは1％が用いられているが, このような固定観念にとらわれることなく, 適切な有意水準の値についても更に検討していきたい.

謝　辞

原稿を精査し多くの有益なコメントを下さった遠洋水産研究所の平松一彦博士と所内の査読者の方々に深く御礼申し上げる. また, 検定手法に関する技術的なアドバイスをいただいた東京大学農学部の岸野洋久博士と東京水産大学の北門利英氏に深く感謝申し上げる.

なお, 本論文は東京大学海洋研究所共同利用シンポジウム「生態学および水産学における統計的方法の新しい展開」(平成11年12月8日)で報告した内容の理論的な部分をベースに執筆したものである.

最後に、今回の内容を文章として残すよう強く薦めて下さった東京大学海洋研究所の松宮義晴博士に厚く御礼申し上げる。

文 献

Akaike, H. 1973: Information theory and an extention of the maximum likelihood principle. *2nd International Symposium on Information Theory.* Petrov, B.N., and Csaki, F. (eds.), Akadimiai Kiado, Budapest, p. 267-281.

Hannan, E.J., and B.G. Quinn. 1979: The determination of the order of autoregression. *J. Royal Statist. Soc. Ser. B*, **41**: 190-195.

Hilborn, R., and C.L. Walters. 1992: *Quantitative fisheries stock assessment.* Chapman and Hall, 570 p.

Hiramatsu, K., Y. Matsumiya, and S. Kitada. 1994: Introduction of suitable stock-recruitment relationship by a comparison of statistical models. *Fish. Sci.*, **60**: 411-414.

平松一彦. 1995: 統計モデルによるCPUE標準化. 漁業資源研究会議北日本底魚部会報, **28**: 87-97.

Linhart, H. 1988: A test whether two AIC's differ significantly. *South African Statist. J.*, **22**: 153-161.

Matsumiya, Y. 1990: AIC introduced to parsimonious modeling of capture-mark-recapture studies. *Nippon Suisan Gakkaishi*, **56**: 839.

McCullagh, P., and J.A. Nelder. 1989: *Generalized linear models.* 2nd ed. Chapman and Hall, 511 p.

坂元慶行・石黒真木夫・北川源四郎. 1983: 情報量統計学, 共立出版, 236 p.

Schwarz, G. 1978: Estimating the dimension of a model. *Ann. Statis.*, **6**: 461-464.

Shimodaira, H. 1998: An application of multiple comparison techniques to model selection. *Ann. Inst. Statist. Math.*, **50**: 1-13.

Sugiura, N. 1978: Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist. -Theor. Meth.*, **7**: 13-26.

Yamakawa, T., Y. Matsumiya, M. Nishimura, and S. Ohnishi. 1994: Expanded DeLury's method with variable catchability and its application to catch-effort data from spiny lobster gillnet fishery. *Fish. Sc.*, **60**: 59-63.

付録 情報量規準AICの導出について

AICの導出については，坂元ら(1983)を初めいくつかの文献で紹介されているが，いずれもかなり簡略化して書かれているため，数学ないし物理系出身でない一般の読者が内容を追っていくことは極めて厄介である．そこで，本付録では導出過程での式変形を省略せず，完全な形での証明を与える．

まず，標本ベクトルとその実現値，パラメーターベクトルをそれぞれ

$$Y=(Y_1,\cdots,Y_n),\ y=(y_1,\cdots,y_n),\ \Theta=(\theta_1,\cdots,\theta_n)\ (n>p)$$

とし，標本ベクトルYの従う分布の密度関数を$f(y|\Theta)$とおく．また，標本ベクトルYの要素 が互いに独立であることを仮定しておくと，

$$Y_i\sim p.d.f.f(y_i|\Theta)\,(i=1,\cdots,n),\ f(y|\Theta)=\prod_{i=1}^{n}f(y_i|\Theta)$$

と書き表せる．このとき，未知パラメーターΘに関するモデルMを考え，MLE(maximum likelihood estimator, 最尤推定量)，尤度関数，対数尤度をそれぞれ$\hat{\Theta}=\hat{\Theta}(Y), L(\Theta|Y), l(\Theta|Y)$とおくと，このモデルMに対するAICは

$$\mathrm{AIC(M)}=-2l(\hat{\Theta}|y)+2p$$

と表される．ここではこの式の導出について考える．

次に，ベクトルZを(Yと独立であるような)Yの未来の観測変量，Θ^*を真のパラメーター(i.e.平均対数尤度$E_z[l(\Theta|Z)]=\int l(\Theta|z)f(z|\Theta)dz$を最大にするパラメーター)とすると，Zの真の分布と予測分布の近さを測るための１つの基準であるKullback-Leibler情報量の期待値は

$$E_y\left[\int f(z|\Theta^*)\log\frac{f(z|\Theta^*)}{f(z|\hat{\Theta})}dz\right] \qquad \text{(A1)}$$
$$=\int f(z|\Theta^*)\log f(z|\Theta^*)dz-E_y\left[\int f(z|\Theta^*)\log f(z|\hat{\Theta})dz\right]$$

と書き下せる．以下に示すように，AIC最小のモデルを選択することは(A1)を最小にするような$f(z|\hat{\Theta})$を求めることに相当している．(A1)式の右辺第１項は$f(z|\hat{\Theta})$と無関係であるため，(A1)式の最小化はすなわち(A1)式の右辺第２項の最大化になり，

((A1)式の右辺第２項)=
$$E_y\left\{E_z[\log f(Z|\hat{\Theta})]\right\}=E_y\left\{E_z[l(\hat{\Theta})|Z]\right\}$$
($\because l(\Theta|Z)=\log L(\Theta|Z)=\log f(Z|\Theta)$)

である．ここで，密度関数モデル$f(z|\Theta)$はΘを適当に選ぶ(i.e. $\Theta=\Theta^*$とする)ことによって真の密度関数モデルg(z)が得られる場合を想定する．すなわち，

$f(z|\Theta^*)=g(z)$ を満たす真のパラメーター Θ^*と真の密度関数g(・)が存在する場合について考える．

次に，$E_z[\log f(Z|\hat{\Theta})]$を真のパラメーター$\Theta^*$のまわりで２次の項までTaylor展開すると

$$E_z[\log f(Z|\hat{\Theta})] = \int\{\log f(z|\hat{\Theta})\}f(z|\Theta^*)dz$$
$$\approx \int\{\log f(z|\Theta^*)\}f(z|\Theta^*)dz + (\hat{\Theta}-\Theta^*)\left[\int\{\frac{\partial}{\partial\Theta}\log f(z|\Theta^*)\}f(z|\Theta^*)dz\right]$$
$$+\frac{1}{2}(\hat{\Theta}-\Theta^*)^t\left[\int\{\frac{\partial^2}{\partial\Theta\partial\Theta^t}\log f(z|\Theta^*)\}f(z|\Theta^*)dz\right](\hat{\Theta}-\Theta^*)$$
$$= \int\{\log f(z|\Theta^*)\}g(z)dz + (\hat{\Theta}-\Theta^*)\left[\int\{\frac{\partial}{\partial\Theta}\log f(z|\Theta^*)\}g(z)dz\right]$$
$$+\frac{1}{2}(\hat{\Theta}-\Theta^*)^t\left[\int\{\frac{\partial^2}{\partial\Theta\partial\Theta^t}\log f(z|\Theta^*)\}g(z)dz\right](\hat{\Theta}-\Theta^*)$$
$$= E_z[\log f(Z|\Theta^*)] + (\hat{\Theta}-\Theta^*)E_z\left[\frac{\partial}{\partial\Theta}\log f(z|\Theta^*)\right]$$
$$+\frac{1}{2}(\hat{\Theta}-\Theta^*)^t E_z\left[\frac{\partial^2}{\partial\Theta\partial\Theta^t}\log f(z|\Theta^*)\right](\hat{\Theta}-\Theta^*)$$

となる．但し，記号tは行列やベクトルの転置を表す．ここで，$E_z[\log f(Z|\Theta)]$ は $\Theta=\Theta^*$ で最大値をとることから

$$\frac{\partial}{\partial\Theta}E_z[\log f(Z|\Theta^*)] = E_z\left[\frac{\partial}{\partial\Theta}\log f(Z|\Theta^*)\right] = 0$$

となる．なお，このような微分と積分の交換条件は，あらかじめ正則条件として仮定しておく．また，

$$J(\Theta^*) = -E_z\left[\frac{\partial^2}{\partial\Theta\partial\Theta^t}\log f(Z|\Theta^*)\right],$$
$$I(\Theta^*) = E_z\left[\frac{\partial}{\partial\Theta}\log f(Z|\Theta^*)\frac{\partial}{\partial\Theta^t}\log f(Z|\Theta^*)\right] \text{ (Fisher情報行列)}$$

とおくと

$$E_z[\log f(Z|\hat{\Theta})] \approx E_z[\log f(Z|\Theta^*)] - \frac{1}{2}(\hat{\Theta}-\Theta^*)^t J(\Theta^*)(\hat{\Theta}-\Theta^*) \quad (A2)$$

と書ける．ここで$g(z)=f(z|\Theta^*)$ のとき$I(\Theta^*)=J(\Theta^*)$であることと

$$\sqrt{n}(\hat{\Theta}-\Theta^*) \to N(0, I^{-1}(\Theta^*)) = N(0, J^{-1}(\Theta^*)) \ (\text{as } n\to\infty),$$
$$\sqrt{n}(\hat{\Theta}-\Theta^*)^t J(\Theta^*)\sqrt{n}(\hat{\Theta}-\Theta^*) \to \chi_p^2 \ (\text{as } n\to\infty) \quad (A3)$$

より

$$E_y[\sqrt{n}(\hat{\Theta}-\Theta^*)^t J(\Theta^*)\sqrt{n}(\hat{\Theta}-\Theta^*)] = p,$$
$$E_y[(\hat{\Theta}-\Theta^*)^t J(\Theta^*)(\hat{\Theta}-\Theta^*)] = p/n \quad (A4)$$

となる．ここで(A2)の両辺でYに対する期待値を考えると

$$E_y\{E_z[\log f(Z|\Theta^*)]\} \approx E_y\{E_z[\log f(Z|\hat{\Theta})]\} - \frac{p}{2n} = E_z[\log f(Z|\hat{\Theta})] - \frac{p}{2n} \quad (A5)$$

$$E_z[\log f(Z|\hat{\Theta})] = \left\{E_z[\log f(Z|\hat{\Theta})] - \frac{1}{n}\sum_{i=1}^n \log f(Y_i|\Theta^*)\right\}$$
$$+\left\{\frac{1}{n}\sum_{i=1}^n \log f(Y_i|\Theta^*) - \frac{1}{n}\sum_{i=1}^n \log f(Y_i|\hat{\Theta})\right\} + \frac{1}{n}\sum_{i=1}^n \log f(Y_i|\hat{\Theta}) \quad (A6)$$

と変形出来る．ここで

$$\frac{1}{n}\sum_{i=1}^n \log f(Y_i|\Theta^*) \approx \frac{1}{n}\sum_{i=1}^n \log f(Y_i|\hat{\Theta}) + \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\Theta}\log f(Y_i|\hat{\Theta})(\Theta^*-\hat{\Theta})$$
$$+\frac{1}{2}(\hat{\Theta}-\Theta^*)^t\left\{\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\Theta\partial\Theta^t}\log f(Y_i|\hat{\Theta})\right\}(\hat{\Theta}-\Theta^*)$$

であり，

$$\sum_{i=1}^n \frac{\partial}{\partial\Theta}\log f(Y_i|\hat{\Theta}) = 0 \quad (\because)\ \hat{\Theta}\text{が}\Theta\text{のMLE}),$$
$$\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\Theta\partial\Theta^t}\log f(Y_i|\hat{\Theta}) \to -J(\Theta^*) \ (\text{as } n\to\infty)$$
$$(\because)\ n\to\infty\text{のとき}\hat{\Theta}\to\Theta^*(\hat{\Theta}\text{の一致性}))$$

を用いると

$$\frac{1}{n}\sum_{i=1}^n \log f(Y_i|\Theta^*) - \frac{1}{n}\sum_{i=1}^n \log f(Y_i|\hat{\Theta}) \approx -\frac{1}{2}(\hat{\Theta}-\Theta^*)^t J(\Theta^*)(\hat{\Theta}-\Theta^*) \quad (A7)$$

となる．(A7)を(A6)に代入すると，

$$E_z[\log f(Z|\hat{\Theta})] = \left\{E_z[\log f(Z|\hat{\Theta})] - \frac{1}{n}\sum_{i=1}^n \log f(Y_i|\Theta^*)\right\}$$
$$-\frac{1}{2}(\hat{\Theta}-\Theta^*)^t J(\Theta^*)(\hat{\Theta}-\Theta^*) + \frac{1}{n}\sum_{i=1}^n \log f(Y_i|\hat{\Theta}) \quad (A8)$$

となり，

$$E_y\left\{E_z[\log f(Z|\hat{\Theta})] - \frac{1}{n}\sum_{i=1}^n \log f(Y_i|\Theta^*)\right\}$$
$$= E_z[\log f(Z|\hat{\Theta})] - \frac{1}{n}E_y\left[\sum_{i=1}^n \log f(Y_i|\Theta^*)\right] = 0$$

に注意して(A8)のYに対する期待値をとると下の(A9)のようになる．

$$E_z[\log f(Z|\hat{\Theta})] = E_y\{E_z[\log f(Z|\hat{\Theta})]\}$$
$$\approx -\frac{1}{2}E_y[(\hat{\Theta}-\Theta^*)^t J(\Theta^*)(\hat{\Theta}-\Theta^*)] + \frac{1}{n}E_y\left[\sum_{i=1}^n \log f(Y_i|\hat{\Theta})\right]$$
$$\approx -\frac{p}{2n} + \frac{1}{n}E_y\left[\sum_{i=1}^n \log f(Y_i|\hat{\Theta})\right] \quad (A9)$$

よって，(A5)と(A9)より

$$E_y\{E_z[\log f(Z|\hat{\Theta})]\} \approx \frac{1}{n}E_y\left[\sum_{i=1}^n \log f(Y_i|\hat{\Theta})\right] - \frac{p}{n}$$

となり，両辺n倍すると

$$nE_y\{E_z[\log f(Z|\hat{\Theta})]\} \approx E_y\left[\sum_{i=1}^n \log f(Y_i|\hat{\Theta})\right] - p \quad (A10)$$

と表される．(A10)式の左辺をEMLL(expected mean log-likelihood，期待平均対数尤度)と呼び，

$$\text{EMLL} = E_y\{nE_z[\log f(Z|\hat{\Theta})]\} = nE_y\{E_z[\log f(Z|\hat{\Theta})]\}$$

における括弧内の値

$$E_z[\log f(Z|\hat{\Theta})] = \int\{\log f(z|\hat{\Theta})\}f(z|\Theta^*)dz$$

が大きいほど良いモデルになっている．このEMLLは個々の標本の実現値に依存せず，データの真の分布とデータ数，およびモデルのみで決定される．また，(A10)式は $\sum_{i=1}^n \log f(Y_i|\hat{\Theta}) - p$ がEMLLの近似的な不偏推定量になっていることを意味する．

これまで見てきたように，AICの導出は確率変数(標本)を用いて期待値ベースで行ったが，実際の計算はデータ(標本の実現値)によって行われる．具体的には，与え

られたデータ $y=(y_1,\cdots,y_n)$ に対する最大対数尤度

$$l(\hat{\Theta}\,|\,y)=\sum_{i=1}^{n}\log f(y_i\,|\,\hat{\Theta})$$

（定義は $l(\hat{\Theta}\,|\,y)=\max_{\Theta\subset\Omega} l(\Theta\,|\,y)$ で与えられる）

を考えてEMLLの漸近的な不偏推定量 $l(\hat{\Theta}\,|\,Y)-p$ を（−2）倍したものが情報量規準AICであり，

AIC=

$$-2l(\hat{\Theta}\,|\,y)+2p=-2\times(\text{MLL})+2\times(\text{自由パラメーター数}) \quad \text{(A11)}$$

と定義される．そして，AICの値が小さくなるほど，真の分布と候補となる予測分布とのKullback-Leibler の距離が小さくなるため，良いモデルであると考えられる．

このように，AICの導出にはいくつかの近似が用いられている．小標本の場合に問題が起こるのは，主に(A2)で最尤推定量の漸近正規性を用いてカイ二乗近似を行った部分であり，c-AIC(correction of AIC)ではこの部分に工夫を凝らして，exactにカイ二乗分布に従う統計量を使用している．また，大標本の場合に偏りが生じるのは主に(A4)で期待値を用いた評価を行っているためと考えられる．HQでは重複対数の法則により

$$\left|\sqrt{n}(\hat{\Theta}-\Theta^*)^t J(\Theta^*)\sqrt{n}(\hat{\Theta}-\Theta^*)\right|<p\log\log n$$

とexactに評価することによって一致性を持つように修正している．