特集にあたって

松嶋　敏泰

本特集のテーマである統計モデル選択についてあまり親しみのない読者の方も多いと伺ったので，モデル選択とはどんな問題か簡単な例題で説明してみたい．

その日の最高気温xを用いて，あるビアガーデンの来店者数yを予測したいというようなニーズがあったとしよう．このように思いついたのは，過去の経験からxとyとの間には何らかの関係があるのではと考えたからであろう．そして，予測を行うための第1歩として，過去のn対のデータ$(x_i, y_i), i = 1, 2, \cdots, n$を用いて，$x$と$y$の関係を説明する数式が導き出せないか考えてみたい．

xとyの関係を説明する式としてすぐ思い浮かぶのは，以下のような線形回帰式で、両者の関係を記述することではなかろうか．

$$y = a_0 + a_1 x + e.$$

ここで、eは平均0，分散σ^2(未知)の正規分布に従う確率変数である．

最小2乗法で，この式の3つのパラメータa_0, a_1, σ^2をデータから推定することによって，未知の係数がなくなりxとyの関係式は一意に定まることになる．この求まった関係式を用いて，来店者数の予測などの様々な意思決定を行うことが可能となる．

この関係式はxとyが線形関係にあることを前提としたモデル上で求められたものである．ところが，両者の関係は次のようなl次の多項式(l次の多項式回帰モデル)でしか表すことができないかもしれない．

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_l x^l + e.$$

さらに，この多項式でもxとyの関係はうまく表現できず，指数関数やその他のもっと複雑な関数でしか表すことができないかもしれない．

以上のようにどのモデルを用いて解析を行うかは，この問題に限らず統計解析において最初に決めておかなければならない問題である．一般的には，その現象が起る物理的メカニズム等，固有技術や経験的知識からモデルが主観的に決定されることが多い．

それに対して，データのみによって客観的にモデルを決定しようとする場合もある．この問題が本特集のテーマとなっているモデル選択問題である．データからのみモデルを決定するといっても，まったく何も無いところからモデルをつくり出すわけではなく，考慮に入れるモデルのクラスはあらかじめ決定されており，その中からモデルを選択することになる．

例えば，この例題ではxとyの関係を表すモデルのクラスとして，多項式回帰モデルのクラスを仮定しているとしよう．この多項式回帰モデルクラスの中からあるl次の回帰モデルを選択(同時にパラメータも推定)することは，一つの典型的なモデル選択問題である．

l次の回帰モデルを用いることが決まってしまえば，そのパラメータ$a_0, a_1, \cdots, a_l$は次式で示す誤差Sを最小化する最小2乗法によって推定され，回帰式はパラメータを含めて完全に決定される．

$$S = \sum_i (y_i - a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_l x_i^l)^2.$$

この誤差Sは，x_iが与えられたもとでのyのモデル上での平均値と実現値の2乗誤差を表している．これはモデルのデータに対する適合度の一つの基準と考えられる．

最小2乗法によって決定されたl次の回帰式はl次の回帰モデルの中では，誤差Sの評価基準のもとで最適なモデルとなっている．またこれは，以下の(対数)尤度を最大化するという基準においても最適なモデルとなっている．

$$L = \log P((x, y)^n | a^l, \sigma, m_l).$$

多項式回帰モデルのクラスの中の各モデル間の比較にも，この適合度の基準をそのまま用い，誤差Sを最小化するモデルを選択すればれば，モデル選択の問題は解決しそうに思える．しかし問題はそれほど簡単

ではない.

なぜなら，ある回帰モデルはそれより低次のモデルを含んでいるので，次数lが高くなればモデルの自由度は増し，与えられたデータに対する誤差Sは必ず小さくなる．1次の回帰モデルが直線，2次のモデルが2次曲線とモデルの次数が高くなるにつれて次数の高い曲線を表現できることとなり，データに対する適合度が良くなることは直感的にも明らかであろう．極端な場合を考えると，$n-1$次の回帰モデルを使えば$S=0$の回帰式をつくれることになるので，$n-1$次の回帰モデルが最適なモデルとして選択されることとなる．

本当にこの最高次のモデルがいいモデルなのであろうか．与えられたデータだけに過度に適合したモデルは，未知のデータの予測に利用できるのだろうか．$n-1$次の回帰モデルの回帰モデルはn個の係数パラメータを用いて表される．これでは与えられたn個のデータをそのまま記録しておくのと記憶容量的に何ら変わらず，本来，データの特徴を抽象化して表すためにあるモデルが，データそのものを残しておくのと全く変わらないことになってしまう．

もう一度，モデルの良さの評価基準とは何かということを考え直さなければならなくなってしまったようだ．そこで少し視点を変えて，もう少し一般的なモデルの選択問題を考えてみたい．

ここまで考えてきたモデルは，統計モデルであったが，工学，社会科学など様々な分野では様々なモデルが用いられている．数値データから統計モデルを推論する統計モデル選択問題を一般化すると，事例からその事例全体の特徴を表す法則を導き出す問題ととらえられ，このような問題は帰納推論（学習）問題と呼ばれている．

例えば，ある言語の文章の例からその言語の文法を導き出す問題は代表的な帰納推論の問題であり，データは例文に対応し，モデルは文法によって表現されていることになる．また，ニュートンはリンゴの落ちる事例を観察することで，万有引力の法則（ニュートン力学モデル）を帰納推論したわけである．

宇宙の力学モデルとして，その後相対性理論をはじめ様々なモデルが提案されているわけであるが，アインシュタイン，ホーキング等多くの天才物理学者たちは，宇宙のモデルはシンプルであるはずだという信念のもとにモデルを構築している．シンプルなモデルほどいいモデルであるという考え方は，古くから哲学者の間で述べられていたことで，14世紀のスコラ学者の名にちなんでオッカムの剃刀あるいはケチの原理などとして知られている．

先ほどまでの回帰モデルの選択では，モデルの良さをデータに対する適合度のみで評価していた．この基準のみでなくモデルの単純さという基準も加えてモデルの良さを測ることも必要なのではあるまいか．しかし，この2つの評価基準は一般に両立しないことは，回帰モデルの選択問題を考えれれば明らかであろう．次数の低いモデルはモデルは単純であるが適合度は低くなる．逆に次数の高いモデルは複雑であるが適合度は高くなる．

一般にこの2つの基準はトレードオフの関係にあり，どちらの基準に対しても最適なモデルは存在しない．そこで，両基準を合わせた総合的な基準を用いることが考えられる．例えば，ある程度適合度があり，それほど複雑でない中庸なモデルを良いモデルと考えるわけである．

この2つの評価基準を総合したモデル選択の基準を明瞭な式で初めて表したものが，赤池により1970年代に提案されたAIC(Akaike Information Criterion)である．AICは次式で定義され，この基準を最小化するモデルを選択することが提唱されている．

$$AIC = (\text{対数尤度}) + (\text{モデルのパラメータ数}).$$

右辺第1項はデータに対する適合度，第2項はモデルの単純さを表していると解釈できる．例えば，多項式回帰モデルの場合のAICは以下のように表される．

$$AIC = \log P((x,y)^n|a^l,\sigma,m_l) + (l+2).$$

AICは真の分布とモデルとの距離をKullback-Leibler情報量で測った場合，最も近いモデルを良いモデルとするという仮定の下に漸近不偏推定量として導き出されたものであるが，得られた式はモデルの適合度と単純性という2つの視点からの総合的基準と解釈することが可能である．

2つの視点からのトレードオフ関係を見事に表現したAIC基準は，その後提案された多くのモデル選択基準に強い影響を与えた．モデル選択基準としてBIC(Bayes Information Criterion)やMDL(Minimun Discription Length)等様々な視点から数多くの提案がなされているが，どれもモデルの適合度と単純性という2つの視点からの解釈が可能である．

これらのモデル選択基準の応用範囲は非常に広い．

例であげた多項式回帰モデルの選択と類似の問題としては，自己回帰モデルの次数を決定する問題や，重回帰モデルにおいてどの説明変数を回帰式の中に取り入れるかという変数選択問題がある．

またヒストグラムのセルの区間の決定問題も典型的なモデル選択問題といえよう．セルの区間を細かくして，各セルに1つか2つぐらいのデータしか出現しないのでは，ヒストグラムの意味をなさないし，逆に区間を広くしすぎてセルの数が少なくなってしまっても分布の特徴を表せなくなってしまう．このように考えていくと，統計モデルを利用する現場で頻繁に直面する問題に対し，モデル選択基準が有効であることがご理解いただけるであろう．

また，最近は純統計的問題ばかりではなく，例えばニューラルネットワークの分野でもモデル選択基準が中間ノード数決定などに用いられている．ニューラルネットの中間ノードの数が多ければ，自由度が増し多様な関数を表現できるが，逆に学習データに過度に適合した重みパラメータになり，誤差分散をを多く含んでしまう．そこで，適切なノード数決定のためにいくつかのモデル選択基準を用いた方法が提案されている．さらには，もっと一般的な帰納推論の分野においてもモデル選択基準と関連した様々な研究が盛んになってきている．

モデル選択基準が広汎な分野において利用されていることが以上でご理解いただけたと思う．モデル選択の研究は上記の問題を解決するのみにとどまらず，良いモデルとは何かという，統計学の，あるいは工学の本質的問題の考察を含んでいる奥の深いテーマといえる．

本特集は，統計モデル選択に関する4編の解説論文からなっている．

最初の松嶋による解説は，モデルの適切さの評価を，モデルの利用目的からとらえた統計モデル選択の概要となっている．

2番目の論文は，この分野の創始者である赤池氏御自身にAICの解説をしていただいた．AICの基本的考え方，導出の過程，そしてAICの影響を受けてその後に提案されたMDL，BICのモデル選択基準との対比を通じて，AICの本来の意味について論じていただいた．読者の方々にはAICの理解とともに，古典的検定論，推定論の枠組みを超えてモデルの比較を行うことを初めて可能にしたAICの意義を，歴史的視点からも再確認していただけると思われる．

3番目の論文は，確率的コンプレキシティと学習理論について山西氏に解説をしていただいた．データの記述長を最小にするモデルが良いモデルであるというMDL基準は，Rissanenによって2段階符号化を用いて提案された．その後MDLは様々な形に発展し，近年は確率的コンプレキシティと呼ばれる概念にまとめられるようになってきた．また，計算論的学習理論にモデル選択基準を適用することで，従来検討されなかった視点からモデル選択基準に関する新しい知見が得られてきている．今回の解説論文では，このような情報理論，学習理論の両分野の視点からモデル選択の解説をお願いした．

4番目の論文は，経営工学におけるモデル選択について関氏に解説をしていただいた．モデル群の中から1つのモデルを選択する場合に，1対のモデルに対し検定を繰り返すことで最後に1つの仮説を選択することも可能であるが，その比較回数は莫大であり，与えられた危険率のもとでの判定基準の設計は非常に困難であろう．このような多重検定の問題に対してのモデル選択の有用性や，逆にモデル選択を検定的に見た場合の誤り率などについて，応用例を交えて解説していただいた．

以上のように，この特集で執筆いただいた論文は一編一編が個性的で，いままでのモデル選択の解説論文とは一味違った内容となった．また，それぞれが異なった立場から独創的視点でモデル選択について論じていただいたため，モデル選択問題の根底に流れる興味深いテーマが，より一層浮き彫りになったのではと思われる．

最後になりましたが，お忙しいなか原稿の執筆を快諾していただいた執筆者の方々に心より感謝いたします．