回帰分析における

説明変数選択のための諸基準

1. はじめに

回帰分析の応用にあたって，もっとも頭を悩まされる問題は，説明変数の取捨選択である．モデルの定式化が，指定以前にはっきり決まっているなどということはめったにない．いくつかの候補変数群が与えられ，あれこれ試行錯誤をくりかえした後，最良と思われる回帰式が，最終的にひとつ選ばれる．試行錯誤の過程においては，主観的判断と客観的判断が入り混じる．

私がこの小論で論じようとしているのは，回帰式の変数選択に用いられるさまざまな客観的基準は，おのおのいかなる形式的合理性を背景とするものかという点．さらに，諸基準間の比較についてである．

いずれの基準も，読者にとってはなじみ深いものであろうし，またそれを実用された経験も豊富であろう．しかし，そうした統計的手法がどういう「意味」をもつのかについては，必ずしもよく知られていないと思う．こうした点についての理解を深めるうえで，この小論が多少ともお役にたてば幸いである．

叙述をなるたけ平易にするために，式の導出過程は原論文を参照していただくことにし，手法の「意味」についての説明に多くの紙幅をさくことにしたい．またこの小論は，私自身がやってきた仕事を中心にまとめたもので，いわゆるサーヴェイを意図するものではないことを，あらかじめお断りしておきたい．

2. 先験情報の活用

可能な説明変数群として，k 個の変数がリストされているとしよう．これらの変数の全部または一部をとりこんだ回帰式は，都合 2^k-1 通りありうる(たとえば $k=10$ とすれば1,023通り)．あくまで客観主義の立場にたって，“最良”な回帰式（説明変数の組合せ）を選択しようとすれば，原則として，可能な 2^k-1 本の回帰式をぜんぶ推定してみないといけない．そのためには，どういう順序で推定すればよいか，すなわち，どういうルールで変数の出し入れをやればよいかについて，さまざまな方法が提案されている([7], [8])．

その場合，なるべく体系的であり計算機にのせやすいこと，さらに計算の能率がよいこと，などがルールの望ましさの基準となる．

ともあれ，変数の出し入れのルールを計算機に記憶させておけば，従属変数と k 個の説明変数群の観測値系列を与えるだけで，自動的に 2^k-1 本の回帰式の推定結果がうちだされる．人間のやることは，これらの回帰式を相互に比較して，ベストとおぼしきものを選択することである．計算時間に何の制約もなければ，こうした手続きは（少なくとも主観的判断要素が入りにくいという意味で）望ましいであろう．実際，米国の文献などをみると，こうした手続きのための計算プログラムの，開発が盛んなようである．

しかしながら，わずか10個の変数から適切な組合せをえらぶために，1,023本の回帰式を推定するなどということは，どう考えても，時間と費用

の浪費ではないか．そこで，多少の客観性は犠牲にしても，もう少し能率的な方法はないものかということになる．

リスト・アップされたk個の変数は，必ずしも無差別ではなく，なんらかの基準にしたがって，“重要度(レレバンス)”に関する一定の順序づけを与えることができよう．目的が構造分析であれば，現象のふるまいに関する先験的理論情報にもとづき，“効いてる”変数は何かについて，あらかじめ多少は知っている．また予測が目的ならば，前もって観測しやすい（コストも安く誤差も少ない）変数が優先されるはずである．こうした順序づけは，完璧になされる必要はない．

たとえば，10個の変数のうち3個は絶対に落とせない．残りの7個の変数から，いくつかを追加的に選択したいというような場面には，しょっちゅう出くわす．これだけの先験情報があれば，可能な回帰式の本数を，一挙に1,023から128に減らせる．相当な節約ではないか．

多項式回帰や自己回帰の場合には[1]，変数の順序づけがほぼ確定している．こうした場合には，可能な回帰式の本数を，大幅に節減できる．すなわち一般に，p次の項が式に入れば，p次以下の項は必ず式に含まれる，とするのが自然であろう．したがって，多項式の次数が高々kであるという先験情報があれば，k個の回帰式を推定するだけでことが足りる．

このように変数が順序づけられている場合，逐次的に，1次から出発して2次，3次と順々に，あるいは逆に，k次から出発して$k-1$次，$k-2$次と順々に推定してゆき，一定の停止ルール（たとえば自由度修正重相関係数が減少したらうちきる）にしたがって，機械的に変数選択を行なうことができる．通常の回帰分析においても，変数の「重要度(レレバンス)」についての先験的順序づけが可能であれば，同様の方法が適用可能である．

また，追加された変数の寄与度を示すなんらかの統計量にもとづいて，逐次的に変数選択を行なう方法もありうる．変数を逐次的に追加してゆき一定の規則に従って停止する変数増加法，逆に変数を逐次的に除去してゆく変数減少法，それらを兼ねあわせた変数増減法などがある．これらの方法については，奥野他［19，pp. 137～152］にくわしく解説されている．この節のはじめに述べた“総なめ”式方法に比べれば，はるかに効率的であると同時に，あくまで「データをして語らしめる(letting data speak themselves)」という立場を守っているのが，これらの方法の特徴である．

3. 予備検定

回帰係数の有意性検定にもとづいて変数選択する方法を予備検定（preliminary test）という．予備検定にもとづく変数選択法の推測統計的意味づけについては，古くからさまざまな文脈において議論がなされている（[10]，[11]，[20]）．要は，同一の標本データを用いて，モデル（ないし変数）の選択と，しかる後の推定を行なうという一連の手続きが，推定結果にどのくらいのバイアスをもたらすか，さらに，（何もしない時に比べて）平均2乗誤差をいかほど低減させうるか，といった問題が論ぜられる．

もっとも標準的な問題設定は以下のとおり．通常の線形正規回帰モデル

$$(3.1)\quad y=X\beta+u=X_1\beta_1+X_2\beta_2+u,\qquad u\sim N(0,\ \sigma^2 I)$$

において，p個の変数X_1は絶対におとせない「核変数（core variables）」であり，q個の変数X_2は“効いてるかどうか不確かな”「任意変数（optional variables）」であるとしよう．たとえ$\beta_2=0$（X_2はまったく効いていない）としても，除去せずにそのまま推定すれば，（上記のモデルが真であるかぎり）β_1とβ_2の最小2乗推定量は

1) $y_t=\alpha+\beta_1 x_t+\beta_2 x^2_t+\cdots+\beta_p x_t{}^p+u$という型の回帰式のことを多項式回帰という．$y_t=\alpha+\beta_1 y_{t-1}+\beta_2 y_{t-2}+\cdots+\beta_p y_{t-p}+u$という型の回帰式のことを$p$次の自己回帰という．

不偏である．しかし，余計な変数X_2を含めたことにより，β_1 の推定値の標準誤差を，あたら大きくすることになる．逆にX_2を除去すれば，$\beta_2=0$でないかぎり，β_1 の推定と y の予測にバイアスが生じてくる．しかしその分，推定値と予測値の標準誤差は小さくなる．

そこで，つぎのような手続きがふまれる．X_2を含めた回帰式をひとまず推定し，その結果から，帰無仮説 $\beta_2=0$ を対立仮説 $\beta_2\neq0$ にたいして検定し帰無仮説が棄却されれば X_2 を含め，しからざるとき X_2 を排除する．一般性を失うことなく，$X_1'X_2=0$ と仮定すれば，上に述べた予備検定の手続きは，β_2 を，

$$(3.2)\quad \hat{\beta}_2=\begin{cases} b_2, & F>c \text{ ならば,} \\ 0, & F\leq c \text{ ならば,} \end{cases}$$

y の予測式を，

$$(3.3)\quad \hat{y}=\begin{cases} x'_1b_1+x'_2b_2, & F>c \text{ ならば,} \\ x'_1b_1, & F\leq c \text{ ならば,} \end{cases}$$

とするものである．ここで $(b_1,\ b_2)$ は $(\beta_1,\ \beta_2)$ の最小2乗推定量，

$$(3.4)\quad F=\frac{y'X_2(X'_2X_2)^{-1}X'_2y}{y'[I-X(X'X)^{-1}X']y}\div\frac{q}{n-p-q}$$

は，$\beta_2=0$ のとき自由度 $(q,\ n-p-q)$ の F 分布に従う確率変数であり，c は適当に選ばれた有意点である．$c=0$ ならば常に X_2 を含めることになり，予測 $\hat{y}$ は不偏だが分散は相対的に大きい．$c=\infty$ ならば常に X_2 を排除することになり，$\hat{y}$ は片寄った予測になる．バイアスとバラツキの両方を結合した基準として，平均2乗誤差を望ましさの基準としよう．任意変数 X_2 が一個しか存在しない $(q=1)$ 場合，異なる有意点 c に対応する平均2乗誤差は，**図 1** のような振舞いを示す．

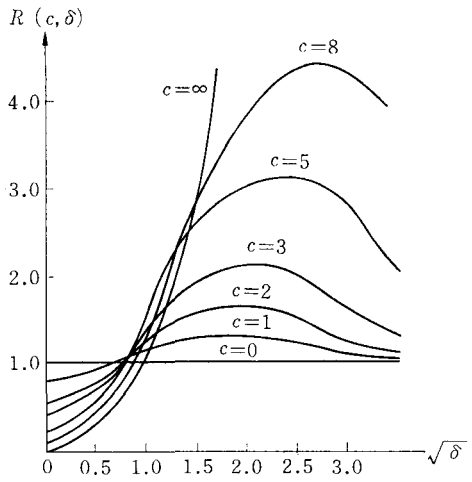

図 1 予備検定の平均2乗誤差

この図からまずわかるのは，平均2乗誤差を一様に小さくするような有意点は存在しないこと．さらに，通常の有意水準(5％または10%)でやると，非心度 $\delta(=\beta^2_{p+1}/\sigma^2)$ の値のいかんによって，相当大きな平均2乗誤差を覚悟しないといけない．そこで，決定理論におけるリグレットという基準をもちこむことにしよう．すなわち，平均2乗誤差を，有意点 c と非心度 δ を変数とする危険(リスク)関数とみなし，それを $R(c,\delta)$ と書く．c という有意点を選んだことにより被るリグレットは，

$$(3.5)\quad r(c,\delta)=R(c,\delta)-\min_{c}R(c,\delta)$$

と定義される．δ は，決定理論においていうところの「自然の状態(state of nature)」である．そこで，δ に関する最大リグレット $\max_{\delta} r(c,\delta)$ を最小にする c をもって“最適”とすることにしよう．かくして定義されるミニマクス・リグレット有意点は，自由度 $n-p-q$ と q に依存する．(くわしい数表は Sawa and Hiromatsu[14]に与えられている)．大ざっぱにいって，自由度が極端に小さくないかぎり，ミニマクス・リグレット有意点は，自由度のいかんにかかわらず，ほぼ一定値1.88前後である．ということは，最適な有意水準が，自由度とともに大幅に変動することを意味する(**表 1** を参照せよ)．

4. 予備検定に対する批判

予備検定の適用に対して，統計理論の立場から，つぎのような批判がなされる．「2乗誤差を損失関数とするとき，予備検定の結果として導かれる推定量 $\hat{\beta}_2$ は，非許容的 (inadmissible) であ

表 1　ミニマクス・リグレット有意点 ($q=1$の場合)

自由度	最適点	5%	10%	20%	30%
10	1.893	2.228	1.812	1.372	1.093
20	1.882	2.086	1.725	1.325	1.064
30	1.879	2.042	1.697	1.310	1.055
40	1.877	2.021	1.684	1.303	1.050
60	1.877	2.000	1.671	1.296	1.046
120	1.876	1.980	1.658	1.289	1.041

参考のために，5%，10% 等の有意水準に対応する有意点を併記した．

る」．別の言葉でいいかえれば，$\hat{\beta}_2$ の平均2乗誤差は，修正スタイン推定量，

$$(4.1)\qquad \beta_2^*=\left[1-\frac{c}{F}\right]^+ b_2$$

の平均2乗誤差よりも一様に（パラメータ値のいかんにかかわらず）大きい[2]．ただし c は，$0<c<2(q-2)(n-p-q)/[q(n-p-q+2)]$ となる定数であり，a^+ は $a<0$ ならば $a^+=0$，$a\geq 0$ ならば $a^+=a$ を意味する．（くわしくは［4］，［17］，［18］を参照せよ）．このことの意味は以下のとおりである．第1，予備検定にもとづく推定という常套手続きは，用いるべきでない．なぜなら，それよりも明らかにベターな推定法が存在するのだから．第2，予備検定つきの推定量の統計的性質をこれ以上理論的に吟味するのは意味がない．なぜなら，非許容的なものの中でベストなのは何か，といった類の問は所詮意味がない．

かくして，上記のやや驚くべき結果が証明されて以降，予備検定つき推定といういわゆる推測過程[20]に関する，統計理論家たちの関心は，急速にさめてしまったようである．以来，この問題に関連した論文は，もっぱら応用統計関係の雑誌に刊行の場を移したようである．たとえ非許容的な手法であっても，それが実際によく用いられているのだから，そうした手法に関する議論には相応の意味が認められてしかるべきではないか．また，スタイン流の推定量（最小2乗法を適用して得られた推定値に，1より小さい数をかけて短縮(shrink)させる）を現実の応用の場で用いるのは，どうも気持が悪いではないか．このような感想を抱かれる読者は少なくあるまい．（だからこそ，本誌の特集が組まれるのであろう）．

また，もう一つの抗弁として，つぎのような反論がありうる．私たちがやろうとしているのは，β_2 の推定ではなくて，モデルの選択(または識別)なのである．つまり，y を X_1 のみで説明するモデルと，y を (X_1, X_2) で説明するモデルを比較して，いずれか一方を選択しようとしているのであって，β_2 の推定という観点からの批判は，いささか的外れである．

5. 重相関係数の修正

そこで「回帰モデルの選択」という観点から，説明変数選択の問題を見なおしてみよう．そのためのもっとも基本的な統計量は，残差平方和 RSS_p と，その変換である重相関係数 R である．すなわち，回帰モデル (3.1) において，それらはおのおの

$$(5.1)\qquad RSS_{p+q}=y'[I-X(X'X)^{-1}X']y$$

$$R^2_{p+q}=1-RSS_{p+q}/\sum_{i=1}^{n}(y_i-\bar{y})^2$$

で与えられる．R^2_{p+q} が大きいほど（1に近いほど），回帰式のあてはまりは良好といえる．

とりあえず，二つのモデルが包含関係にある(nested)場合について考えよう．すなわち，

$$(5.2)\quad M1: y=X_1\beta_1+u,\quad u\sim N(0,\ \sigma^2 I)$$

$$M2: y=X_1\beta_1+X_2\beta_2+u,\ u\sim N(0,\sigma^2 I)$$

を比較する．前者が後者のスペシャル・ケースであるという意味で，両者の関係を包含(nested)であるという，$M1$ の R を R_p と書き，$M2$ の R を R_{p+q} と書くことにすれば，X_2 が何であれ $R_{p+q}\geq R_p$ という不等式が成立し，単に R の大小によってモデルの良し悪しを比較するのは無意味なことが，すぐにわかる．

2) このことが成立するためには $q\geq 3$ であること，さらに $X'X=I$ であることが必要などの制約はある．

一般に，説明変数を逐次的に追加していくとき，自由度（＝標本のサイズ−説明変数の個数）の低減を代償に，R の値をいくらでも大きくすることができる．自由度が低減するということは，推定値や予測値の分散が増大することを意味し，それ自体としては好ましくない．こうしたトレード・オフの関係を加味して，R になんらかの修正を加えてやる必要がある．

さまざまな修正の仕方がありうる．もっともよく用いられるのは，

$$(5.3)\quad \bar{R}_p^2=1-\frac{n-1}{n-p}(1-R_p^2)$$

という修正である[3]．通常，$\bar{R}$ のことを自由度修正重相関という．$\bar{R}$ は変数の追加とともに単調増加するわけではなく“効かない”変数を追加すると，かえってその値は小さくなる．$M1$ と $M2$ を $\bar{R}^2$ の大小によって比較するのは，予備検定つき推定(3.2)において，$c=1$ にするのと同じである．($q=1$ のときは，有意水準がおよそ30%強の F 検定を行なっていることになる)．回帰分析の応用の場では，$\bar{R}$ 最大化の決定方式がもっともよく用いられているようである．

このほか，説明変数群が多変量正規分布にしたがうことを仮定したうえで，予測の平均2乗誤差を最小化するという立場からの基準として，

$$(5.4)\quad \bar{\bar{R}}^2=1-\frac{n-2}{n-p-1}\cdot\frac{n-1}{n-p}(1-R^2)$$

という修正方法も提案されている．自由度が再修正されるわけである([21])．

6. 情報量基準

モデル選択の一般理論として，赤池弘次氏の情報基準(AIC)というのがある([1],[2],[3])．真の確率分布 $g(y)$ とモデル $f(y|\theta)$ との「距離」を，Kullback−Leibler の情報量

$$(6.1)\quad I(f:g)=\min_{\theta}\int g(y)\log\frac{f(y|\theta)}{g(y)}dy$$

によって測る，という考え方から導かれたものである．上記の量が小さい（真の確率分布との距離が近い）ほど，モデル $f(x|\theta)$ は望ましいとされる．大ざっぱにいって，AIC は，モデルの情報量の漸近的不偏推定量として導かれる統計量である．モデルの尤度関数を $L(\theta|y)$ とすれば，

$$(6.2)\quad \mathrm{AIC}=-2\log L(\hat{\theta}|y)+2p$$

となる．ただし $\hat{\theta}$ は θ の最尤推定値であり，p はモデルに含まれる未知パラメータの個数である．右辺の第1項は“尤度の最大値”に−2をかけたものであり，モデルのあてはまりのよさを測る．第2項は，パラメータの増加に対するペナルティーと解釈できる．

かくして「AIC の小さいモデルほど望ましい」ということになる．すなわち「データへのあてはまりがよくて，パラメータ節約的なモデル」が好ましいとされる．「AIC 最小化」のモデル選択原理を，MAIC (minimum AIC) という．

さて，MAIC を回帰モデル $M1$ と $M2$ の選択に適用すれば，ただちにつぎのような決定方式が導かれる．

$$(6.3)\quad F\leq[\exp(2q/n)-1](n-p-q)/q$$

ならば $M1$ を，しからざるとき $M2$ を選ぶ．上式の右辺を MAIC 有意点とよぶことにしよう．

くわしい解説は省略せざるをえないが，同様の考え方にもとづき，Sawa [15] が導いた情報量基準によると，つぎのような決定方式が結果する．$W=[1+qF/(n-p-q)]^{-1}$ とするとき，

$$(6.4)\quad n\log W-2(p+2)W+2W^2+2(p+q+1)<0$$

ならば $M1$ を採択し，しからざるとき $M2$ を採択する．上の不等式によって定義される有意点のことを MBIC 有意点とよぶことにする[4]．

表 2 の MAIC 有意点と **表 3** の MBIC 有意点を比較すると，つぎの点に気づく．第1，両者は漸近的に同等であり，有意点の漸近値は2である．第2，有意水準でみると，MBIC について

3) 右辺の修正係数の分子に，$n-1$ のかわりに n とされることもある．いずれにせよ，本質的には違わない．

は15～16%とほぼ一定値なのに対し，MBICのほうは20～15%の間を変動する．第3，MBICのほうが，よりいっそうパラメータ節約的である．

以上に紹介した情報量基準は，“真”の確率分布と，想定されたモデルの“隔り”の推定値を，モデル選択の基準としようとするものである．先に述べた予測の平均2乗誤差を危険関数とした決定方式に比べると，情報量基準は，変数の追加（パラメータの増加）に対して，より節約的である．しかし，通常の有意性検定（5%または10%有意水準）に比べれば，より放漫（prodigal）ではある．奥野他[19, p. 139]によると，有意点を2（情報量基準の漸近値）にとるのは，経験的にも，適切と思われるとのことである．

表 2　MAIC有意点と有意水準（$q=1$）

n \ p	2	3	4	5	10
10	1.573(.253)	1.329(.293)	1.107(.341)	.885(.400)	—
12	1.633(.233)	1.452(.263)	1.270(.297)	1.088(.337)	—
16	1.732(.211)	1.598(.230)	1.464(.252)	1.332(.275)	.666(.452)
20	1.788(.199)	1.682(.213)	1.578(.228)	1.471(.245)	.947(.356)
30	1.860(.184)	1.793(.192)	1.724(.201)	1.654(.211)	1.309(.267)
50	1.918(.173)	1.877(.177)	1.836(.182)	1.796(.187)	1.593(.214)
100	1.960(.164)	1.940(.166)	1.918(.170)	1.899(.172)	1.798(.184)
200	1.980(.160)	1.971(.162)	1.960(.164)	1.949(.164)	1.899(.170)
500	1.991(.158)	1.988(.160)	1.985(.160)	1.980(.160)	1.960(.162)
1000	1.997(.158)	1.994(.158)	1.991(.158)	1.991(.158)	1.980(.160)

表 3　MBIC有意点と有意水準（$q=1$）

n \ p	2	3	4	5	10
10	2.709(.144)	3.298(.119)	4.145(.097)	5.126(.086)	—
12	2.531(.146)	2.941(.125)	3.542(.102)	4.376(.081)	—
16	2.350(.149)	2.952(.133)	2.921(.116)	3.371(.096)	7.607(.040)
20	2.262(.151)	2.522(.139)	2.641(.125)	2.914(.110)	6.222(.034)
30	2.158(.153)	2.250(.146)	2.359(.137)	2.484(.128)	3.656(.071)
50	2.088(.155)	2.137(.151)	2.190(.146)	2.100(.154)	2.641(.112)
100	2.042(.156)	2.065(.154)	2.088(.152)	2.111(.150)	2.247(.138)
200	2.019(.156)	2.031(.156)	2.042(.154)	2.053(.154)	2.111(.148)
500	2.008(.158)	2.014(.156)	2.016(.156)	2.019(.156)	2.042(.154)
1000	2.005(.158)	2.005(.158)	2.008(.156)	2.011(.156)	2.019(.156)

7. Mallows の C_p 基準

もう一つ実用されることの多い基準として，Mallows の C_p 基準というのがある．この基準の導出に関しては，必ずしも適切な文献が見あたらないので，ややくわしく説明しておこう[5]．

平均値が変動する確率変数 Y に関する n 個の観測値から成る確率ベクトル y を，いくつかの説明変数によって“説明”したいとする．

(7.1)　$E(y)=\eta,\quad V(y)=\omega^2 I$

を仮定する．η の値を知りたいのだが，このままではどうにもしようがない．そこで，回帰モデル，

(7.2)　$y=X\beta+u,\ E(u)=0,\ V(u)=\sigma^2 I$

を想定する．すなわち，平均ベクトル η は，p 個のベクトル $X=(x_1, \cdots, x_p)$ で張られる線形部分空間に属するものと想定してみる．$X\beta$ は X の列で張られる部分空間への η の射影である．したがって，

(7.3)　$\beta=(X'X)^{-1}X'\eta$

となる．さて β の最小2乗推定量 $b=(X'X)^{-1}X'y$ を用いて，未知の定数ベクトル η を，

(7.4)　$\hat{y}=Xb=X(X'X)^{-1}X'y$

によって推定する．推定の平均2乗誤差は，

4)　AICとBICの基本的相違点は以下のとおり．BICの場合，$M1$ の情報量基準も $M2$ の情報量基準も，「より複雑なモデル $M2$ が“真”に近い」という仮定のもとに評価されるのに対し，AICの場合，$M1$ の情報量の評価にあたって，「$M1$ がほぼ“真”である」と仮定される．くわしくは Sawa [15] を参照．

5)　以下の説明は，Mallows [12] によって与えられた C_p 統計量に関する，筆者なりの解釈である．

$$(7.5)\quad \Delta_p=E\|\hat{y}-\eta\|^2=E\|X(X'X)^{-1}X'u\|^2+\|\eta-X\beta\|^2=p\omega^2+\eta'(I-X(X'X)^{-1}X')\eta=p\omega^2+SSB_p$$

となる．右辺の第2項は，η を X の列で張られる空間に射影したときの垂線の長さの平方であり，モデル(7.2)の偏りの2乗和とみることができる．

ところで，残差平方和 RSS_p の期待値は，

$$(7.6)\quad E(RSS_p)=(n-p)\omega^2+SSB_p$$

となることが，たやすく示される．したがって，

$$(7.7)\quad RSS_p+(2p-n)\omega^2$$

の期待値は Δ_p に等しい（ω^2 を既知とすれば，Δ_p の不偏推定量である）．Δ_p は η の推定の平均2乗誤差であるから，その値が小さければ小さいほど，モデルとしては望ましいことになる．(7.5)の右辺の第1項はパラメータ数の増加に対するペナルティーであり，第2項は回帰式の近似度のよさをあらわすという点，前節に述べた情報量基準と相通ずるところがある．

さて，以上のような考え方を背景として，Mallows は，

$$(7.8)\quad C_p=\frac{RSS_p}{\hat{\omega}^2}+2p-n$$

をもって，モデル選択の基準にすべきであるという．$\hat{\omega}^2$ は未知の分散 ω^2 の推定値である．いかにして ω^2 を，推定すべきかについて，完全に納得的な方法を提案することはできない．「もっとも複雑なモデルの分散の不偏推定量をもって，ω^2 の推定値とする」というのが，考えうる限りにおいて，もっとも納得のいく推定方法であろう．

モデルが包含関係（nested）にある場合，C_p にもとづく決定方式は，やはり F 統計量にもとづく決定方式であり，有意点を常に2とするものである．想定されたモデルが“真”であるということは，(7.5)式の右辺の第2項がゼロということである．このとき RSS_p の期待値は $(n-p)\omega^2$ となり，($\hat{\omega}^2$ の確率的変動を無視すれば) C_p の期待値は p となる．この点に着目すれば，横座標に説明変数の個数(p)を目盛り，縦軸に C_p を目盛ったグラフを作図するという方法が提案される．45°線に近いほど「近似度」は高く，かつまた原点に近いほど望ましい．

Mallows の C_p 基準も，漸近的には AIC と同等になる．しかし，Mallowsのアプローチは，分布型に対する仮定がおかれていないという長所をもっている．

8. 不偏な決定方式

さて以上において，変数選択のための基準をいくつか紹介してきたが，いずれもそれなりの形式的合理性を背景としており，一概にどの基準が良いとか悪いとか論ずることはできない．ともあれ，比較の対象となるモデルが包含関係にある場合，いずれも予備的F検定に帰着する．差異は，有意点のとり方にのみ関わる．

そこで，回帰モデル(7.2)を想定することのリスクを，Mallows の C_p で測るとして，$M1$ と $M2$ を比較してみよう[6]．

$\Delta_p\leq\Delta_{p+q}$ のときは $M1$ が，$\Delta_p>\Delta_{p+q}$ のときは $M2$ が望ましい，と考えることに異論はあるまい．Δ_p や Δ_{p+q} はもとより未知である．$F>c$ または $F\leq c$ に応じて $M2$ または $M1$ を選択するという決定方式について，

$$P(F\leq c|\Delta_p\leq\Delta_{p+q})\geq.5$$
$$P(F>c|\Delta_p>\Delta_{p+q})\geq.5$$

の2条件が満たされるなら，c を有意点とする決定方式は不偏であるということにする．F 分布の連続性によって，上記の2条件は，

$$P(F\leq c|\Delta_p=\Delta_{p+q})=.5$$

と同値である．検定統計量 F は，$SSB_{p+q}=0$ のとき（$M2$ が真であるとき），非心度 $\delta=SSB_p/\omega^2$ の非心 F 分布に従う．ところで条件 $\Delta_p=\Delta_{p+q}$ は，$\delta=q$ と同値であることが簡単に示される．これ

6) 正規性の仮定のもとに，Kullback—Leibler の情報量を基準にとっても，同様の結果が導かれる．

表 4　不偏決定の有意点

d.f. \ q	1	2	3	4	5
10	1.388	2.686	3.258	3.568	3.756
12	1.357	2.628	3.190	3.489	3.675
16	1.320	2.557	3.105	3.397	3.576
20	1.300	2.513	3.056	3.342	3.519
30	1.272	2.462	2.989	3.272	3.445
50	1.250	2.421	2.941	3.218	3.389
100	1.234	2.390	2.904	3.179	3.349
200	1.225	2.375	2.887	3.158	3.327
500	1.221	2.365	2.876	3.147	3.316
1000	1.221	2.362	2.873	3.144	3.312

$d.f.$＝自由度

より，不偏な決定方式を与える c は，自由度 $(q, n-p-q)$，非心度 q の非心 F 分布のメディアンにほかならない．こうして求まる不偏有意点は，表 4 に見るとおりである．（くわしくは［16］を参照せよ）．

表 1 ～ 4 を比較してみると，いくつかのおもしろい事実がよみとれる．$q=1$ の場合に限って見てみよう．情報量基準にしろ C_p 基準にしろ，いずれもより簡単なモデル（説明変数の少ないモデル）$M1$ のほうに片寄っている，すなわち，$M1$ と $M2$ が無差別（$\triangle_p=\triangle_{p+q}$）のとき，$M1$ を選ぶ確率が 1/2 以上である．自由度修正重相関 $\bar{R}$ にもとづく決定は，$q=1$ のとき，ほぼ不偏である．

9. 包含関係にない場合

モデルが包含関係にない場合でも，AIC や $\bar{R}$ はそのまま適用できる．しかし BIC 基準や C_p については，未知の分散 ω^2 をいかにして推定すべきか，というやっかいな問題が生じてくる．たとえば，

(9.1)　$y=X_1\beta_1+u$

(9.2)　$y=X_2\beta_2+u$

を比較する場合，両方を含むモデル，すなわち，$X_1 \cup X_2$ を説明変数とするモデルを推定して，その不偏分散推定値を $\hat{\omega}^2$ とすることが考えられる．

こうした場合，いずれの基準に従うにしても，予備的検定との関連をつけることはむずかしい．というよりは，予備的検定をエクザクトに行なうことからして不可能である．Cox［5］,［6］は帰無仮説としてのモデルと対立仮説としてのモデルが包含関係にない場合の尤度比検定について論じている．尤度比の対数に -2 をかけた統計量が，包含関係にある場合は近似的に χ^2 分布に従うけれども，しからざるときは，こうしたことが成り立たない．

そこで尤度比の分布を正規近似して，検定方式を導こうというのが，Cox の考え方である．回帰モデルの変数選択や関数型の選択のために，Cox の方法は有効と思われる．しかし紙幅の関係上，ここでその問題に深入りする余裕はないので，可能性を指摘するだけにとどめておこう．

10. 数 値 例

回帰分析の応用例として引用されることの多い Hald［9］のデータに，以上述べきたった諸基準を適用してみよう．従属変数と 4 個の説明変数に関する13個の観測値データは，表 5 に見るとおりである．可能な15本の回帰式について，必要な統計量が表 6 にまとめられている．

C_p 基準の計算に必要な ω^2 の推定値は，すべての説明変数を含んだもっとも大きなモデルの不偏分散推定値を用いることにした．このほかたとえば，すべての可能な回帰式の不偏分散推定値を比べてみて，その最小値を $\hat{\omega}^2$ とする，という方法も考えられる．いずれにせよ，$\hat{\omega}^2$ のとり方によって，変数選択の結果が影響されるという点は，C_p を実用化するうえでの難点といえよう．

AIC による序列と C_p による序列とは，ほぼ一致している．これらの基準が，漸近的に同等であることから，予想される結果といえよう．$\bar{R}$ による序列と，AIC または C_p による序列との間には，かなりの差が認められる．すでに述べたよう

表 5 Hald のデータ系列

	X_1	X_2	X_3	X_4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

X_1=3 $CaO \cdot Al_2O_3$ 量
X_2=3 $CaO \cdot SiO_2$ 量
X_3=4 $CaO \cdot Al_2O_3 \cdot Fe_2O_3$ 量
X_4=2 $CaO \cdot SiO_2$ 量
Y=セメント 1g 当たり発熱量

表 6 Hald のデータにもとづく回帰式

説明変数	RSS	R^2	AIC	C_p	$\bar{R}^2$	$d.f.$
(1)	1265.7	.534	98.4(13)	200.5(14)	.492(13)	11
(2)	906.4	.666	92.1(10)	140.5(12)	.636(10)	11
(3)	1939.4	.286	104.0(15)	313.2(15)	.221(15)	11
(4)	883.9	.675	93.7(11)	136.7(11)	.645(12)	11
(1,2)	57.9	.979	58.3(1)	.68(1)	.975(4)	10
(2,3)	415.4	.847	85.9(9)	60.4(9)	.816(9)	10
(1,3)	1227.1	.548	100.0(14)	196.1(13)	.458(14)	10
(1,4)	74.8	.972	63.6(6)	3.5(6)	.966(6)	10
(2,4)	868.9	.680	95.5(12)	136.2(10)	.616(11)	10
(3,4)	175.7	.935	74.7(8)	20.4(8)	.922(8)	10
(1,2,3)	48.1	.982	59.9(3)	1.04(3)	.976(2)	9
(1,3,4)	50.8	.981	60.6(4)	1.50(4)	.975(3)	9
(1,2,4)	48.0	.982	59.9(2)	1.02(2)	.976(1)	9
(2,3,4)	73.8	.973	65.5(7)	5.33(7)	.964(7)	9
(1,2,3,4)	47.9	.982	61.8(5)	3.00(5)	.974(5)	8

たとえば(1, 2)は，X_1 と X_2 を含む回帰式という意味である．AIC, C_p, $\bar{R}^2$ の欄のカッコの中の数字は，おのおのの基準による回帰式のよさの順序づけである．

に，$\bar{R}$ にもとづく決定方式は，他の基準と比べて，変数の追加に対して寛容である．そのため，AIC と C_p が (1,2) を選ぶのに対し，$\bar{R}$ は (1,2,4)を選ぶ．しかしながら，15本の回帰式の「順序づけ」に関するかぎり，諸基準間に大差は見られない．ちなみに AIC と $\bar{R}$ による「順序づけ」の間の順位相関係数は0.97である．

予備的検定にもとづく逐次選択法は，変数群にどういう先験的序列を与えるかによって，結果に大差が生じてくる．たとえば，従属変数との単相関の大きさによって (4,1,2,3) という序列を与えたとしよう．(4)→(4,1)→(4,1,2)→(4,1,2,3)という順序で，AIC（または C_p または $\bar{R}$）が減少する限り前に進む，という決定方式に従うとしよう．いずれの基準によるとしても，x_4 のみを説明変数とする式 (4) が選ばれてしまう．逆に，(4,1,2,3)→(4,1,2),→(4,1)→(4) と進むことにすれば，(4,1,2)が選択される．式(4)は，全体の順序づけでは下位（AIC と C_p では11位に $\bar{R}$ では12位)にランクされているにもかかわらず，前者の逐次決定方式によると選択されることになる．奥野他[19, pp. 137～8]による「変数増加法」によると，(4,1,2)が選ばれる．また「変数減少法」だと(2,1)が選ばれる．

参考文献

[1] Akaike, H. (1970) "Statistical Predictor Identification," *Ann. Inst. Statis. Math.*, Vol. 22, pp. 203—217.

[2] Akaike, H. (1972) "Information Theory and an Extension of the Maximum Likelihood principle," *Problems of Control and Information Theory*, AKADEMIAI KIADO (Publishing House of the Hungarian Academy of Sciences), pp. 202—212.

[3] Akaike, H. (1974) "A New Look at Statistical Model Identification," *IEEE Transactions on Automatic Control*, Vol. 19, pp. 716—722.

[4] Cohen, A. (1965) "Estimates of the Linear Combination of Parameters in the Mean Vector of a Multivariate Distribution," *Annals of Mathematical Statistics*, Vol. 46, pp. 78—87.

[5] Cox, D.R. (1961) "Tests of Separate Families of Hypotheses," *Proc. 4th Berkeley Symp.*, Vol.

1, pp. 105—123.

[6] Cox, D. R. (1962) "Further Results on Tests of Separate Families of Hypotheses," *J. R. Stat. Soc., B*, Vol. 24, pp, 406—424.

[7] Furnival, G.M.(1971) "All Possible Regressions with Less Computation," *Technometrics*, Vol. 13, pp. 403—408.

[8] Garside, M. J. (1965) "The Best Sub-Set in Multiple Regression Analysis," *Appl. Stat.*, Vol. 14, pp. 196—200.

[9] Hald, A. (1952) *Statistical Theory with Engineering Applications*, Wiley, New York.

[10] Larson, Harold J. and T. A. Bancroft (1963) "Sequential Model Building for Prediction in Regression Analysis, I," *Annals of Mathematical Statistics*, Vol. 34, pp. 462—479.

[11] Larson, Harold J. and T.A. Bancroft(1963b) "Biases in Prediction by Regression for Certain Incompletely Specified Models," *Biometrika*, Vol. 50, pp. 391—402.

[12] Mallows, C. L. (1973) "Some Comments on Cp," *Technometrics*, Vol. 15, pp. 661—675.

[13] Sawa, T. (1968) "Selection of Variables in Regression Analysis," *Economic Studies Quarterly*, Vol. 19, pp. 55—63.

[14] Sawa, t. and T. Hiromatsu (1973) "Minimax Regret Significance Points for a Preliminary Test in Regression Analysis," *Econometrica*, Vol. 41, pp. 1093—1101.

[15] Sawa, T. (1977) "Information Criteria for the Choice of Regression Models," *Econometrica*, in press.

[16] Sawa, T. and K. Takeuchi (1977) "Unbiased Decision Rules for the Choice of Regression Models,"

[17] Sclove, Stanley L. (1968) "Improved Estimators for Coefficients in Linear Regressions," *Journal of the American Statistical Association*, Vol. 63, pp. 597—606.

[18] Sclove, S. L., C. Morris and R. Radhakrishnan (1972) "Non Optimality of Preliminary-Test Estimators for the Multinormal Mean," *Annals of Mathematical Statistics*, Vol. 43, pp. 1481—1490.

[19] 奥野忠一ほか (1971)『多変量解析法』, 日科技連出版社.

[20] 北川敏男 (1958)『推測過程論』現代応用数学講座B—10a, 岩波書店.

[21] 佐和隆光 (1968) 「予測効率による回帰モデルの選択」,『季刊理論経済学』17巻3号, pp. 65—69.

[22] 佐和隆光 (1970)『計量経済学の基礎』, 東洋経済新報社.

OR手帳

文献の整理法

オフィスで毎日作成される文献の整理には誰しも悩みながら, これは, という解決法がない. そこで実務的に見てまぁ満足と思われる方法として, 国際機関の文献整理法を参考のためご紹介します.

(1) すべての文献に必ず組織コード, 部会コード, 年, 一連番号の4種および改訂番号から成る identifier (以後文献コードという)を付す.

(2) あとから文献を取り出すために, 組織コード, 部会コード別にファイルを作り, 年, 一連番号, 改訂番号の順に綴じる.

(3) 会議の案内(議事次第)に各議事に上述の参照すべき文献コードを付す. なお, 議事の順と文献コードの順がなるべく対応するように, 文献コードは時により予約される.

(4) 上述の議事次第だけを別にまとめて綴っておくと, これが内容から文献コードへの索引帳になるので, 議事次第がくわしいほど, 有用な索引となる.

非常に単純であるが, なかなか実用的ファイル法である. 難点は厚い文献が多い場合, すぐにファイルがパンクすることであるが, 予算が許せば, マイクロフィッシュのような形で保存するとスペースが節約できる. またこれは60ページ程度の文献でも1枚のコピーですむのでいっそう便利になると思う.

また, 共同研究のような場合, 一つの文献が複数の部会や組織の活動に関連することもあるが, この場合, 一つの文献に複数の文献コードを付し, コピーをそれぞれの部会や組織のファイルに綴じておくことが一つの工夫です (もちろん, 文献X→Y参照というメモだけを綴じておいてもよいが). (入沢 元)

フォーラム