推定と予測における直観的な論点の吟味

—— 回帰分析の場合 ——

竹　内　　啓

1　問題の所在

経済分析の統計的方法の体系としての計量経済学の方法は，同時方程式体系の母数推定論を中心として，きわめて精緻なものとして完成されていることは，いまさらいうまでもない。

しかしながら，現実の経済分析においては，精密な確率的モデルを前提とする厳密な方法を，現実のデータや，それが反映している経済構造のあり方に照らして見るとき，あまりにも sophisticated であって，形式的な厳密性が非現実的な基礎の上に立っているという感じを抱かせることは，決して少なくない。

勿論私は，計量経済学的方法を根本から否定しようとする考え方[1]には同調できないし，また想定されたモデルと現実との間に多かれ少なかれギャップが存在することは，モデルが一種の“理想化された関係の体系”を表現するものである以上，いわば最初から前提されているのであって，そのことは直ちにモデル分析の有効性を損うものではないと思っている。しかしそれにもかかわらず，モデルの極度の精密さが，果してその有効性を高める上に真に必要なものであるかという疑問は残るであろう。

もう一つの問題は，精密なモデル化が困難であることが明白な場合，しかもそこにおいて何らかの定量的分析が要求される場合も少なくないことである。このような問題を，計量経済学の対象の外にあるものとして，それを直観的な，純然たるアド・ホックな方法に委ねてしまえば，計量経済学の体系のいわば理論的純粋性は保たれるかもしれないが，しかしそれは現実の経済分析の要求によく応えるものではなくなるであろう。

ここでいうまでもないことではあるが，モデルと現実とのギャップは，モデルの仮定をゆるめ，ただモデルを“一般化”することによって埋めることはできないことを強調しておきたい。そのようにして表面的にモデルを現実に対応させることができても，むやみに一般的なモデルは分析の道具としての有効性を失ってしまう。

ここで問題は，ある程度一般的な条件の下で，ある程度の有効性を保つような方法を見出すこと，すなわち広い意味で robust な方法を考えることである。robust な方法ということばはふつう確率分布の型が“標準型”(多くの場合正規分布が想定される)から離れた場合にも，効率を失わない方法という意味に用いられるが，ここではより一般に標準的な状況からのいろいろな方向へのズレに対して考えねばならない。

robust な方法を求める際の問題状況の特徴はかつて Allan Birnbaum が指摘したことがあるが[2]，その fuzzyness，すなわちその限界が不明確であるという点にある。すなわち標準的な状況からの乖離の可能性について，その大体の方向と，程度とはわかっており，従って考えられる程度のズレに対してあまり効率を失わないような方法を求めることが問題になるが，その方向や限界について，明確な定義はできないというのが，その標準的な場合なのである。問題の要点はむしろ漠然としたものであっても，ある程度の方向や範囲が想定されているという点にあり，従って例えば分布型についての仮定を全く置かないノンパラメト

1)　計量経済学に対するマルクス経済学の立場からの批判はすでに1955年に出た広田純・山田耕之介「計量経済学批判」『近代経済学批判講座』第3巻(日本評論社)に基本的な論点は展開されている。

2)　personal discussion による。

リック法とは，異る考え方に立って問題が論ぜられなければならない。

実はわれわれが現実のデータ解析にあたって事前に持っている情報は，多かれ少かれ fuzzy なものである。われわれがモデルを構成するにあたって，その情報の一部は確固としたものとみなしてモデルの中に組み入れ，また一部は不確かなものとして一応除外し，またもしベイズ的方法を用いるならば一部は母数の事前分布という形で定式化する。このようにして fuzzy なものに一応明確な区切りをつけたものがモデルなものである。もしこのことが適切に行い得ないとされる場合には，モデルを構成することなく，fuzzy な情報をそのまま用いて直観的な方法が用いられることになる。

しかしこの2つの場合の区別は絶対的ではない。どのような場合においても，モデルは全く疑問の余地がないということはないし，またモデルの中に組みこまれなかった情報が全くないということもあり得ない。逆にどのように fuzzy な情報の下であっても，全くモデル化が不可能ということもない。多くの場合その差は程度の問題にすぎない。従って一見あいまいな情報をなるべく数学的に客観的な論理の中にとり込むことが必要である。

このような情況を理論的にとり扱うには，モデルについて一種の二重構造を想定しなければならないことが多い。すなわち“真の構造”についてのある程度一般的な，しかし精密には規定されていない前提と，それについて仮定される“モデル”すなわち必らずしも正確ではないが，一連の精密な仮定とである。そうして後者から導びかれる方法が，より一般的な前提の下でどのような性質を持つか，そうしてそのことと基準として，モデル自体の適切さをどのようにはかるかを考えねばならない。この問題の一つの解は赤池氏の情報量基準によって与えられる[3]。それはあるモデルの適切さを

AIC＝－対数尤度＋推定される母数の数

で計るもので，これを最小にするモデルが，最も望ましいとされる。それはモデルのデータに対する「あてはまり」のよさと，母数の数によって表わされるモデルの複雑さの間のバランスをはかろうとするものである。この基準の合理化，或いはそれ以上立ち入った議論にはここではふれないが，ここで“モデル”と真の分布との違いが最初から考慮に入れられていることに注目したい。

以下では経済分析においても最もふつうに使われる手法である回帰分析について，問題を論じよう。

2 回帰分析における Mallows の Cp 統計量

いま，ある経済変量 Y の変化を，いくつかの説明変数を用いて“説明”することを考えよう。

いま被説明変数 Y，および考えられる説明変数の候補 $x_1, x_2, \cdots$ についての時系列データ

$$Y_t, x_{1t}, x_{2t}, \cdots, t=1, 2, \cdots T$$

が与えられたものとしよう。

分析の目的は，これらのデータを用いて，適当な説明変数をえらんで推定式

$$\hat{Y}=\hat{\alpha}+\hat{\beta}_1 x_{j1}\cdots+\hat{\beta}_p x_{jp}$$

を求めることである。

このときふつうに行われる方法は，まづいくつかのモデル

$$M_i;\ Y_t=\alpha_i+\beta_{i1}x_{i1t}+\cdots+\beta_{ip}x_{ipt}u_t \quad t=1\cdots T$$

$i=1, 2\cdots$ を考え，その中から最も適当なものをえらび，その母数を推定して，Y の推定式を定めるという方法である。

この方法は実は論理的な難点をふくんでいる。というのはそれぞれのモデルの下での母数の推定は(当然のことながら)そのモデルが正しいという前提の下で行われる。しかし実はいくつかのモデルを比較する場合には，正しいモデルは2つ以上はあり得ないはずである。更に厳密にいえば真に“正しい”モデルは一つも存在せず，ただより真に近いモデルとそうでないモデルとが存在するだけであると考えねばならないであろう。そこで現実には多くのモデルについての計算結果をいろいろ比較して，適当と思われるものをえらぶことになる。そうしてモデルをえらんだ後には，それが多くの中からえらばれたものであるということは不問にして，あたかもそれが最初から仮定された

3) 『数理科学』特集情報量基準 1976 年 3 月号。

ものであるかのように見なされるのがふつうである。

このようなモデル選択においては，決定係数を基準とするほか，係数推定値の符号や大きさの尤もらしさ等から直観的に判断がなされるが，その論理は必らずしも明確でないことが多い。そこでこの点をもう少し論理的に扱うことにしよう。そのためにまず一般的な構造として，

$$Y_t=\eta_t+u_t \qquad E(u_t)=0 \qquad t=1\cdots T \qquad (1)$$

と想定し，次に η_t を

$$\eta_t=\alpha_i+\beta_{i1}x_{i1t}+\cdots+\beta_{ip}x_{ipt}+\zeta_{it} \qquad (2)$$
$$\sum_t x_{ijt}\zeta_{it}=0 \qquad j=1\cdots p$$

と表現することにすれば，モデル M_i の“偏り”は

$$b_i^2=\sum\zeta_{it}^2$$

で表わすことができる。この値の小さいモデルが望ましいモデルである。

ここで $E(u_t^2)=\sigma^2 \quad E(u_tu_{t'})=0 \quad t\neq t' \qquad (3)$

とし，形式的にモデル M_i をあてはめて推定を行えば，$\beta_{i1}\cdots\beta_{ip}$ の最小2乗推定量 $\hat{\beta}_{i1}\cdots\hat{\beta}_{ip}$ は，やはり不偏推定量になり，またその分散共分散行列は，偏り ζ_i が存在しない場合と同じになる。また，

$$\hat{Y}_{ti}=\hat{\alpha}_i+\hat{\beta}_{i1}x_{i1t}+\cdots+\hat{\beta}_{ip}x_{ipt}$$

とおくと，

$$\sum E(Y_t-\hat{Y}_{ti})^2=(T-p-1)\sigma^2+b_i^2$$

となるから，$Q_i=\sum(Y_t-\hat{Y}_{ti})^2$ とするとき

$$\hat{\sigma}_i^2=Q_i/(T-p-1)$$

は過大な推定量になる。

ところで，一般に説明変数を増せば，偏り b_i^2 は減少するであろう。しかし逆にそれによって Y_t の“理論値” $\hat{Y}_t$ の分散は大きくなるであろう。そこでこの2つの傾向のつり合いを考えねばならない。

Mallows はこの点に関連して次のような概念を導入した[4]。いま Y_t^0 を Y_t と同じ期待値を持ち，かつ $Y_1\cdots\cdots Y_T$ とは独立な分布に従う変数とする。すなわち，

$$Y_{t'}^0=\eta_{t'}+u_{t'} \quad t=1,\cdots T$$

であって $u_{t'}$ $t'=1\cdots T$ は u_t $t=1\cdots T$ とは互いに独立であると仮定する。

そうして Y_t^0 の値を $\hat{Y}_{ti}$ を用いて予測するときの平均2乗誤差を考えると，その和は，

$$W_i=\sum E(Y_t^0-\hat{Y}_{ti})^2=(n+p+1)\sigma^2+b_i^2$$

となる。従って，

$$W_i=E(Q_i)+2(p+1)\sigma^2$$

となる。そこで W_i が小さいようなモデルが望ましいと考えることができる。そこで W_i の推定量として

$$\hat{W}_i=Q_i+2(p+1)\hat{\sigma}^2$$

或いは，$C_p=Q_i/\hat{\sigma}^2+2(p+1)$

を用いる。$\hat{\sigma}^2$ は何らかの形で求めた σ^2 の推定量である。C_p は Mallows の C_p 統計量といわれている。

いくつかのモデル M_i $i=1,2,\cdots$ の中から最もよいモデルをえらぼうとするとき，それぞれについて C_p 統計量を計算して，それが最小になるようなモデルをえらぶことが考えられる。

この方式については，いろいろな技術的な吟味もなされており，また母数の数 $p+1$ の係数2が最も適当であるかどうかについても議論の余地がある[5]。しかし少なくとも第一次的近似としては，この方法はモデル選択に一つの合理的な基準を与えるといってよい。

ここで一方における説明変数の組が，他方のそれの部分集合となっているような2つのモデル，

$$Y=\alpha+\beta_1x_1+\cdots+\beta_px_p+u$$

および

$$Y=\alpha+\beta x_1+\cdots\beta_px_p+\beta_{p+1}x_{p+1}+\cdots+\beta_qx_q+u$$

の比較において，問題を仮説

$$\beta_{q+1}=\cdots=\beta_q=0$$

の検定という観点から考えることもできよう。このとき，ふつうに用いられる F 統計量は，

$$F=\frac{(Q_p-Q_q)}{(q-p)\hat{\sigma}^2} \qquad \hat{\sigma}^2=\frac{Q_q}{n-q-1}$$

と表わされるから，Mallows の C_p 統計量を用い

4) Mallows, C. L., “Some coments on C_p,” *Technometrics*, 15 (1973). ただし彼は C_p 統計量をこれより数年早く発表しており，この前からそれを用いた他の人の論文が現われている。

5) 竹内啓「回帰分析における変数選択基準について」『経済学論集』42巻2号，1976年。

る方法は($\hat{\sigma}^2$ を上記の値に等しくおくと)$F \gtrless 2$ に応じて仮説をすてたり，受容したりすることに対応する。

2という限界は F 検定のふつうに用いられる水準に対応する棄却限界よりも小さい。このようないわゆる検定—推定型の多段階推測(或いは推測過程)の問題において，仮説の棄却限界はふつうの5%水準よりも小さく，10～30%水準に対応する値にとらねばならないことは，以前から知られていたことであり，また実際 F 検定の限界を2に定めるということも，これまで経験的に採用されて来たところであるが，上の議論はその一つの根拠を与えている。

ただしここで注意すべきことは，Mallowsの考えを適用する場合，候補となるモデルすべてについて同時に比較を行わねばならないこと，また σ^2 の推定量 $\hat{\sigma}^2$ はほぼ偏りのないものであること，すなわち十分多数の説明変数を用いたモデルの残差から計算されねばならないことである。従って説明変数を逐次的に1つづつ増したり減らしたりする場合に，F 検定を機械的に適用し，2という値を基準として用いることは適当でない。特に一つづつ説明変数をふやして行く，いわゆる前進法は説明変数の少なすぎるモデルで止まってしまう危険性がある[6]。

3 回帰式の意味の再吟味

しかしこのような理論にもなお疑問の余地が残る。

先にのべた構造においては，Y_t の期待値 η_t と，説明変数 x_{1t}, x_{2t}……とを結びつける式(2)は，全く形式的に定義されており，そこには特に“因果関係”は仮定されていないし，説明変数 x_i の係数 β_i が，その変数の Y におよぼす影響の強さを直接表現しているわけでもない。このことは，説明変数の係数の定義が，説明変数と“偏り”とが直交するという条件(2)に依存していることから明らかである。

従って係数 β_i の値はそのモデルにおいて当の変数 x_i のほかにどのような変数がふくまれるかによって変わるのである。

もし特定の変数の直接の影響の大きさをはかることが必要であり，従ってその変数の係数 β_i が独立の意味を持つならば，直交条件(2)はなり立つとは限らない。従ってまた推定量 $\hat{\beta}_i$ は偏りを持つことになる。しかしその偏りはまたモデルにふくまれる他の変数によって影響されるから，その意味を解釈することが困難になる。

β_i の意味に立ち入ることをさけて，Y の“予測”のみを問題としたのは，論理的な困難をさける巧妙な方法ということができるかもしれないが，しかしなお問題は残されている。というのは，予測される時点において，説明変数 x_{it} の値が，データにおける値と全く一致するという仮定の不自然さは，便宜的なものとして容認するとしても，説明変数の値の組が一致すれば，その期待値も一致すると考えてよいか否かには疑問があるからである。上記の議論では，予測時点において，すべての説明変数の値がデータ時点の値に一致するとしたのであるが，このような仮定の下では Y の期待値もまた両者において等しくなると想定してもよいかもしれない。しかし，一度特定のモデルを選択してしまえば，その式はその中に現われなかった説明変数の値は考慮せずに用いられるであろう。その式をあてはめるための条件として，その中に現われない説明変数の値を前提することは，不自然でもあり，またできる限り簡単なモデルをえらぶという要求に反することになるであろう。

従って一部の説明変数の値を定めれば，他の説明変数の値が一義的に決定されてしまうような場合(例えば，多項式モデルにおいて，各次数の項，或いはそれと同等な直交多項式のそれぞれを説明変数とみなすとすれば，それらの間には関数関係が成立する)を除けば，上のような接近法はなお不十分であるといわねばならない。

このように考えると，結局説明変数の間には何らかの形で一定の関係が存在し，そうしてその関係は予測期においてもなり立つと想定しなければ，予測に関連して変数選択の問題を考えることができないことがわかる。実際極端な場合として，デ

6) 5)に同じ。

ータ期間においては説明変数の間に完全な線型関係が成立する一方，予測期間においてその関係がもはや成り立たず，しかも Y はこれらすべての変数によって影響されるとすれば，適切な予測量を求めることはそもそも最初から不可能になる。

4　説明変数の種類区分

ここで更に立ち入って論ずるためには説明変数の種類をわけて考えることが必要になる。まず基本的に制御可能な変数と制御不能な変数との区別が重要である[7]。制御可能な変数の中では，自由に動かし得る変数の数，すなわちその自由度は一般に最初から与えられている。そうしてまたそれらの変数の一部をモデルから除くとしても，それはそれらの変数を動かしても Y の値に影響を与えないであろうということを想定するだけで，それらを全く考慮しないわけにはいかない。それらの値もとにかく特定の値に定めておかねばならない。Y の値を特定の方向に動かそうとして制御変数の値を変えるとするならば，データに照らしてあまり影響がないように思われる制御変数の値は，データにおける平均値に等しく取っておくのが最も安全であろう。これに対してもしそのような変数を何らかの理由によってデータ期間における値から大きく動かさねばならないとき，その結果を正確に予測することは非常に困難になる。

一方説明変数が制御不能であるときには，回帰式にもとづいて予測を行うということは，説明変数の値を知ることを前提として Y の値を予測することを意味する。そうしてモデルの中にふくまれる変数を制限するということは，説明変数の一部を知って予測を行おうとすることにほかならない。しかしその場合には他の変数がどうなっているかについての情報は得られないであろう。従って予測に用いられる変数以外の変数が大きく変動して，それによって Y の値に予期しない変動が生ずるというようなことはないものとしなければならない。

$x_1, x_2, \cdots$ が制御可能でない変数の場合，それらを更に Y に対して明確な因果関係を持つと思われる変数と，Y に影響を与えるような状況を直接或いは間接に表現していると思われる変数とに分けて考える必要がある。前者についてはその変動が Y に影響をおよぼさないということが，すなわち因果関係があるという想定が誤りであったことが明確に示されない限り，モデルから除くべきではない。後者については，実はそれは直接 Y に影響をおよぼすのではなく，Y に影響をおよぼすような，直接には観測されない変量によって影響され，それを反映しているものとみなすことができるであろう。

以上のことをまとめると，モデルは一般に次のように表わされる。

$$\begin{aligned} E(Y_t)=\eta_t=\alpha&+\beta_1 x_{1t}+\cdots+\beta_r x_{rt}\\ &+\beta_{r+1}\tilde{x}_{1t}+\cdots+\beta_{r+s}\tilde{x}_{st}\\ &+\gamma_1\xi_{1t}+\cdots+\gamma_k\xi_{kt}\end{aligned}$$

ここに $x_1\cdots\cdots x_r$ は制御可能な変数，$\tilde{x}_1\cdots\cdots\tilde{x}_s$ は因果関係を表わす変数，$\xi_1\cdots\cdots\xi_k$ は未知の構造を表わす変数である。

そうして $\xi_1\cdots\cdots\xi_k$ に対しては，他の“説明変数” $x_1'\cdots x_q'$ が存在して，

$$x_{jt}'=c_{j1}\xi_{1t}+\cdots+c_{jk}\xi_{kt}+v_{jt}\quad j=1\cdots q$$

と表わされるものとしよう。

一般にこのような，モデルにおいて r, s, k はあまり大きくないものとしなければならない。しかも更に $\beta_1\cdots\beta_{r+s}$ のいくつかは絶対値が0に近いので，その項は無視してもよい程度であると想定されよう。

そうすると結局問題は $x_1\cdots x_r$ の中の若干を固定し，$\tilde{x}_1\cdots\tilde{x}_s$ の中の一部を除き，また $x_1'\cdots x_q'$ の中からいくつかを $\xi_1\cdots\xi_p$ の代わりに用いて，

$$\begin{aligned}\hat{Y}=\hat{\alpha}&+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_r x_r\\ &+\hat{\beta}_{r+1}\tilde{x}_1+\cdots+\hat{\beta}_{r+s}\tilde{x}_s\\ &+\hat{\gamma}_{i1}x_{i1}'+\cdots+\hat{\gamma}_{ip}x_{ip}'\end{aligned}$$

という形の式を構成することである(ただしここで $\hat{\beta}_1\cdots\hat{\beta}_{r+s}$ のうち一部は最初から0とおく)。

そうすると説明変数の一部を除いたことの影響は，制御可能な変数についてはその値をデータにおける平均に等しく固定しておく限り問題はない。また因果関係を表わす変数については，それを無

7)　この点に関しては竹内啓『計量経済学の研究』東洋経済新報社，1973年第，12章参照。

視することは偏りを生ずるが，その大きさを理論的に考慮する上で特別な点はない。

問題は最後の部分である。議論を簡単にするために $x_1 \cdots x_{r+s}$ を無視すると

$$\begin{aligned} Y_t &= \alpha + \gamma_1 \xi_{1t} + \cdots + \gamma_k \xi_{kt} + u_t \\ x_{jt}' &= c_{j1}\xi_{1t} + \cdots + c_{jk}\xi_{kt} + v_{jt} \end{aligned} \tag{4}$$

となる。ここで u_t と v_{jt} とは互いに独立であると考えてよいであろう。また v_{jt} $j=1 \cdots q$ も互いに独立と仮定しておく。このことは Y, x_j の相関を表わす部分は，すべて ξ で表現されるものと仮定すれば一般性を損わない。

いま ξ は未知であり，その個数 k も未知であるとすると，上記の構造は因子分析のモデルに一致する。

ここで2つの方法が考えられる。一つはまず $x_1' \cdots x_q'$ についてふつうの因子分析の手法[8]を用いて因子スコアを求めてそれを $\hat{\xi}_{it}$ とし，次に Y のこれらの値に対する回帰式を計算し，

$$\hat{Y} = \hat{\alpha} + \hat{\gamma}_1 \hat{\xi}_{1t} + \cdots + \hat{\gamma}_k \hat{\xi}_{kt}$$

とする。更に $\hat{\xi}_{it}$ の $x_1' \cdots x_q'$ に対する回帰式を

$$\hat{\xi}_{it} = d_{i1}x_{1t}' + \cdots + d_{iq}x_{qt}'$$

とし，これを先の推定式に代入すれば，

$$\hat{Y}_t = \hat{\alpha} + \hat{\beta}_1' x_{1t}' + \cdots \hat{\beta}_q x_{qt}'$$

という形の“予測式”が得られる。

この方法は C. R. Rao が推称するところであって[9]，彼は $\hat{\xi}$ に時々用いられるような主成分分析法による主成分ではなく，因子スコアを用いることを主張しているが，その理由は上記のような構造から導びかれるであろう。ただしここで因子スコアをもう一度 x' で表現しておくことは，予測に用いるために必要である。またこの過程の途中で，適当に“座標軸の回転”を行って，最後に得られた $\hat{Y}$ の式においてなるべく多くの変数の係数を0に近くすることが望ましいであろう。

第2の方法は x, Y をともに一度に座標変換して，Y の因子荷重から $\hat{\gamma}_1 \cdots \hat{\gamma}_k$ を推定することである。そうして次に上記と同様に ξ を x' の一次式で表わして代入し，Y の予測式を求めることができる。どちらがよりよいかは一概にはいえないようである。

より一般に，制御可能な変数，および明確な因果関係を持つ変数が存在する場合には，まず Y のこれらの変数に対する回帰式を計算し，その残差部分について上記の方法を適用すればよい。

5 多重共線性と“縮小”推定量

現実のデータ解析において，回帰分析に際し多重共線性が生ずることは極めて多い。またそれに関していくつかの方法が提案されている。特に最近ではいわゆる Ridge estimator などという，推定量の絶対値を小さくする手法が論ぜられている[10]。それは最小2乗法における正規方程式

$$\boldsymbol{M}_{xx}\hat{\boldsymbol{\beta}} = \boldsymbol{M}_{xy}$$

の代わりに，この係数行列を拡大して，

$$(\boldsymbol{M}_{xx} + k\boldsymbol{I})\hat{\boldsymbol{\beta}}^* = \boldsymbol{M}_{xy}$$

の解を推定量とするものである。多重共線性のために x 変数のモーメント行列 $\boldsymbol{M}_{xx}$ が特異に近くなるとき，最小2乗解は大きく変動するが，ridge 推定量 $\hat{\boldsymbol{\beta}}^*$ はあまり大きな値をとることがなくなる。

この方法についても Mallows の C_p 統計量を用いて，k の適当な大きさを定めたりすることも可能である。

しかしこの場合にも説明変数の性質が明確でないならば，このようなことを行うことの意味がはっきりしない危険性がある。

また Mallows の場合と同じ基準を用いて，最小2乗推定量以外の推定量 $\hat{\beta}_i^*$ を用いて

$$Y_t^* = \bar{Y} + \hat{\beta}_1^*(x_{1t} - \bar{x}_1) + \cdots + \hat{\beta}_p^*(x_{pt} - \bar{x}_p)$$

とし，予測誤差の平均平方の和を計算すれば，

$$\begin{aligned} \sum E(Y_t^0 - \hat{Y}_t^*)^2 &= (\boldsymbol{\beta}^* - \boldsymbol{\beta})'\boldsymbol{M}_{xx}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) + b^2 \\ &\quad + \text{trace}\, \boldsymbol{M}_{xx}\Sigma_{\hat{\beta}^*} + (n+1)\sigma^2 \end{aligned}$$

という形に表わされる。ただし $\boldsymbol{\beta}^*$ は $\hat{\beta}_i^*$ の期待値，$\boldsymbol{\beta}$ は β_i をそれぞれ成分とするベクトル，$\Sigma_{\hat{\beta}^*}$ は $\hat{\boldsymbol{\beta}}^*$ の分散共分散行列である。また b^2 はモデル自体の偏りの2乗和である。

8) 竹内啓・柳井晴夫『多変量解析の基礎』東洋経済新報社，1972年参照。

9) 1975年来日の際の講義。

10) Hoerl, A. E. and Leslie, R. N.,“Ridge regression: biased estimators in non-orthogonal problems,” *Technometrics*, *12*, 1970.

説明変数を直交変換して $\boldsymbol{M_{xx}}=\boldsymbol{I}$ としておけば，

$$\sum_i (Y_t^0-\hat{Y}_t^*)^2=\sum_i(\beta_i^*-\beta_i)^2+\sum V(\hat{\beta}_i^*)$$
$$+b^2+(n+1)\sigma^2$$
$$=\sum_i E(\hat{\beta}_i^*-\beta_i)^2+b^2+(n+1)\sigma^2$$

となる。そこで $\sum_i E(\hat{\beta}_i^*-\beta_i)^2$ をなるべく小さくする推定量が望ましい推定量である。

ここでもし誤差 u が正規分布に従うものと仮定すれば，$\boldsymbol{M_{xx}}=\boldsymbol{I}$ のとき最小2乗推定量 $\hat{\beta}_i$ は互いに独立に平均 β_i 分散 σ^2 の正規分布に従い，そうして $p>2$ ならば，Stein の有名な推定量[11]

$$\hat{\beta}_i^*=\max\left\{0,\ 1-\frac{(p-2)f\hat{\sigma}^2}{(f+2)\sum\hat{\beta}_i^2}\right\}\hat{\beta}_i$$

を用いれば，つねに

$$\sum E(\hat{\beta}_i^*-\beta_i)^2<p\sigma^2=\sum E(\hat{\beta}_i-\beta_i)^2$$

となることが示される。ただしここで $\hat{\sigma}^2$ は σ^2 の不偏推定量，f はその自由度である。仮説 $\beta_i\equiv 0$ を検定する場合の F 統計量を F にすれば，上の推定量は

$$\hat{\beta}_i^*=\max\left(0,\ 1-\frac{(p-2)f}{(f+2)F}\right)\hat{\beta}_i$$

となる。Stein 推定量はいろいろな形に変形することができるが，いずれにしても最小2乗推定量を縮小する(shrink)ならば，その比率が $\hat{\beta}_i$ の値に無関係に定められる ridge 推定量よりも，F 比に依存して定める Stein 型の推定量の方がより適当であると思われる。

Stein 型の推定量はモデルの中の変数の一部の係数にのみ適用することもできる。たとえば一組のダミー変数(季節変動を表わす項など)の係数の推定量を $\hat{\gamma}_1\cdots\hat{\gamma}_k$ とするとき，もしこれらが互いに独立ならば，これを次のように“縮小”して用いるのが適当である。

$$\hat{\gamma}_i^*=\max\left(0,\ 1-\frac{(k-2)f\hat{\sigma}^2}{(f+2)\sum\hat{\gamma}_i^2}\right)\hat{\gamma}_0$$

しかし Stein 型の推定量にしても，性質の異る説明変数をすべて一まとめにして適用することは好ましくない。

多重共線性の処理にあたっては，それぞれの変数の特質を考慮することが重要であって，ただ数理形式的な方法によってのみそれを扱うことは適当でないと思う。

なお多重共線性が生ずる場合，しばしば“おかしい”推定値が現われる。しかしある値がおかしいと判定されることは，係数に対してある程度事前の情報が存在していることを意味する。すなわちわれわれは係数の存在範囲についてある程度の知識は持っていることが多い。もしそのような情報，或いは知識を事前分布という形で表現することができるならば，Bayes の方法が適用できるであろう。いま $\boldsymbol{\beta}$ の事前分布が平均ベクトル $\boldsymbol{b}$ 分散行列 $\boldsymbol{\Sigma}$ を持つとすれば，その事後平均はほぼ

$$(\boldsymbol{M_{xx}}\hat{\sigma}^2+\boldsymbol{\Sigma})\hat{\boldsymbol{\beta}}^*=\boldsymbol{M_{xy}}\hat{\sigma}^2+\boldsymbol{\Sigma}\boldsymbol{b}$$

或いは

$$(\boldsymbol{M_{xx}}+\hat{\sigma}^{-2}\boldsymbol{\Sigma})\hat{\boldsymbol{\beta}}^*=\boldsymbol{M_{xy}}+\hat{\sigma}^{-2}\boldsymbol{\Sigma}\boldsymbol{b}$$

をみたす。従って $\boldsymbol{b}=\boldsymbol{0}$ とすれば，これは ridge type の推定量になる。或いはモデルを書変えて，

$$y_t-\boldsymbol{b}'\boldsymbol{x}_t=\alpha+(\boldsymbol{\beta}-\boldsymbol{b})'\boldsymbol{x}_t+u_t$$

という形に変形し，これに ridge 推定量の考え方を適用すると考えてもよい。ridge 推定量はむしろ Bayes 推定量の一つと理解する方がその意味が明確になるかもしれない。

或いはまた上記のように変換したモデルに Stein の方法と適用することも考えられる。

Bayes 法に対する批判は，主として事前分布の想定に主観性が入るということにある。しかしながら ridge 法などは，それ以上に恣意的な面をふくんでいるのであって，たまたまあるデータに適用して尤もらしい結果が得られたとしても，その判定は Bayes 法の場合以上に主観的といわざるを得ない。私は Bayesian ではないが，もし事前の判断を推定の中に取り入れるとするならば，論理の明確な Bayes 法による方が，Ridge 法などのあまりにも ad hoc な方法を用いるよりも適当であると思う。

6　変数変換の問題

ところで，被説明変数 Y に，予測するに当って，適当な変数変換を行ってから回帰分析を適用することもある。変換を主要な理由としては，非線形

11)　回帰分析の場合については C. Stein “Multiple regression” in *Contribution to Probability and Statistics*, ed. by Olkin et al., Stanford, 1960.

関係の線形化，誤差分散の均一化の2つが考えられる。対数変換などは，この2つの目的に同時に合致すると考えられることが多いが，それがつねに同時に解決されるとは限らない。また第3の理由として，誤差分布の正規分布からの極端な隔り，とくに非対称性を除くという目的も考えられる。正の値のみをとる変数を対数変換する目的の一つはこの点にもある。しかしこの目的も上記の目的と同時に達成できるとは限らない。

例えば Y がポアソン分布に従うと考えられる離散量であるときには，$\log Y$ について線形モデルを考えるのが母数の構造としては最も適当であるが，分散を均一にするためには $\sqrt{Y}$ を，正規分布に近づけるためには $Y^{2/3}$ を用いなければならない。

ここでいろいろな変数変換の適切さを比較することを考えよう。そのためには先の構造を僅かに一般化して(3)の代りに，

$$E(u_t^2)=\sigma_t^2 \qquad E(u_t u_{t'})=0 \qquad (5)$$

と仮定する。もし観測値が互いに独立ならば，(1), (2), (5)で表わされるモデルはどのように変換を行った値についてもあてはまることはいうまでもない。

いま簡単のために説明変数がすべて互いに直交するとしよう。そうするとモデル M_i をあてはめて最小2乗法により推定を行ったときの係数の推定量は，

$$\hat{\beta}_{ij}=\sum x_{ijt}Y_t/\sum x_{ijt}^2$$

となる。これはやはり β_{ij} の不偏推定量になるが，その分散は今度は

$$(\sum x_{ijt}^2\sigma_t^2)/(\sum x_{ijt}^2)^2$$

になって，簡単にならない。そうしてまた

$$E\{\sum(Y_t-\hat{Y}_t)^2\}=\sum\sigma_t^2+b_i^2-\sum_j(\sum x_{ijt}^2\sigma_t^2)/(\sum x_{ijt}^2)$$

となるから，これを，$\sum\sigma_t^2+b_i^2-\sum_j\bar{\sigma}_{ij}^2$ と表わそう。次に先の場合と同様にして Y_t^0 を Y_t と等しい平均分散を持ち，かつそれを独立な量とすると，

$$E(\sum(Y_t^0-\hat{Y}_t)^2\}=\sum\sigma_t^2+b_i^2+\sum_j\bar{\sigma}_{ij}^2$$

となることがわかる。従って

$$E\{\sum(Y_t^0-\hat{Y}_t)^2\}=E\{\sum(Y_t-\hat{Y}_t)^2\}+2\sum_j\bar{\sigma}_{ij}^2$$

となることがわかる。ここで $\bar{\sigma}_{ij}^2$ は σ_t^2 の加重平均であるが，その値は x_{ijt}^2 と σ_t^2 との関係に依存して定まるから，一般にはそれについて何もいうことはできない。従って Mallows の C_p 統計量にあたるものを計算することはできない。勿論分散比が既知であれば，その逆比を加重として加重最小2乗法を適用することができる。そのときには，

$$E(\sum(Y_t^0-\hat{Y}_t)^2/\sigma_t^2)=E(\sum(Y_t-\hat{Y}_t)^2/\sigma_t^2)+2(p+1)$$

が成立するから，Mallows の考え方を僅かに修正して適用することでができる。しかも $\sum_j\bar{\sigma}_{ij}^2$ が未知の場合その代わりにいいかげんな推定量を用いると，偏った結論に導びかれる危険性が大きい。

分散の構造についての2つのモデル

1 $E(u_t^2)=\sigma^2$, 2 $E(u_t^2)=c_t\sigma^2$ (c_t は既知)

を比較するには，もはや平均2乗誤差は使えないから対数尤度を比較しなければならない。正規を仮定するとモデル2の下での対数尤度は，

$$2\log\lambda^{(2)}=\text{const}-n\log\hat{\sigma}^{*2}-\sum\log c_t$$

となる。ただし $\hat{\sigma}^{*2}$ は加重最小2乗法を適用した場合の σ^2 の推定量である。モデル1の下での対数尤度は，

$$2\log\lambda^{(1)}=\text{const}-n\log\hat{\sigma}^2$$

となる。従って

$$(\prod c_t)^{1/n}\hat{\sigma}^{*2}>\hat{\sigma}^2$$

ならば，分散の不均一性を考慮しないモデル1の方がよく，逆の不等号が成立すれば，分散比についての仮定2を入れた方がよいことになる。

分散の不均一性は，その構造について明確な仮定を想定しない限り，データからは検出されないことが多く，そうしてそれを見落すと，モデルの適切さについて誤った判断を下す危険性が大きい。従ってデータの性質についての考慮から，その構造についてのあり得べきパターンを注意深く考慮しなければならない。そうしていくつかの可能性についてモデルを作って比較する必要がある。

変数変換において，分散の不均一性には大きな注意を払わねばならない。多くの場合変数変換に

ついて1)分散の均一性を重視して非線形モデルを構成するか，2)線形モデルにして分散の不均一性を許容するかの2つの方法がある。前者の場合は非線形最小2乗法を，後者の場合は加重最小2乗法を用いなければならない。コンピューターがあれば計算の複雑さは本質的な困難とはならない。しかし結果の評価は，多くの場合1)の考え方をとった方が容易になる。

被説明変数 Y に直接回帰モデルをあてはめた場合と，$Y'=\phi(Y)$ と変換して，Y' に線形モデルをあてはめた場合との2つを比較する一つの方法は，その対数尤度を比較することである。すなわち Y についてのモデルの対数尤度を $\log\lambda^{(1)}$, Y' についてのモデルの対数尤度を $\log\lambda^{(2)}$ とするならば，Y 自体についての第2のモデルの下での尤度はヤコビアンの因子を加えて，

$$\log\lambda^{(2)}+\sum\log|\phi'(Y_t)|$$

となる。そこでこの値と $\log\lambda^{(1)}$ を比較してどちらがより適当であるかを判定することができる。回帰モデルの場合，

$$\log\lambda=\text{const}-n\log\hat{\sigma}$$

となるから，このような計算は容易である。

ただしこの場合2つのモデルにおいては当然に誤差分散の構造についての想定が異って来ることに注意を要する。誤差分散についての別個のチェックなしに上記のような方法だけでモデルを選択することは危険である。

なお Y 自体についてのモデルと，変換した値についてのモデルにおいて推定される母数の数に差があるときには，赤池情報量(AIC)との考え方を応用して，それぞれの対数尤度に母数の数を加えて比較を行えばよい。

変数変換を行うこと自体の評価は行われることが少ないが，AIC などを用いる評価は，少なくとも第一次的接近として有効であろう。

7　非正規性と最小2乗法の是非

ところでこれまで推定法はほとんど最小2乗法によるものとして来た。ところで最小2乗法は誤差分布の正規性を前提とし導びかれるものであるから，このことは間接的に誤差の正規性を想定していることを意味する。“縮小法”を論じたときにも正規性の仮定の下で議論を進めた。

しかし回帰式の誤差項が正規分布に従うという仮定は，あまり根拠のないものであるといわざるを得ないであろう。この点を強調すると推定法としての最小2乗法の妥当性も疑わしいと思われるかもしれない。もし誤差分布が正規分布より著しくスソが長い(確率密度が0に近づく近づき方がおそい)と想定されるならば，最小絶対偏差法，すなわち，

$$d=\sum|Y_t-\alpha-\beta_{i1}x_{i1}t-\cdots-\beta_{ip}x_{ipt}|/n$$

を最小にする値を推定値とする方法を適用することも考えられる。そうして最小2乗法と最小偏差法のどちらをとるべきかについては比

$$\sqrt{\frac{2}{\pi}}\hat{\sigma}/d$$

を基準とすることができる。正規分布の場合この値がほぼ1になるから，これが1より著しく大きい，例えば2以上ならば，平均偏差法を用いる方がよい。

このようなタイプのいわゆる robust な推定量を求める問題は，P. Huber らによって論ぜられているが[12]，しかし問題の形式的な処理よりも，残差分析を用いる方が実際的である。それには

$$\hat{u}_t=Y_t-\hat{Y}_t$$

或いは，それを基準化した，

$$\hat{u}_t'=\hat{u}_t/a_t \quad \text{ただし } a_t=E(\hat{u}_t^2)/\sigma^2$$

を用いる方がよい。$\hat{u}_t'$ の中で著しく大きいものがあれば，それを“異常値”とみなすことができる。その棄却限界としては，ふつうの Thompson-Smirnoff の棄却限界の値を $\sqrt{1-1/n}$ 倍したものを用いれば近似的には十分である。このような“異常値”はそれが本当に他の値にあてはまる式からズレた値であるか，または分布が正規分布でないかの，2つの理由によって生じ得る。しかしいずれの場合にもその値を除いて最小2乗推定を行えば，精度のよい推定量が得られるであろう。しかしこのようにして予測値を計算したとき，その信頼性は見かけより落ちることに注意しなけれ

12)　P. Huber, “Robust estimation,” Lecture note, Yale University, 1971.

ばならない。すなわち予測される値が“異常値”になるかもしれないからである。“異常値”による検定は，コーシー分布のような非常にスソが長い分布に対しては(標本数があまり小さくない限り)検出力がかなり高くなるので，このような方法はかなり有力である。すなわち“異常値”に注意することにより(そうしてこのことはデータにつきまとうミスを発見するためにも怠ってはならないことである)，非正規性から生ずる最小2乗推定量の効率の低下は，かなり防ぐことができる。

現実の問題においては，データから明らかに見られる非正規性を見落さない限り，非正規性の問題はそれほど重大でないといってよい。

8 むすび

以上にのべたような問題点は，より複雑な問題，例えば同時方程式体系や，時系列の場合に生ずるものと，本質的には共通であると考えられる。そのような場合にも，より簡単なモデルとより複雑なモデルとの間の選択，変数変換の是非，非正規性の問題等が生ずる。そこでやはり C_p 統計量や尤度比，或いはそれを拡張したAIC統計量などの考え方を利用することができるであろう。またその際，あれこれの手法を形式的に適用するだけでなく，各変数の意味と性質，分析や予測の目的を具体的に考慮して，適切(relevant)な結論を導くことが大切である。このことは統計的推測のすべての場面を通じていわれるべきことであるが，その際何らかの意味で指針となるような数学的理論を構成することが大切であると思う。ここでのべたことはそのための一つの手がかりとなり得るであろう。

(東京大学経済学部)