回帰における準共線性

田中浩光

Near Collinearity in Linear Regression

Hiromitsu TANAKA

序

回帰分析では，結果として得られる推定回帰式の解釈に困難をきたすことが多い．また，手許の回帰データに対し，特定の観測値（観測個体）・観測特性項目の削除，あるいは新たな追加から推定回帰式が一変することも往々にして経験する．推定結果の不安定さが問われることになる．分析の目的が予測にある場合でも，この難点を避けることは容易ではない．推定回帰式の脆弱さは対象とする母集団の規定，観測特性項目の選定，標本抽出の設計，観測機構，推測方式などに因ると考えられる．推定回帰式に実質的な解釈を付与するには，とくに標本抽出過程，観測機構に起因する影響の評価に関心を払わなければならない（田中 (1993b)）．

無理な解釈を強いる，これらの原因として説明変数間に相関の強い場合が考えられる．所謂，多重共線性の問題が惹起する．とくに，観測研究の場に限局すると，データは計画・管理の側面から離れて収集の側面が強く，その説明変数行列は非直交，あるいは準特異になることもまれでない．Mason (1975) に依れば，多重共線性の生起には観測研究にまつわるモデルの不完全規定，サンプリングの不備，そして物理的制約が主因であるとしている．

多様な回帰問題においても回帰平面である説明変数行列の果す役割が推測の側面で本質的であり，その振舞は鋭敏である．したがって，この説明変数行列，とくに準共線性の問題は避けて通れないものとなる．

本論では，回帰における多重共線性の問題を準共線性の問題として整理し，従来の対応と異なり，悪条件の問題として位置づけることなく功罪両面について論旨を展開する．とくに，準共線性の特性を積極的に活用することが，モデルの妥当性確認に有用となることを主張する．2節では回帰モデルを与え，併せて悪条件の問題をとりあげる．3節では準共線性の問題を診断・治療の視点から整理する．4節では非直交性の影響を緩和する視点に立ち，従来の枠組の中で基軸となる最小二乗推定量の代替として縮小推定量を導入する．ここでは，3回帰問題をとりあげ，準共線性の影響を検討する．第1の問題では回帰係数の最小二乗推定量を凌ぐ偏推定量の構成，第2では標的母数に関する先験情報の活用か否かの二律問題を有する予備検定推定量の構成，第3では影響力のある観測値の検出統計量の構成について考察する．5節では説明変数行列が回帰問題に対し決定的役割を果すことに留意し，準共線性に絡む若干の疑問と示唆を提示する。

回帰モデルと悪条件

一般に，線形回帰分析では，次のモデルが仮定される．

$$\underline{Y}=X\underline{\beta}+\underline{\varepsilon} \tag{1}$$

ここに，Yは応答変数からなる$n\times 1$ベクトル，Xは説明変数からなる$n\times p$行列，$\underline{\beta}$は未知パラメータからなる$p\times 1$ベクトル，そして$\underline{\varepsilon}$は誤差を表す$n\times 1$ベクトルである．さらに，応答変数および各説明変数の平均はゼロと仮定し，$\underline{\varepsilon}$の各成分は相互に独立に平均ゼロ，分散σ^2の正規分布に従うとする．Press (1972) は種々の回帰モデルを観測機構に応じて5通りに分類している．標準回帰モデルがモデル1に，説明変数・応答変数が同時分布すると仮定した上で応答変数が説明変数上の条件付きのもとで観測されるとする条件付回帰モデルはモデル2に分類される．いま，n個の観測個体が得られるとして，条件付回帰モデルは

$$\underline{Y}|X=X\underline{\beta}+\underline{\varepsilon},\quad \underline{\varepsilon}\sim N(\underline{0},\ \sigma^2 In) \tag{2}$$

と与えられる．ここに，$N(\underline{0},\ \sigma^2 In)$は平均ベクトル$\underline{0}$，対角要素に分散$\sigma^2$をもつ多変量正規分布を示す．すなわち，同時分布が多変量正規分布であれば，条件付回帰モデル(2)は標準回帰モデルと同等になる．通常，$\underline{\beta}$の推定量として最小二乗推定量

$$\underline{\hat{\beta}}=(X'X)^{-1}X'\underline{Y} \tag{3}$$
$$=S^{-1}X'\underline{Y} \tag{4}$$

が採用される．

Xの特異値分解は

$$X'=\sum_{l=1}^{p}\lambda_l^{\frac{1}{2}}\underline{V_l}\,\underline{U_l}' \tag{5}$$

と表される．ここに，λ_lは$X'X$の固有値，$\underline{V_l}$はλ_lに対応する$X'X$の固有ベクトル，$\underline{U_l}$はXX'の固有ベクトルである．なお，$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$とする．$T$は転置を示す．

$\underline{\beta}$の平均平方誤差，期待平方長，予測平均平方誤差はそれぞれ

$$\mathrm{MSE}(\underline{\hat{\beta}})=\mathrm{E}(\underline{\hat{\beta}}-\underline{\beta})'(\underline{\hat{\beta}}-\underline{\beta})=\sigma^2\sum_{l=1}^{p}\lambda_l^{-1} \tag{6}$$
$$\mathrm{E}(\underline{\hat{\beta}}'\underline{\hat{\beta}})=\underline{\beta}'\underline{\beta}+\sigma^2 tr(X'X)^{-1}>\underline{\beta}'\underline{\beta}+\sigma^2\lambda_p^{-1} \tag{7}$$
$$\mathrm{MSEP}(y|\underline{x_0})=\sigma^2\{1+\underline{x_0}'(X'X)^{-1}\underline{x_0}\} \tag{8}$$

となる。ここに，$\underline{x_0}$は説明変数の将来の観測点ベクトルである．したがって，極小な固有値があれば，$\underline{\hat{\beta}}$は膨脹して不安定となり，所謂，悪条件(ill-conditioned)の問題を惹起する．条件付回帰モデルが背景にあるときは多重共線性（multi-collinearity）の問題といわれる．多重共線性はXの列のある部分集合について1次従属関係が成立するとき存在すると定義されている(Mason and Gunst(1985))．しかし，実地の場では，とくに観測研究データであれば正確な成立よりも近似的成立を問われることが多く，準共線性(near collinearity)が主要な関心事となる．換言すれば，Xが準特異あるいは非直交となる回帰データを解析対象とすることになる．

準共線性の診断と対処

準共線性の問題が生じると，すなわち説明変数行列Xが準特異あるいは非直交であるとき，データ解析の実践の立場に沿って，診断と対処について整理する．

準共線性の程度を測る尺度には代表的に用いられているものとして，条件数（condition number)，分散膨脹因数（variance influence factor)

$$\kappa(X)=\lambda_1/\lambda_p \tag{9}$$

$$\mathrm{VIF}_i=(1-r_i^2)^{-1} \tag{10}$$

があげられる．ここに，r_iは第i説明変数の重相関係数である。また，Farrar and Glauber (1967)は多重共線性の検出について，多重共線性の生起，程度とパターンを検出する一連の3段階からなる階層的な検定方式を提示している．データ解析の実地では，特異性の検出よりも準特異あるいは非直交の状況把握に有用されている．Farrar and Glauberの検出方式を要約する．最初に，多重共線性の存在の有無を問い，次いで生起とした場合には他の変数集合間での存在を問う．最後に対の説明変数をとりあげて検出する方式である．より具体的には次の通りである．いま，RをX^TXの相関行列の形式で基準化されているとする．このとき，第1段階では説明変数集合全体で検出するを意図して，説明変数が相互に独立とする帰無仮説の下では，検定統計量

$$\chi_0^2=-\left(n-1-\frac{2p+5}{6}\right)\ \ell oge\ |R| \tag{11}$$

が自由度$\frac{1}{2}p(p-1)$のχ^2分布に従うことを利用する．第2段階では，多重共線性が生起しているとして，第i説明変数が他の説明変数と独立とする帰無仮説の下では，検定統計量

$$w_i=(r^{ii}-1)\left(\frac{n-p}{p-1}\right),\quad i=1,2,\cdots,p. \tag{12}$$

が自由度$(p-1,\ n-p)$のF分布に従うことを利用する．ここに，r^{ii}は，R^{-1}の第i対角要素である．第3段階では，第i説明変数絡みで多重共線性の所在が明らかであるとして，第i説明変数と第j説明変数が独立とする帰無仮説の下で検定統計量

$$t_0=\frac{r^{*ij}/(n-p)^{\frac{1}{2}}}{(1-r^{*ij})^{\frac{1}{2}}} \tag{13}$$

が自由度$n-p$のt分布に従うことを利用する．ここに，

$$r^{*ij}=-r^{ij}\cdot/\ \{(r^{ii})^{\frac{1}{2}}(r^{jj})^{\frac{1}{2}}\} \tag{14}$$

であり，r^{ij}はR^{-1}の第$(i,\ j)$対角要素である．

回帰データが悪条件であると認めたとき，データ解析の立場から，その原因に応じた方策が根治的に実施されなければならない(田中(1992))．母集団から偏りを有して標本抽出がなされる偏サンプリングが認められるときは計画された追加観測が要請される．また，母集団分布の端値に相応する標本の抽出ではoutlier検出が必要となる．説明変数が過多のとき，すなわち冗長変数が認められる場合には，固有的に必要な説明変数を除き所謂，説明変数の選択が実施されなければならない．これらの根治対策が，費用・時間的などの何らかの理由で実施困難のときは，データ解析の実践では，観測データの重視の立場と相俟ってすべての観測データを活かす観点から，Xの直交変換による縮約回帰，たとえば標準回帰モデルでの直交回帰分析手法，条件付回帰モデルでの主成分回帰分析手法など，あるいは$\hat{\underline{\beta}}$の原点縮小を図るリッヂ回帰手法，そして頑健回帰手法が有用される．

回帰問題への適用

回帰係数の推定問題では，最小二乗推定量が抱える準共線性の多大な悪影響を調整する視点から偏りのある推定量の代替が提案されている．すなわち線形回帰モデルでの回帰係数の推定に対する評価規準として平均平方誤差を採用するとき，Xの非直交性（準特異性をも含む）を考慮に入れる偏りのある推定量（たとえば，リッヂ回帰推定量，主成分回帰推定量など）は，従来の最小二乗推定量の性能を超えることが知られている．

1）最小二乗推定量$\underline{\hat{\beta}}$の改良

Xが準特異あるいは非直交のとき，$\underline{\hat{\beta}}$の改良を意図して偏りのある縮小推定量が有用されているが，ここでは，リッヂ回帰推定量，Schlove推定量をとりあげる．

リッヂ回帰推定量は適当なリッヂ因数Kのもとで

$$\underline{\hat{\beta}}_{(K)}=(X'X+\mathrm{KI})^{-1}X'\underline{Y} \tag{15}$$

と与えられる（Hoerl and Kennard(1970)）．この$\underline{\hat{\beta}}_{(K)}$は$X$が非直交のときは適当なKを有するとき，

$$\mathrm{MSE}(\underline{\hat{\beta}}_{(K)})\leq\mathrm{MSE}(\underline{\hat{\beta}}) \tag{16}$$

となる．また．Schlove推定量は

$$\underline{\hat{\beta}}^{*}=\left(1-\frac{p-2}{n-p+2}\cdot\frac{(n-p)s^{2}}{\hat{\beta}'\hat{\beta}}\right)\underline{\hat{\beta}} \tag{17}$$

と与えられる（schlove(1972)）．この$\underline{\hat{\beta}}^{*}$は$p\geq 3$のとき，

$$\mathrm{MSE}(\underline{\hat{\beta}}^{*})\leq\mathrm{MSE}(\underline{\hat{\beta}}) \tag{18}$$

となる．

2）先験情報の活用を図る予備検定推定量

回帰係数$\underline{\beta}$に関する先験情報を活かす予備検定推定量の評価について，準共線性の問題を考慮して検討する．いま，$\underline{\beta}$の推定に主眼をおく状況を設定する．このとき，先験情報などにより，$\underline{\beta}$に線形制約$H'\underline{\beta}=\underline{h}$が与えられているとする．ここに，$H'$は$m\times p$行列，$\underline{h}$は$m\times 1$ベクトルである．

このとき，$\underline{\hat{\beta}}$の予備検定推定量は線形仮説$H'\underline{\hat{\beta}}=\underline{h}$を帰無仮説とするF検定の結果を通して従来の制約なしの最小二乗推定量$\underline{\hat{\beta}}$が，あるいは制約つきの最小二乗推定量$\underline{\hat{\beta}}_{R}$かのいずれかを選定することにより与えられる．すなわち，$\underline{\hat{\beta}}$の予備検定推定量$\underline{\tilde{\beta}}$は

$$\underline{\tilde{\beta}}=\mathrm{I}_{(0,c)}(\mathrm{F})\,\underline{\hat{\beta}}_{R}+\mathrm{I}_{(c,\infty)}(\mathrm{F})\,\underline{\hat{\beta}} \tag{19}$$

となる．ここに，$\mathrm{I}_{(a,b)}(\mathrm{F})$はFが区間$(a, b)$内のときは1，区間外のときは0をとる指標関数である．cはF検定の結果を分ける臨界点である．また，

$$\underline{\hat{\beta}}_{R}=\underline{\hat{\beta}}-S^{-1}H(H'S^{-1}H)^{-1}(H'\underline{\hat{\beta}}-\underline{h}) \tag{20}$$

$$\mathrm{F}=(H'\underline{\hat{\beta}}-\underline{h})'(H'S^{-1}H)^{-1}(H'\underline{\hat{\beta}}-\underline{h})/ms^{2} \tag{21}$$

である．なお，Fは自由度$(m, n-p)$，非心度

$$\theta=(H'\underline{\beta}-\underline{h})'(H'S^{-1}H)^{-1}(H'\underline{\beta}-\underline{h})/2\sigma^{2} \tag{22}$$

の非心F分布F（m, $n-p$: θ）に従い，仮説$H^T\underline{\beta}=\underline{h}$の下では中心F分布に従う．

Xが非直交のとき，(16)が成立する．この指摘に基づき，仮説$H^T\underline{\beta}=\underline{h}$が受容できないき，$\underline{\hat{\beta}}$の代替として仮説制約のないリッヂ回帰推定量を組み入れる予備検定推定量は

$$\underline{\tilde{\beta}}_{(K)}=I_{(0,c)}(F)\,\underline{\hat{\beta}}_{R}+I_{(c,\infty)}(F)\,\underline{\hat{\beta}}_{(K)} \tag{23}$$

と与えられる（田中・後藤（1982））．このとき，適当なcとkを選択すれば，

$$MSE(\underline{\tilde{\beta}}_{K})\leq MSE(\underline{\tilde{\beta}}) \tag{24}$$

となる．同様に，縮小推定量である主成分回帰推定量の導入も示唆されることになる．

3）影響観測値の検出統計量

回帰分析の結果に影響を与える観測値の検出について，準共線性の問題を検討する．

Cook（1977）は第i観測値の影響を測る距離尺度として，$\underline{\beta}$の信頼楕円体に基づく，交叉確認形式による影響関係（cook距離統計量）

$$D(i)=(X\underline{\hat{\beta}}-X\underline{\hat{\beta}}_{(i)})^T(X\underline{\hat{\beta}}-X\underline{\hat{\beta}}_{(i)})/ps^2 \tag{25}$$

を提示した．ここに，$\underline{\hat{\beta}}_{(i)}$は第$i$観測値を削除したときの$\underline{\hat{\beta}}$である．

いま，Cook距離統計量が最小二乗推定量を基軸にして構成されていることに着目する．Xの非直交性を考慮し，偏りのある推定量としてSchlove推定量$\underline{\hat{\beta}}^*$を導入する．このとき，調整Cook距離統計量は

$$MD(i)=(X\underline{\hat{\beta}}^*-X\underline{\hat{\beta}}^*(i))^T(X\underline{\hat{\beta}}^*-X\underline{\hat{\beta}}^*(i))/ps^2 \tag{26}$$

と与えられる（田中・後藤(1981)，田中(1993a)）．ここに，$\hat{\beta}^*_{(i)}$は第i観測値を除いたときの$\underline{\hat{\beta}}^*$である．この調整は“悪い”観測値の検出・削除のみならず，“良好な”観測値の検出・保存を意図している．同様に，(16)の成立に着目して，縮小推定量でもあるリッヂ回帰推定量$\hat{\beta}_{(K)}$の導入が考えられる．（田中・後藤(1981)，Walker and Birch(1988)）．リッヂ調整のCook距離統計量は

$$MD_K(i)=(X\underline{\hat{\beta}}_{(K)}-X\underline{\hat{\beta}}_{(K)}(i))^T(X\underline{\hat{\beta}}_{(K)}-X\underline{\hat{\beta}}_{(K)}(i))/ps^2 \tag{27}$$

と与えられる．

若干の問題

多様な回帰問題に対し，広範な守備範囲を有する回帰分析においては，その推測が説明変数行列Xに大きく依存する．実験研究と異なり，説明変数行列（実験研究では配置行列）の直交化を計ることが難しい観測研究下のデータでは，Xが非直交となる準共線性の問題に直面せざるを得ない．

ここでは，準共線性の問題に絡む若干の疑問と示唆を提示する．

1）準共線性は悪条件であるか？

多重共線性の問題は悪条件(ill-conditioned)の問題とも呼ばれ，Mason and Gunst(1985)の定義によれば，冗長な説明変数集合の存在を意味するが，$\underline{\beta}$の推定では$\underline{\hat{\beta}}$の不安定さを緩和すべく改良が計られる．この改良の立場から，Xの列方向，p変数の縮約（Reduction）が主として行われることになる．

準共線性は悪条件であるか？　$\underline{\beta}$の最小二乗推定ではサンプリング機構を無視して推定がなされることから，充分に情報をくみとることができないことに注意する．サンプリング状況が歪みをもつ準共線性では最小二乗推定は結果的に歪みを活用できずに推定性能を改良できない悪条件から脱出できないことになる．サンプリング状況の活用を図る縮約では説明変数集合の冗長部分を無視しているが，モデルの同定（identification），モデル内のパラメータの推定，検定（estimation, test）など各推測段階の目標を含む一連の推測構造を考えるとき，サンプリング状況に即した推測目標のもとで改良がなされるべきとする視点に立ち，冗長部分の活用を探ることは重要な試みと考える．たとえば，情報量規準を導入するなど統一的評価の下で1つの新たな推測方式が展開されなければならない．

2）非直交性の表現

Xが準共線性の状態にあるとき，その非直交な度合，すなわち直交性からの乖離の程度を測る適切な尺度が，より効率的な推測を進める立場からも要請される．非直交性の距離化についてはX自体の直交性からの乖離を方向を規定せずには表現できない．この尺度化は回帰目的に沿う推測目標の評価規準に依拠することで意味を有することから，非直交性は評価規準から離れて絶対尺度では測れないことになる．固有値λ_lは$\underline{\beta}$の推定に対し，平均平方誤差の評価規準のもとで意味をもつことになる．その意味で，特異値分解(5)は推測を展開するときに有力な道具であるが，推測目標を考慮に含む一般化された分解公式が望まれることになる．このような観点に基づき，非直交性の表現ができれば，例えば，影響観測値の検出についても有用となる，Cook統計量が残差の関係項と観測点集合からの乖離を示す梃率の2要素の積形式で分解できることの類推から非直交性の乖離を含む表現は個々の観測値の影響力を構成的に把握することを可能にする．グラフィカル表現を加えて，より実効のあるデータ解析が期待できることになる．

3）観測機構と評価方式

準共線性を伴う推測では観測機構に即した評価がなされなければならない．回帰分析の実践では，同時観測データを扱うことが多い．この場合，条件付回帰モデルを想定するが，一般には標準回帰モデルを想定した標準回帰分析が定型的に実行されている．Xが非直交であることを考慮して偏推定量$\underline{\hat{\beta}}_K$を用いるとき，評価規準として有用されている平均平方誤差は条件付回帰モデルを意識せずに，$\mathrm{MSE}(\underline{\hat{\beta}}_K|\ X)$と表記するのではなく$\mathrm{MSE}(\underline{\hat{\beta}}_K)$として用いられている．種々の回帰問題では，たとえばリッヂ回帰でのリッヂ因数K，あるいは予備検定推定量の構成要素である臨界点αの最適選定問題の場合にもXの条件付きの下での評価を考慮しなければならない．

4）準共線性対策としてのデータ分割

準共線性の問題に対する解決策として，すなわちXが非直交なときに効率的な推測を得るための接近として，Xについての縮約化（reduction），偏化（partial），周辺化（marginal）などを図る，所謂Xの列方向での$p \to q\ (\leq p)$の次元縮小の視点と，計画された追加観測もしくは影響観測値，はずれ値の検出・削除のXの行方向での観測データの調整の視点に大きく分けることができる．この接近は概して良好に作動しているが，これらの方向と異なり，得られたXをX^*と変形させることなく，“観測データの活用”の立場からXの分割を考える．いま，X，$\underline{Y}$が

$$X=\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{pmatrix},\qquad \underline{Y}=\begin{pmatrix} \underline{Y}_1 \\ \underline{Y}_2 \\ \vdots \\ \underline{Y}_K \end{pmatrix}$$

とK個に分割することを考える．ここに，$\underline{Y}_1, \cdots, \underline{Y}_K$の分割は$X_1, \cdots, X_K$に対応する．このときの最適な分割は個々の$X_i$，$i=1, \cdots, K$に対する直交行列からの乖離度合，たとえば条件数を評価し，その総和を最小にする．

5）準共線性と非直交

準共線性の問題が惹起するときXは準特異あるいは非直交となると謂われている．準共線性が非直交(準特異)の程度，如何なる出現パターンで生じているかが明らかではない．とくに，条件付回帰モデルを想定する同時観測データの回帰推測では，同時分布の相関行列の規定が重要となる．モンテカルロ実験を通して，特定の相関行列がXに如何なる程度の，そして出現形式の非直交をもたらすことを把握することが準共線性の影響の解消に有用となる．

要　　約

本論では，回帰における多重共線性の問題を準共線性の問題として整理する．準共線性の問題を，Xが準特異あるいは非直交となるときの従来の対応と異なり，すなわち悪条件の問題として位置づけることなく功罪両面について論旨を展開する．とくに，準共線性の特性を積極的に活用することが，モデルの妥当性確認に有用となることを主張する．

本論の考察から，説明変数行列が回帰問題に対し決定的役割を果たすことに留意し，準共線性に絡む若干の疑問と示唆を提示する．

- 準共線性は悪条件であるか？サンプリング状況の歪み，すなわち冗長変数の情報を推測に有用となるべく汲みとるべきである．
- Xが準共線性の状態にあるとき，直交性からの乖離の程度を測る適切な尺度化が効率的な推測を進める意味で重要である．とくに，尺度化においては回帰目的に沿う推測目標の評価規準に依拠することに留意する．
- 準共線性の対処では，Xの列方向での次元縮小とXの行方向での観測データの調整（追加・削除）の二方向に分けることができる．列方向の調整として，“観測データの活用”の立場からデータの分割も準共線性の影響を解消する1つの有効な対策となる．

文　　献

1) Farrar, D F. and Glauber, R. R.: Review of Economics and Statistics, 49 (1967)
2) Mason, R. J, Gunst, R. F. and Webster, J. T · Comm Statist. 4, 277-292 (1975)
3) Cook, R D: Technometrics, 19, 15-18 (1977)
4) Mason, R J and Gunst, R. F.: Technometrics, 27, 401-407 (1985)
5) Press, S. J · Appled Multivariate, Holt, Rinhart and Winston (1972)
6) 田中浩光：名古屋女子大学紀要（家政・自然），**38**, 119-127 (1992)
7) 田中浩光：名古屋女子大学紀要（家政・自然），**39**, 63-70 (1993a)
8) 田中浩光：第61回日本統計学会講演報告集，290-291 (1993b)