行列表現による重回帰分析 (1)

新村 秀一

1. はじめに

広義の多変量解析[1]のなかで，重回帰分析は最も重要なモデルの1つであり，実用性も高い．このため数多くの良書が出版されている．本講座では，理論の記述に適した行列表記を用いて各種統計量を導くとともに，理解しやすい数値例を示して計算手順を示すことにする．

行列表記を用いることの利点は，重回帰分析の全体的な視野に立つ整理ができることである．行列表記に慣れておられない読者も恐れずに慣れることに努力していただきたい．

2. データ

以下のデータは，応答変数 y と x_1 から x_4 までの4個の説明変数からなる7個の観測データである．

y：分娩までの経過時間の自然対数による表示

x_1：子宮口開大度

x_2：陣痛間欠時間

x_3：胎児心拍数

x_4：陣痛持続時間

4個の説明変数はある観測時点において計測され，応答変数はその時点から分娩までの経過時間を示す．次の7個の時系列データは同一母体からのものである．

このデータに，多重共線性の説明に用いる変数 x_5 を追加する．

x_5：x_2+x_4．ただし，最初のデータのみ，この値に2をさらに加える．

No.	y	x_1	x_2	x_3	x_4	x_5
1	7.390	8	29	150	18	49
2	7.300	5	4	144	20	24
3	7.215	7	9	134	30	39
4	7.162	7	18	150	40	58
5	5.193	7	54	130	30	84
6	4.654	7	7	130	30	37
7	2.708	5	8	120	100	108
m	5.946	6.571	18.429	136.857	38.286	57.000
σ	1.813	1.134	17.859	11.423	28.176	29.462

(1)

ここで，データの各列をベクトルとみなし，次の行列を以下の議論で主として用いる．

$$D=(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4) \tag{2}$$

$$X=(\boldsymbol{1},\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4)$$

〔注〕$\boldsymbol{1}$はすべての要素が1の列ベクトル．他の列ベクトルと同じ扱いをするため $\boldsymbol{x}_0$ と表わす． ▲

3. 重回帰モデルの定義とパラメータの推定

重回帰モデルは，変数のレベルで表わすと，応答変数 y，説明変数を $x_i(i=1,\cdots,p)$ と表わして，

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon \tag{3}$$

と表わされる．ここで ε は誤差である．なお，説明変数は確率変数でも決定論的変数でもよいが，確率変数の場合には，その実現値は正確に測定されるものと仮定する．

これをデータのレベルで表わすと式(4)で説明される．

$$y_i=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\cdots+\beta_px_{pi}+\varepsilon_i \quad (i=1,\cdots,n) \tag{4}$$

ここで，n はサンプル数，p は説明変数の個数を示す．

これを，さらに行列表記すれば式(5)になる．

$$\underset{\boldsymbol{y}}{\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix}}=\underset{X}{\begin{bmatrix}1 & x_{11}\cdots x_{p1}\\ \vdots & \vdots\\ 1 & x_{1n}\cdots x_{pn}\end{bmatrix}}\underset{\boldsymbol{\beta}}{\begin{bmatrix}\beta_0\\ \vdots\\ \beta_p\end{bmatrix}}+\underset{\boldsymbol{\varepsilon}}{\begin{bmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_n\end{bmatrix}} \tag{5}$$

ただし，これらのモデル中に現われる誤差 ε_i について，

1) 多変量解析とは相互に相関のある多くの特性値の問題を分析する手法であるので，重回帰分析は特性値が1つしかないことから厳密な定義では多変量には入らない．しかし多変量解析に大いに関係のある分析手法であることは明らかなので，広義では多変量解析の中に入れることもある．

しんむら しゅういち 住商コンピュータサービス㈱

以下の仮定を置く．

i) 不偏性：ε_i の期待値は零である．

$E(\varepsilon_i)=0$

ii) 等分散性：ε_i の分散は i の値によらず一定である．$V(\varepsilon_i)=\sigma^2$

iii) 独立性：誤差 ε_i が互いに独立である．

$\varepsilon_i \perp\!\!\!\perp \varepsilon_j (i \neq j)$

iv) 正規性：誤差は正規分布をする．

以上をまとめると，誤差 ε_i は平均0，分散 σ^2 の正規分布をすることになる．すなわち，$\varepsilon_i \in N(0,\sigma^2)$ になる．行列表記でまとめると，$E(\boldsymbol{\varepsilon})=\mathbf{0}, Var(\boldsymbol{\varepsilon})=E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')=\sigma^2 E$ になる．

［例］今回のデータ(1)を式(5)にあてはめれば，$n=7$, $p=4$ の重回帰モデル $\boldsymbol{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ になる．

$$\begin{bmatrix}7.390\\7.300\\\vdots\\2.708\end{bmatrix}=\begin{bmatrix}1&8&\cdots&18\\1&5&&20\\\vdots&\vdots&&\vdots\\1&5&&100\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1\\\vdots\\\beta_4\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_7\end{bmatrix} \tag{5'}$$

▲

次の誤差平方和 (SSE) を最小にする未知母数 $\boldsymbol{\beta}$ の推定値 $\hat{\boldsymbol{\beta}}$ を求める次の方法を最小二乗法という．

$$\begin{aligned}\mathrm{SSE}&=\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \qquad (6)\\&=(\boldsymbol{y}-X\boldsymbol{\beta})'(\boldsymbol{y}-X\boldsymbol{\beta})\\&=\boldsymbol{y}'\boldsymbol{y}-\boldsymbol{y}'X\boldsymbol{\beta}-\boldsymbol{\beta}'X'\boldsymbol{y}+\boldsymbol{\beta}'X'X\boldsymbol{\beta}\\&=\boldsymbol{y}'\boldsymbol{y}-2\boldsymbol{\beta}'X'\boldsymbol{y}+\boldsymbol{\beta}'X'X\boldsymbol{\beta}\end{aligned}$$

〔注1〕$\boldsymbol{y}'X\boldsymbol{\beta}$ はスカラー量であるので，その転置行列である $\boldsymbol{\beta}'X'\boldsymbol{y}$ と等しくなる．

〔注2〕誤差 $\boldsymbol{\varepsilon}(=\boldsymbol{y}-X\boldsymbol{\beta})$は，最小二乗法で得られた推定値 $\hat{\boldsymbol{\beta}}$ から計算される残差 $\boldsymbol{\mu}(=\boldsymbol{y}-X\hat{\boldsymbol{\beta}})$ と区別すべきだが，本稿では誤差に統一して扱う．▲

SSE を最小にする $\hat{\boldsymbol{\beta}}$ を求めるために，式(6) を $\boldsymbol{\beta}$ で偏微分して零と置く．ベクトル微分を知らない方は〔注3〕を見られよ．

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\beta}}(\mathrm{SSE})&=\frac{\partial}{\partial\boldsymbol{\beta}}(\boldsymbol{y}'\boldsymbol{y}-2\boldsymbol{\beta}'X'\boldsymbol{y}+\boldsymbol{\beta}'X'X\boldsymbol{\beta})\\&=-2X'\boldsymbol{y}+2X'X\boldsymbol{\beta} \qquad (7)\\&=0\end{aligned}$$

この式を満たす $\hat{\boldsymbol{\beta}}$ は極値であるが，最大値か最小値かは次の2階微分で決まる．

$$\begin{aligned}\frac{\partial^2}{\partial\boldsymbol{\beta}^2}(\mathrm{SSE})&=\frac{\partial}{\partial\boldsymbol{\beta}}(-2X'\boldsymbol{y}+2X'X\boldsymbol{\beta}) \qquad (8)\\&=2X'X>0\end{aligned}$$

行列微分において，2階微分が正定値の場合，推定値 $\hat{\boldsymbol{\beta}}$ は最小値になる．$X'X$ が正則の場合，必ず正定値になることは，ここでは天下り的に仮定する(文献[4])．

〔注3〕式(6)を通常の式で表わせば次式になる．

$$\begin{aligned}\mathrm{SSE}&=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\\&=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{1i}-\cdots-\beta_px_{pi})^2\end{aligned}$$

これを $\beta_k(k=1,\cdots,p)$ で微分すれば，

$$\frac{\partial}{\partial\beta_k}(\mathrm{SSE})=\sum_{i=1}^{n}2(y_i-\beta_0-\beta_1x_{1i}\cdots-\beta_px_{pi})(-x_{ki})$$

これを零と置いて得られる p 個の連立方程式の解 β_k は，次の2次微分が正になるので最小値を与える．ただし，すべてのデータは零でない．

$$\frac{\partial^2}{\partial\beta_k{}^2}(\mathrm{SSE})=\sum_{i=1}^{n}2x_{ki}{}^2>0$$

▲

以上から，推定値 $\hat{\boldsymbol{\beta}}$ は次の正規方程式を解いて求まる．

$$X'X\hat{\boldsymbol{\beta}}=X'\boldsymbol{y} \quad \text{（正規方程式）} \tag{9}$$

$$\hat{\boldsymbol{\beta}}=(X'X)^{-1}X'\boldsymbol{y} \quad \text{（解）}$$

〔注4〕実際の重回帰分析のアルゴリズムは，行列 $\left(\begin{array}{c|c}X'X & X'\boldsymbol{y}\\\hline \boldsymbol{y}'X & \boldsymbol{y}'\boldsymbol{y}\end{array}\right)$ の $X'X$ の対角要素を掃き出すことにより，$X'\boldsymbol{y}$ の場所に $\boldsymbol{\beta}$ の推定値が求まる．▲

［例］平方和・積和行列 $X'X$ は次のとおりである．

$$X'X=\begin{array}{c}\\1\\2\\3\\4\\5\end{array}\begin{array}{c}\begin{array}{ccccc}\text{定数項}&x_1&x_2&x_3&x_4\end{array}\\\begin{pmatrix}7&46&129&958&268\\46&310&908&6328&1654\\129&908&4291&17722&4222\\958&6328&17722&131892&35400\\268&1654&4222&35400&15024\end{pmatrix}\end{array} \tag{10}$$

また，$X'X$ の逆行列，行列 $X'\boldsymbol{y}$，推定値 $\hat{\boldsymbol{\beta}}$ は次のとおりである．

$$(X'X)^{-1}=\begin{array}{c}\\1\\2\\3\\4\\5\end{array}\begin{array}{c}\begin{array}{ccccc}\text{定数項}&x_1&x_2&x_3&x_4\end{array}\\\begin{pmatrix}56.936&-1.086&-0.009&-0.328&-0.122\\-1.086&0.243&-0.006&-0.004&0.003\\-0.009&-0.006&7.2\mathrm{E}{-4}&2.6\mathrm{E}{-4}&3.9\mathrm{E}{-5}\\-0.328&-0.004&2.6\mathrm{E}{-4}&0.002&5.9\mathrm{E}{-5}\\-0.122&0.003&3.9\mathrm{E}{-5}&5.9\mathrm{E}{-5}&4.5\mathrm{E}{-4}\end{pmatrix}\end{array}$$

$$X'\boldsymbol{y}=\begin{pmatrix}-5.790\\-0.046\\-0.010\\0.097\\-0.028\end{pmatrix}$$

$$\hat{\boldsymbol{\beta}}=\begin{pmatrix}\hat{\beta}_0\\\hat{\beta}_1\\\hat{\beta}_2\\\hat{\beta}_3\\\hat{\beta}_4\end{pmatrix}=\begin{pmatrix}-5.790\\-0.046\\-0.010\\0.097\\-0.028\end{pmatrix} \tag{9'}$$

▲

データ行列 D の各列から，その列の平均を引きさったものを偏差行列 D_d とよぶことにする．この時，$D_d'D_d$ は偏差平方和積和行列になる．D の各列の平均値を行ベクトル M の要素とすれば，$D'D$ と $D_d'D_d$ の関係は次のとおりになる．

$$D_d'D_d=D'D-nM'M \tag{11}$$

[例] $D'D$ は式(10)で求めた $X'X$ の1行1列を省いたものに等しくなる.

$$7*M'M=\begin{array}{c}\begin{matrix}x_1 & x_2 & x_3 & x_4\end{matrix}\\\begin{bmatrix}302.286 & 847.714 & 6295.430 & 1761.140\\847.714 & 2377.290 & 17654.600 & 4938.860\\6295.430 & 17654.600 & 131109.000 & 36677.700\\1761.140 & 4938.860 & 36677.700 & 10260.600\end{bmatrix}\end{array}$$

よって,

$$D_d'D_d=\begin{bmatrix}7.714 & 60.286 & 32.571 & -107.143\\60.286 & 1913.710 & 67.429 & -716.857\\32.571 & 67.429 & 782.857 & -1277.710\\-107.143 & -716.857 & -1277.710 & 4763.430\end{bmatrix} \tag{11'}$$

▲

これを自由度 $(n-1)$ で割ったものがデータの分散共分散行列 V_d になる.

$$V_d(v_{ij})=D_d'D_d/(n-1) \tag{12}$$

[例] 行列(11')より分散共分散行列は次のとおり.

$$V_d(v_{ij})=\begin{bmatrix}1.286 & 10.048 & 5.429 & -17.857\\10.048 & 318.952 & 11.238 & -119.476\\5.429 & 11.238 & 130.476 & -212.952\\-17.857 & -119.476 & -212.952 & 793.905\end{bmatrix} \tag{12'}$$

▲

この行列の (i,j) 要素 v_{ij} を (i,i) 要素 v_{ii} と (j,j) 要素 v_{jj} の積の平方根で割った $v_{ij}/\sqrt{v_{ii}v_{jj}}$ は変数 x_i と x_j の相関係数 r_{ij} になる. 同様に, D_d の (i,j) 要素を d_{ij} とした場合, $d_{ij}/\sqrt{d_{ii}d_{jj}}$ も r_{ij} になる.

[例] (11')または(12')より次の相関行列 R が求まる.

$$R=\begin{array}{c}\begin{matrix}x_1 & x_2 & x_3 & x_4\end{matrix}\\\begin{bmatrix}1.000 & 0.496 & 0.419 & -0.559\\0.496 & 1.000 & 0.055 & -0.237\\0.419 & 0.055 & 1.000 & -0.662\\-0.559 & -0.237 & -0.662 & 1.000\end{bmatrix}\end{array}$$

▲

以上の行列による表現は, 元のデータ x_i を平均 $\bar{x}_i$ と平方和 $S_{x_ix_i}$ を用いて式(13)で規準化したことに等しい.

$$x_i'=\frac{x_i-\bar{x}_i}{\sqrt{S_{x_ix_i}}} \tag{13}$$

同様にして, y を平均 $\bar{y}$ と平方和 S_{yy} で規準化したものを y' とすれば, 式(3)は次の式(14)になる.

$$y'=\beta_0'+\beta_1'x_1'+\beta_2'x_2'+\cdots+\beta_p'x_p'+\varepsilon' \tag{14}$$

$$\beta_0'=0,\ \beta_i'=\beta_i\frac{\sqrt{S_{x_ix_i}}}{\sqrt{S_{yy}}}$$

〔注〕 変換後の各変数の平均が零より定数項は零になる. また x_i を α 倍すればその係数は $1/\alpha$ 倍になる. ▲

よって, 式(14)の正規方程式と解は次のとおりになる. ただし, D と y はデータ(1)を式(13)で規準化した後のものをあらためて D_1 と y とおく.

$$R\hat{\beta}_1=D'y \quad \text{(正規方程式)} \tag{15}$$

$$\hat{\beta}_1=R^{-1}D'y \quad \text{(解)}$$

このことから, 重回帰分析と重相関分析を一度に行なうことができる(文献[5]).

[例] 規準化データによる重回帰式は式(15)により次式で表わされる.

$$\hat{y}=-0.029x_1-0.097x_2+0.612x_3-0.435x_4 \tag{16}$$

変数 x_1 が他の説明変数と独立であると考えれば, これが1標準偏差動いた時, $\hat{y}$ は -0.029 偏差だけ影響を受ける. ▲

4. 分散分析表

重回帰分析の結果の評価には分散分析表が用いられる[1].

(5)の行列 X を $(p+1)$ 個の n 次元列ベクトル x_i から構成されているものとする.

$$X=(x_0x_1,\cdots,x_p) \tag{17}$$

この列ベクトルで張られる n 次元空間の部分空間 $L(X)$ を考える.

$$L(X)=\{X\alpha=\alpha_0x_0+\alpha_1x_1+\cdots+\alpha_px_p \quad \alpha\in R^{p+1},\ x_i\in R^n\} \tag{18}$$

この時, n 次元空間の点 y から $L(X)$ へ下した垂線の足を図1に示すように $X\hat{\beta}$ とする. この変換行列 Q を y の $L(X)$ への射影行列とよぶことにする.

$$Qy=X\hat{\beta}(=X(X'X)^{-1}X'y) \tag{19}$$

$L(X)$ への垂線は, $y-X\hat{\beta}$ で表わされ, $L(X)$ 内のすべてのベクトルに垂直である.

$$X'(y-X\hat{\beta})=0 \tag{20}$$

これを変形すれば式(9)と同じ正規方程式が得られる.

$$X'X\hat{\beta}=X'y \tag{9''}$$

図1からわかるとおり, 直角三角形に対するピタゴラスの定理を適用すれば, ベクトル y の長さの二乗は, 重回帰モデルの予測値ベクトル $\hat{y}(=X\hat{\beta}=Qy)$ の長さの二乗と誤差ベクトル $e(=y-X\hat{\beta})$ の長さの二乗とに分解される.

$$y'y=\hat{y}'\hat{y}+e'e \tag{21}$$

これを次のような形で表にまとめたものを分散分析表(修正前)とよぶ.

分散分析表(修正前)

	D. F.	平方和	平均平方和	F値
回帰	$p+1$	$\hat{y}'\hat{y}$	$S_1=\hat{y}'\hat{y}/(p+1)$	S_1/S_2
誤差	$n-p-1$	$e'e$	$S_2=e'e/(n-p-1)$	
全体	n	$y'y$		

(22)

1)分散分析表の理解を助けるため, 以下で射影行列(文献2)を導入する. 射影行列 Q は, $Q'=Q, Q^2=Q, QX=X(X\in L(X))$, $\mathrm{rank}\,Q=\mathrm{rank}\,X$ の性質をもつ.

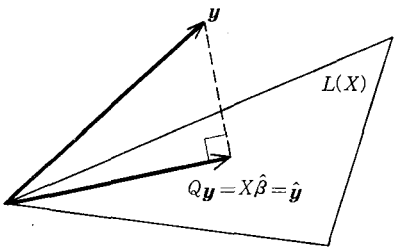

図 1　射影子の幾何学表現

ただし，D. F.は自由度を示し，行列Xの列数が回帰の，行数から列数を引いたものが誤差の自由度を表わす．F値は自由度$(p+1, n-p-1)$のF分布にしたがう．

［例］ $y=\beta_0+\sum_{i=1}^{4}\beta_i x_i+\varepsilon_i$ に対する分散分析表は次のとおり．

	D. F.	平方和	平均平方和	F値
回帰	5	264.706	52.941	42.218*
誤差	2	2.507	1.254	
全体	7	267.213		

(22′)

▲

このF検定は，次の帰無仮説H_0を検定することに等しい．

$$H_0: \beta_0=\beta_1=\cdots=\beta_p=0 \tag{23}$$

この検定は現在考えているモデルが$y=\varepsilon$のモデルと比較して有意か否かの検定であり，当然すぎて有効な情報をもたらさない．そこで，すべての回帰モデルのベースとして次の定数項モデルを考えることにする．

$$\begin{aligned} y_i&=\bar{y}+\varepsilon_i \quad (i=1,\cdots,n) \\ &=\hat{\beta}_0+\varepsilon_i \end{aligned} \tag{24}$$

このモデルに対応する帰無仮説H_0'と対立仮説H_1'は次のとおり．

$$\begin{aligned} &H_0': \beta_0=\bar{y}, \beta_1=\beta_2=\cdots=\beta_p=0 \\ &H_1': \beta_0=\hat{\beta}_0, {}^{\exists}\beta_i\neq 0 \quad (\text{for } i=1,\cdots,p) \end{aligned} \tag{25}$$

これらの関係を図2に示す．すなわち，分散分析表(22)は回帰平方和として$\sum\hat{y}_i{}^2$を表わすのに対し，モデル(24)をベースにした回帰平方和は，$\hat{y}_i$の偏差平方和$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$になる．このことは，分散分析表(22)の回帰平方和と全体の平方和から中心効果$n\bar{y}^2$を差し引き，自由度をpと$(n-1)$に修正した次の分散分析表を求めたことになる．

分散分析表(修正済み)

	D. F.	平方和	平均平方和	F値
回帰	p	$\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}-n\bar{y}^2$	$S_1=(\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}-n\bar{y}^2)/p$	S_1/S_2
誤差	$n-p-1$	$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$	$S_2=\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/(n-p-1)$	
全体	$n-1$	$\boldsymbol{y}'\boldsymbol{y}-n\bar{y}^2$		

(26)

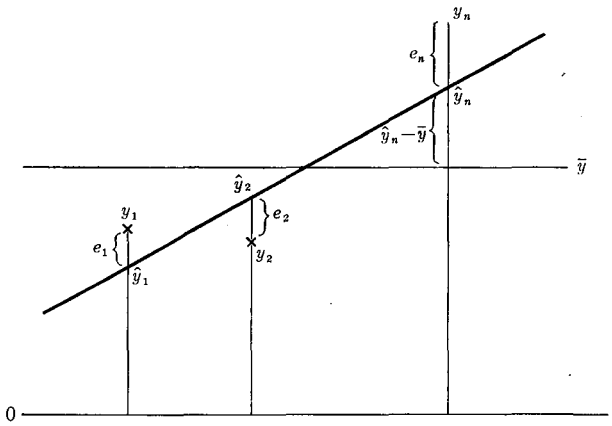

図 2　修正項$n(\bar{y})^2$の幾何学表現

$$R^2=(\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}-n\bar{y}^2)/(\boldsymbol{y}'\boldsymbol{y}-n\bar{y}^2)$$

［例］ $y=\beta_0+\sum_{i=1}^{4}\beta_i x_i+\varepsilon_i$の修正済み分散分析表は$\bar{y}=5.946$として次のとおりになる．

	D. F.	平方和	平均平方和	F値
回帰	4	17.221	4.305	$3.434<F_2^4(0.05)$
誤差	2	2.508	1.254	
全体	6	19.728		

(26′)

$R^2=0.873$ ▲

誤差の平均平方和1.254は，データのバラツキを示す分散σ^2の推定量s^2であるので，その平方根はσの推定量sになる[1)]．

$$s=\sqrt{1.254}=1.120 \tag{27}$$

一方，応答変数yと予測値$\hat{y}$の相関係数は重相関係数とよばれ，その平方は多重決定係数または寄与率とよばれR^2で表わされるが，修正済み回帰平方和と全体平方和の比に等しい．

$$\begin{aligned} R^2&=(\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}-n\bar{y}^2)/(\boldsymbol{y}'\boldsymbol{y}-n\bar{y}^2) \\ &=pS_1/\{pS_1+(n-p-1)S_2\} \end{aligned} \tag{28}$$

このR^2値は，式変形により，平均回帰平方和S_1と平均誤差平方和S_2の比で表わされるので，分散分析表によるF検定と，決定係数R^2に対する検定は型式が違っても本質的に同じであるので，一方を行なえば，他方を行なう必要はない．

5. パラメータの各種統計量

パラメータ$\boldsymbol{\beta}$の推定値$\hat{\boldsymbol{\beta}}$の期待値は次式で与えられる．

1)不偏推定ではない．

$$E(\hat{\boldsymbol{\beta}})=E((X'X)^{-1}X'\boldsymbol{y}) \tag{29}$$
$$=(X'X)^{-1}X'E(\boldsymbol{y})$$
$$=(X'X)^{-1}X'E(X\boldsymbol{\beta}+\boldsymbol{\varepsilon})$$
$$=(X'X)^{-1}X'X\boldsymbol{\beta}$$
$$=\boldsymbol{\beta}$$

$\boldsymbol{y}$ の分散行列 $\mathrm{Var}(\boldsymbol{y})$は，$\varepsilon_i\sim N(0,\sigma^2)$と$\varepsilon_i \perp\!\!\!\perp \varepsilon_j (i\neq j)$であるので，次式になる．

$$\mathrm{Var}(\boldsymbol{y})=E((\boldsymbol{y}-X\boldsymbol{\beta})(\boldsymbol{y}-X\boldsymbol{\beta})') \tag{30}$$
$$=E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$$
$$=\sigma^2E$$

推定値 $\hat{\boldsymbol{\beta}}$ の分散行列は，次式になる．

$$\mathrm{Var}(\hat{\boldsymbol{\beta}})=\mathrm{Var}((X'X)^{-1}X'\boldsymbol{y}) \tag{31}$$
$$=(X'X)^{-1}X'\cdot\mathrm{Var}(\boldsymbol{y})\cdot X(X'X)^{-1}$$
$$=(X'X)^{-1}X'X(X'X)^{-1}\sigma^2$$
$$=(X'X)^{-1}\sigma^2$$

［例］ σ^2 は平均誤差平方和 $s^2=1.254$ により推定されるので，$(X'X)^{-1}s^2$ が $\mathrm{Var}(\hat{\boldsymbol{\beta}})$の推定値になる．

$$\mathrm{Var}(\hat{\boldsymbol{\beta}})=\begin{array}{c}\begin{matrix}\text{定数項} & x_1 & x_2 & x_3 & x_4\end{matrix}\\ \begin{pmatrix} 71.385 & -1.361 & -0.012 & -0.411 & -0.153\\ -1.361 & 0.305 & -0.008 & -0.005 & 0.004\\ -0.012 & -0.008 & 9.1\text{E-}4 & 3.3\text{E-}4 & 4.9\text{E-}5\\ -0.411 & -0.005 & 3.3\text{E-}4 & 0.003 & 7.4\text{E-}4\\ -0.153 & 0.004 & 4.9\text{E-}5 & 7.4\text{E-}4 & 5.7\text{E-}4\end{pmatrix}\end{array} \tag{31'}$$

この(ij)要素を，(ii)要素と(jj)要素の積の平方根で割って，推定値 $\hat{\boldsymbol{\beta}}$ の相関行列 $R(\hat{\boldsymbol{\beta}})$ が求まる．

$$R(\hat{\boldsymbol{\beta}})=\begin{array}{c}\begin{matrix}\text{定数項} & x_1 & x_2 & x_3 & x_4\end{matrix}\\ \begin{pmatrix} 1.000 & -0.292 & -0.047 & -0.890 & -0.758\\ -0.292 & 1.000 & -0.468 & -0.161 & 0.333\\ -0.047 & -0.468 & 1.000 & 0.198 & 0.068\\ -0.890 & -0.161 & 0.198 & 1.000 & 0.569\\ -0.758 & 0.333 & 0.068 & 0.569 & 1.000\end{pmatrix}\end{array}$$

参考として，モデル $y=\beta_0+\sum_{i=1}^{5}\beta_i x_i+\varepsilon$ での推定値 $\hat{\boldsymbol{\beta}}$ の相関行列は次のようになる．

$$\begin{array}{c}\begin{matrix}\text{定数項} & x_1 & x_2 & x_3 & x_4 & x_5\end{matrix}\\ \begin{pmatrix} 1.000 & -0.452 & -0.541 & -0.910 & -0.558 & 0.540\\ -0.452 & 1.000 & 0.408 & 0.055 & 0.433 & -0.424\\ -0.541 & 0.408 & 1.000 & 0.446 & 0.999 & -0.999\\ -0.910 & 0.055 & 0.446 & 1.000 & 0.454 & -0.440\\ -0.558 & 0.433 & 0.999 & 0.454 & 1.000 & -0.9996\\ 0.540 & -0.424 & -0.999 & -0.440 & -0.9996 & 1.000\end{pmatrix}\end{array}$$

両相関行列を対比してわかることは，x_5 と x_2, x_4 が高い相関をもつのは当然として，x_5 をモデルに入れたことにより x_2 と x_4 の間にも高い相関が認められるようになった． ▲

$(X'X)^{ii}$ を $(X'X)^{-1}$ の i 番目の対角要素とすれば，$\hat{\beta}_i$ の標準偏差 $\mathrm{stderr}(\hat{\beta}_i)$ と t 統計量は次式で与えられる．

$$\mathrm{stderr}(\hat{\beta}_i)=\sqrt{(X'X)^{ii}s^2} \tag{32}$$
$$t=\hat{\beta}_i/\mathrm{stderr}(\hat{\beta}_i)$$

［例］ 式(31′)と式(32)から，$\hat{\boldsymbol{\beta}}$ の標準偏差と t 値は次のとおりになる．

$$\mathrm{stderr}(\hat{\boldsymbol{\beta}})=\begin{pmatrix}8.449\\0.552\\0.030\\0.055\\0.024\end{pmatrix}\quad t(\hat{\boldsymbol{\beta}})=\begin{array}{cc}\begin{pmatrix}-0.685\\-0.084\\-0.327\\1.778\\-1.175\end{pmatrix} & \begin{matrix}\text{定数項}\\ \hat{\beta}_1\\ \hat{\beta}_2\\ \hat{\beta}_3\\ \hat{\beta}_4\end{matrix}\end{array}$$ ▲

6. 多重共線性(multi-collinearity)

ある説明変数が他の説明変数の1次結合でほぼ表わされる時，$\hat{\boldsymbol{\beta}}$ は確定的でなく，多重共線性をもつ．

この時，次の好ましくない情況が発生する（文献[4] pp. 183–184）．

① 推定値は，データの小さな変化に対して不安定である．

② 推定値は大きな標準誤差をもつ．このため，t 検定が棄却できないことが多い．

多重共線性の検出方法としては，リッジ回帰分析（文献[3](pp. 201–206)），主成分分析(文献[3])，分散拡大要因(Variance Inflation Factor, VIF)等がある．これらの方法を以下に解説しよう．

なお，多重共線性が検出された場合，対応としてはバラツキの弱い次元に広く分布するデータを追加するか，多重共線関係にある変数のいくつかをモデルから省くという2つの方法が考えられる．

6.1 分散拡大要因(VIF)

$\hat{\beta}_i$ の VIF_iは，x_iを応答変数として残りのすべての説明変数で回帰して得られる多重決定係数 R_i^2 を用いて次式で表わされる．

$$\mathrm{VIF}_i=1/(1-R_i^2) \tag{33}$$

一応の目安として，VIF が 10 以上の場合に多重共線性が疑われる(文献[3] pp. 201–202)．

［例］ 説明変数が x_1, x_2, x_3, x_4 の4変数の場合，モデル $x_1=\beta_0+\beta_2x_2+\beta_3x_3+\beta_4x_4+\varepsilon$ の決定係数を $R_{1,234}^2$ とすれば，x_1 の分散拡大要因は $\mathrm{VIF}_{1,234}=1/(1-R_{1,234}^2)$ になる．同様に x_2, x_3, x_4 の VIF も計算される．

$$\mathrm{VIF}_{1,234}=1.875$$
$$\mathrm{VIF}_{2,134}=1.385$$
$$\mathrm{VIF}_{3,124}=1.863$$
$$\mathrm{VIF}_{4,123}=2.156$$

になる．

多重共線性のない4個の説明変数の組に，x_5 を追加す

れば,

$$VIF_{1,2345}=\quad 2.287$$
$$VIF_{2,1345}=1008.260$$
$$VIF_{3,1245}=\quad 2.309$$
$$VIF_{4,1235}=2534.162$$
$$VIF_{5,1234}=2724.858$$

と，多重共線関係にある x_2, x_4, x_5 の分散拡大要因は極端に大きくなる．▲

以上みたように多重共線関係にある説明変数の検出は容易に行なえる．しかし，その対応策として，どの変数をどのような基準にもとづいて何個省けばよいかの問題が残る．これを，かりに“多重共線性の解消”問題とよぶが，これは統計論的に決めるべき問題ではなく，その問題の専門分野の知識を参考にして決めるべきであろう．

$\hat{\beta}_i$ の各 VIF_i の値は，$(X'X)^{-1}$ の各 i 番目の対角要素 $(X'X)^{ii}$ の値と比例関係にある．この $(X'X)^{ii}$ は式(32)からわかるとおり，分散 s^2 を $(X'X)^{ii}$ 倍に拡大したものが $\hat{\beta}_i$ の分散になることを示しているので，分散拡大要因とよばれる．

［例］次の簡単なデータを考える．

y	x_1	x_2
1	−1	1
2	0	1
3	1	0
3	2	2

(34)

モデル $y=a_0+a_1x_1+a_2x_2+\varepsilon$ に対して，

$$(X'X)^{-1}=\begin{pmatrix} \frac{3}{4} & 0 & -\frac{1}{2} \\ 0 & \frac{2}{9} & -\frac{1}{9} \\ -\frac{1}{2} & -\frac{1}{9} & \frac{5}{9} \end{pmatrix} \qquad (35)$$

$$\boldsymbol{a}=\left(\frac{9}{4}\quad \frac{7}{9}\quad -\frac{7}{18}\right)'$$

	D. F.	平方和	平均平方和	F
回帰	2	$2\frac{13}{18}$	$\frac{49}{36}$	49
誤差	1	$\frac{1}{36}$	$\frac{1}{36}$	
全体	3	$2\frac{3}{4}$		

$$R^2=\frac{98}{99}$$

モデル $x_1=b_0+b_1x_2+\varepsilon$ に対して，

$$(X'X)^{-1}=\begin{pmatrix} \frac{3}{4} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

$$\boldsymbol{b}=\left(0\quad \frac{1}{2}\right)'$$

	D. F.	平方和	平均平方和	F
回帰	2	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{18}$
誤差	1	$\frac{9}{2}$	$\frac{9}{2}$	
全体	3	5		

$$R^2=\frac{1}{10}$$

$$VIF_{x_1}=1/(1-\tfrac{1}{10})=\frac{10}{9} \qquad (36)$$

モデル $x_2=c_0+c_1x_1+\varepsilon$ に対して，

$$(X'X)^{-1}=\begin{pmatrix} \frac{3}{10} & -\frac{1}{10} \\ -\frac{1}{10} & \frac{1}{5} \end{pmatrix}$$

$$\boldsymbol{c}=\left(\frac{9}{10}\quad \frac{1}{5}\right)'$$

	D. F.	平方和	平均平方和	F
回帰	2	3.2	1.6	$\frac{8}{9}$
誤差	1	1.8	1.8	
全体	3	5.0		

$$R^2=0.64$$

$$VIF\ x_2=1/(1-0.64)=\frac{25}{9} \qquad (37)$$

(35), (36), (37)より，

$$VIF_{x_1}:VIF_{x_2}=(X'X)^{22}:(X'X)^{33}=2:5$$ ▲

6.2 主成分分析の利用

主成分分析は，データが多変量正規分布すなわち確率楕円にしたがうとして，元の変数の作る旧座標系を座標変換により楕円の軸を新座標系として求める手法である．

各説明変数を，平均0（原点移動）と分散1（単位系の違い等による影響を除くため）に規準化したデータ行列 D を考える．この行列の列数（説明変数の数）を p，行数（データ数）を n とする．ここで p 個の重みベクトル $\boldsymbol{a}=(a_1\cdots,a_p)'$ による次の座標変換を考える．

$$\boldsymbol{z}=D\boldsymbol{a} \qquad (38)$$

D の i 行は旧座標系での観測値 i の p 個の座標 D_i であり，$D_i\boldsymbol{a}$ は観測値 i の新座標軸 $\boldsymbol{a}$ での座標を与えるスカラー値である．$\boldsymbol{z}$ はこの新座標系 $\boldsymbol{a}$ での n 個の観測値の新座標値になる．この分散 V_z は，D が規準化されて

いることから次式で表わされ，さらにデータの相関行列を R として次式になる．

$$V_z=\frac{1}{n}z'z=\frac{1}{n}a'D'Da \tag{39}$$
$$=a'(\frac{1}{n}D'D)a=a'Ra$$

ここで，$a'a=1$ の条件で V_z を最大にすることを考える．条件つき極値問題になるので，ラグランジェの未定乗数を λ として，次の φ を最大にする a を求めればよい．

$$\varphi=a'Ra-\lambda(a'a-1) \tag{40}$$

$$\frac{\partial\varphi}{\partial a}=2Ra-\lambda(2a)=0 \tag{41}$$

式(41)は，相関行列 R の固有値問題になる．

$$(R-\lambda E)a=0 \tag{42}$$

ただし，ここで E は単位行列，λ は固有値，a は固有ベクトルである．

一方，$Ra=\lambda a$ の両辺の左側に，a' を乗じれば，

$$V_z=a'Ra=\lambda a'a=\lambda \tag{43}$$

となり，固有値 λ は座標 a でのデータの分散を与える．

相関行列 R の階数が p なら，p 組の固有値 λ_i と固有ベクトル a_i が求まる．固有値の大小順に並べかえて $\lambda_1,\cdots,\lambda_p$ とする．対応する固有ベクトル $a_1,\cdots,a_p$ は，第1主成分軸，…，第 p 主成分軸とよばれる新座標系の係数を与える．このようにして求まった p 個の新座標系で，元のデータ D_i は新座標 $(D_ia_1,\cdots,D_ia_p)$ に変換される．

もし $\lambda_p\fallingdotseq 0$ ならば，第 p 主成分軸上のデータ $D_ia_p(i=1,\cdots,n)$ の分散がほぼ零になり，D_ia_p は一定値とみなせる．元の変数の期待値は零に規準化してあるので，これの合成変数の実現値 D_ia_p の期待値も零になる．すなわち，元の i 番目の変数を x_i とすれば，$a_{1p}x_1+a_{2p}x_2+\cdots+a_{pp}x_p\fallingdotseq 0$ という関係式が求まる．この式が変数 $x_1,\cdots,x_p$ の間の多重共線関係を与えるが，小さな値をもつ a_{ip} を零とみなせば特定の変数間の強い多重共線性を検出できる(文献[3] p.179)．

[例] x_1 から x_5 までの5変数データを主成分分析して，次の固有値が得られた．2.683, 1.526, 0.425, 0.367, 1.6E−4．すなわち，第4主成分までで，全分散の99.9%が説明できる．第5主成分から次の多重共線関係が求まる．

$$0.00010x_1+0.00507x_2+0.00011x_3 \tag{44}$$
$$+0.00803x_4-0.00833x_5\fallingdotseq 0$$

小数第4位以下を零とみなせば次式が求まる．

$$0.00507x_2+0.00803x_4-0.00833x_5\fallingdotseq 0 \tag{45}$$

変数 x_5 の作成過程から次式(46)が期待される．

$$x_2+x_4-x_5\fallingdotseq 0 \tag{46}$$

しかし，実際には式(45)になったのは，データ数が少ないため最初のデータに加えられたバイアスの影響と，データが多変量正規分布から乖離しているためと考えられる．▲

参考文献

1) N.ドレイパー他：応用回帰分析，森北出版，1968
2) 石井吾郎：実験計画法の基礎，サイエンス社，1972
3) S.チャタジー他：回帰分析の実際，新曜社，1981
4) J.ジョンストン：計量経済学の方法，東洋経済新報社，1975
5) 小林龍一：相関・回帰分析法入門，日科技連，1972
6) SAS ユーザーズガイド，SAS Inc., 1982
7) G. E. P. Box & G.M. Jenkins : Time series analysis (forecasting and control), Holden-Day (1970)
8) 新村秀一：多重共線関係の解消とその影響，1983年度ＯＲ学会春季研究発表会，156/157
9) Belsley, D. A., Kuh, E., and Welsch, R.E. (1980) : Regression Diagnostics, New York, John Wiley & Sons
10) Cook, R.D. : Detection of Influential Observations in Linear Regression, Technometrics, 19, 15-18(1977)
11) 竹内　啓：現象と行動のなかの統計数理，新曜社，1972
12) 坂元慶行，石黒真木夫，北川源四郎：情報量統計学，共立出版社，1983

—— ・ —— ・ —— ・ ——

次号の内容は次のとおりです．

7. 平均予測値の分散と信頼区間
8. 観測値 y_i の分散と信頼区間
9. y の予測値と誤差の期待値・分散
10. 誤差（残差）の検討
11. モデルの決定と検定
11.1 フルモデルと縮小モデル
11.2 F検定
11.3 AIC規準と C_p 統計量
11.4 総当り法
11.5 逐次変数選択法
11.6 最終モデルの決定

本稿の作成に際し，小林龍一先生に査読いただき，原稿の不備を指摘していただいた．ここに記して厚くお礼申し上げます．