講　座

行列表現による重回帰分析 (2)

新村　秀一

7. 平均予測値の分散と信頼区間

各々のデータの平均予測値（誤差 ε_i を無視する）とその分散は，個々のデータを $X_i=(1, x_{1i}, x_{2i}, \cdots, x_{pi})$ として次式になる．

$$\hat{y}_i=X_i\hat{\boldsymbol{\beta}} \quad (i=1,\cdots,n) \tag{47}$$

$$\begin{aligned}\mathrm{Var}(\hat{y}_i)&=X_i\,\mathrm{Var}(\hat{\boldsymbol{\beta}})X_i' \\ &=X_i(X'X)^{-1}\sigma^2X_i' \\ &=X_i(X'X)^{-1}X_i'\sigma^2\end{aligned} \tag{48}$$

よって，各予測値 $\hat{y}_i$ の標準偏差は $\sqrt{X_i(X'X)^{-1}X_i'S^2}$ になる．$\hat{y}_i$ に対する $(1-\alpha)$ 信頼区間は，0.5α 水準 t 統計量を $t_{\frac{\alpha}{2}}$ とすれば，上下限信頼区間は次式になる．

$$\begin{aligned}UM_i&=\hat{y}_i+t_{\frac{\alpha}{2}}\sqrt{X_i(X'X)^{-1}X_i'S^2} \\ LM_i&=\hat{y}_i-t_{\frac{\alpha}{2}}\sqrt{X_i(X'X)^{-1}X_i'S^2}\end{aligned} \tag{49}$$

区間 (LM_i, UM_i) は，平均予測値の信頼区間とよばれ，$(1-\alpha)$ の確率で $\hat{y}_i$ はこの区間に含まれる．

［例］$S^2=1.254$，自由度 2 の $t_{0.025}=4.303$ を用いて，95%信頼限界を求めると次のようになる．

y_i	$\hat{y}_i$	$\sqrt{X_i(X'X)^{-1}X_i'S^2}$	LM_i	UM_i
7.390	7.620	0.809	4.139	11.101
7.300	7.366	1.111	2.584	12.148
7.215	5.973	0.695	2.983	8.963
7.162	7.159	0.858	3.467	10.851
5.193	5.142	1.104	0.391	9.893
4.654	5.605	0.866	1.876	9.333
2.708	2.757	1.091	−1.936	7.450

これは，母回帰式 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_3+\hat{\beta}_4x_4$（式(9′)）の95%信頼区間を表わす．次章では，各観測値 y_i の95%信頼区間を考える．

8. 観測値 y_i の分散と信頼区間

個々の観測値 y_i の分散は，$\hat{\boldsymbol{\beta}}$ と $\boldsymbol{\varepsilon}$ が独立であることを考慮して次式になる．

$$y_i=X_i\hat{\boldsymbol{\beta}}+\varepsilon_i \quad (i=1,\cdots,n) \tag{50}$$

$$\begin{aligned}\mathrm{Var}(y_i)&=\mathrm{Var}(X_i\hat{\boldsymbol{\beta}})+\mathrm{Var}(\varepsilon_i) \\ &=X_i(X'X)^{-1}X'_iS^2+S^2\end{aligned} \tag{51}$$

よって，y_i の $(1-\alpha)$ 信頼区間は次式になる．

$$\begin{aligned}UI_i&=\hat{y}_i+t_{\frac{\alpha}{2}}\sqrt{X_i(X'X)^{-1}X_i'S^2+S^2} \\ LI_i&=\hat{y}_i-t_{\frac{\alpha}{2}}\sqrt{X_i(X'X)^{-1}X_i'S^2+S^2}\end{aligned} \tag{52}$$

この信頼区間 (LI_i, UI_i) は，当然のことながら，母回帰モデルの信頼区間 (LM_i, UM_i) を含む．

9. $\boldsymbol{y}$ の予測値と誤差の期待値・分散

$\boldsymbol{y}$ の予測値 $\hat{\boldsymbol{y}}$ と誤差 $\boldsymbol{\varepsilon}$ は次式になる．

$$\hat{\boldsymbol{y}}=X\hat{\boldsymbol{\beta}}=X(X'X)^{-1}X'\boldsymbol{y}=Q\boldsymbol{y} \tag{53}$$

$$\begin{aligned}\boldsymbol{\varepsilon}&=\boldsymbol{y}-\hat{\boldsymbol{y}}=\boldsymbol{y}-X\hat{\boldsymbol{\beta}} \\ &=(E-X(X'X)^{-1}X')\boldsymbol{y}=(E-Q)\boldsymbol{y}\end{aligned} \tag{54}$$

（注）$\boldsymbol{y}$ のそれ自身への射影行列を Q_y とすれば，$Q_y\boldsymbol{y}=\boldsymbol{y}$ より，$Q_y=E$ が直観的にわかる．一方，式(53)と(54)より各射影行列は次の恒等式を満たす．

$$E=Q+(E-Q) \tag{55}$$

$\boldsymbol{y}$ の平方和が，$\hat{\boldsymbol{y}}$ の平方和と $\boldsymbol{\varepsilon}$ の平方和の直和に分解されたのと対応して，$\boldsymbol{y}$ 自身の射影子も，$\hat{\boldsymbol{y}}$ 空間への射影子 Q と誤差空間への射影子 $(E-Q)$ に分解されることがわかる．◢

$\hat{\boldsymbol{y}}$ と $\boldsymbol{\varepsilon}$ の期待値および分散は次式になる．

$$E(\hat{\boldsymbol{y}})=E(X\hat{\boldsymbol{\beta}})=XE(\hat{\boldsymbol{\beta}})=X\boldsymbol{\beta} \tag{56}$$

$$\begin{aligned}E(\boldsymbol{\varepsilon})&=E(\boldsymbol{y}-X\hat{\boldsymbol{\beta}})=E(\boldsymbol{y})-E(X\hat{\boldsymbol{\beta}}) \\ &=X\boldsymbol{\beta}-X\boldsymbol{\beta}=\boldsymbol{0}\end{aligned} \tag{57}$$

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{y}})&=\mathrm{Var}(X(X'X)^{-1}X'\boldsymbol{y}) \\ &=X(X'X)^{-1}X'\cdot\mathrm{Var}(\boldsymbol{y})\cdot X(X'X)^{-1}X' \\ &=X(X'X)^{-1}X'\sigma^2=Q\sigma^2\end{aligned} \tag{58}$$

$$\begin{aligned}\mathrm{Var}(\boldsymbol{\varepsilon})&=\mathrm{Var}((E-Q)\boldsymbol{y}) \\ &=(E-Q)^2\mathrm{Var}(\boldsymbol{y})=(E-Q)\sigma^2\end{aligned} \tag{59}$$

（注）ただし，$E-Q$ は誤差空間への射影行列であり，

しんむら　しゅういち　住商コンピュータサービス㈱

巾等行列であるから，巾等行列の性質$Q^2=Q$，$Q=Q'$を用いた．◢

10. 誤差(残差)の検討

重回帰モデルでは，誤差εが$E(\varepsilon)=0$，$E(\varepsilon\varepsilon')=\sigma^2E$を満たすことを前提としている．そしてこれらの仮定の妥当性を調べるため，誤差を各説明変数等と対にした種々の誤差プロット[1]の検討が重要視されている．誤差がこれらの仮定を満たしていない場合としては，大別して次の3通りが考えられる．

① 誤差ε_iに一定のパターンが認められる場合

一定のパターンをもつ誤差の検出法としては，プロット図とダービン・ワトソン統計量による方法とがある．

前者の例としては，誤差εを特定の説明変数xに対しプロットして放物線等の一定パターンが認められた場合，モデルにx^2の説明変数を追加すればよい．またxが四半期等の時間因子を表わし，誤差が四半期の違いにより層別されるならば，四半期の違いを示すダミー変数をモデルに追加すればよい．

ダービン・ワトソン統計量は，回帰分析における系列相関を検出する．誤差ε_tが1階の自己回帰過程，

$$\varepsilon_t=\rho\varepsilon_{t-1}+n_t,\ |\rho|<1 \tag{60}$$

にしたがうという仮定にもとづいている．ここで，n_tは$E(n_t)=0$, $E(n_t'n_t)=\sigma^2$にしたがう．ダービン・ワトソン統計量dは，

$$d=\frac{\sum_2^n(\varepsilon_t-\varepsilon_{t-1})}{\sum_2^n\varepsilon_t^2} \tag{61}$$

で定義され，帰無仮説$H_0(\rho=0)$を対立仮説$H_1(\rho>0)$に対して検定するため用いられる．$\rho=0$ならば$\varepsilon_t=n_t$となり，εは誤差の仮定を満たすことになる．

1階の自己相関係数ρの推定値は，

$$\hat{\rho}=\frac{\sum_2^n\varepsilon_t\cdot\varepsilon_{t-1}}{\sum_2^n\varepsilon_t^2} \tag{62}$$

で与えられる．dとの間に次の近似式が成り立つ．

$$d\fallingdotseq 2(1-\hat{\rho}) \tag{63}$$

この式から，dは0から4までの値をとることがわかる．$\hat{\rho}=0$で$d\fallingdotseq 2$，$\hat{\rho}=1$で$d\fallingdotseq 0$である．dの値が2に近いほど，ε_tに系列相関がないと言える．このため，ダービン・ワトソンの数表に記載された有意水準(d_L, d_U)を用いて次の検定が行なわれる．

(i) $d<d_L$ならばH_0を棄却する．

(ii) $d>d_U$ならばH_0を棄却しない．

(iii) $d_L<d<d_U$ならば判定不能である．

d統計量が有意な値を示したときは，重回帰モデルに必要な説明変数の欠落が考えられる．これを追加すれば見かけの系列相関がなくなることが多い．一方，真の系列相関がある場合，応答変数yと説明変数xを，$(y_t-\rho y_{t-1})$と$(x_t-\rho x_{t-1})$で変換すればよい．

d統計量の欠点は，2階以上の自己相関を検出できない点にある．これに対しては，種々の誤差プロット図の検討が必要となる．

[例] モデル$y=\beta_0+\sum_{i=1}^{4}\beta_ix_i$で，$d=2.029$, $\hat{\rho}=-0.026$である．$n=15$でモデルのパラメータ数4の有意水準$(d_L, d_U)=(0.69, 1.97)$より，かりにデータ数が7でなく15とした場合，棄却できないことになる．(nが14以下のものは検定できない) ◢

② 誤差が等分散性の仮定を満たさない場合を分散不均一性(heteroscedasticity)とよぶ．この場合通常の最小二乗法による推定値は，不偏ではあるが分散は最小にはならない．データは何らかの重みづけにより変換し，ある種の加重最小二乗法を適用すればよい(文献[3] pp. 108-133)．

③ 特定のデータにかなり大きな誤差が認められた場合，そのデータにもどって詳細な検討が必要である．原因が明確な場合にはダミー変数の導入が考えられる．

プロット図により以上の①②③のパターンの検討が行なえるが，特に③に対しては以下に述べるスチューデント化された誤差(誤差を標準偏差で割ったもの)の詳細な検討が必要である．個々の誤差の分散は式(59)より次式で与えられる．

$$\mathrm{Var}(\varepsilon_i)=(1-X_i(X'X)^{-1}X_i')\sigma^2 \tag{64}$$

スチューデント化された誤差はこれを用い次式になる．

$$\varepsilon_i^s=\varepsilon_i/\sqrt{(1-X_i(X'X)^{-1}X_i')S^2} \tag{65}$$

このε_i^sはスチューデントのtに近似される．この値の大きなデータの悪影響度を調べる方法として次の3尺度(文献[9])がある．

第1の尺度は，この値の大きなデータを1件落としてモデルを再計算する．新しく得られた推定値と元の推定値を比較する．i番目のデータを落とした後で計算される統計量を元の統計量の後にカッコ付(i)で表わす．

$\hat{\beta}_j(i)$ j番目の回帰係数

$S^2(i)$ 平均誤差平方和

$\hat{y}_i(i)$ 予測値$X_i\hat{\boldsymbol{\beta}}(i)$

$\varepsilon_i^s(i)$ 式(65)でS^2の代りに$S^2(i)$を用いる

$(X'X)^{jj}$ $(X'X)^{-1}$の(jj)要素

この時，i番目のデータの欠落による回帰係数への影響

1)標準化誤差を縦軸に，横軸にはyの予測値$\hat{y}$，説明変数x_i，観測値の得られた時刻t等をとればよい．

を次式で計ることにする.

$$\tilde{\beta}_j=(\hat{\beta}_j-\hat{\beta}_j(i))/\sqrt{S^2(i)(X'X)^{jj}} \tag{66}$$

この値を検討することにより,$\hat{\beta}_j$ が i 番目のデータから強い影響を受けているかどうか決めることができる.

[例] データ(34)でモデル $y=\beta_0+\beta_1x_1+\varepsilon$ を考える.

$$(X'X)^{-1}=\begin{pmatrix}0.3 & -0.1\\ -0.1 & 0.2\end{pmatrix} \tag{67}$$

$$\hat{\beta}=(\hat{\beta}_0\hat{\beta}_1)'=(1.9\quad 0.7)' \tag{68}$$

$$\hat{y}=(1.2\quad 1.9\quad 2.6\quad 3.3)' \tag{69}$$

$$\varepsilon=(-0.2\quad 0.1\quad 0.4\quad -0.3)' \tag{70}$$

	D. F.	平方和	平均平方和	F
回帰	1	2.45	2.45	16.33
誤差	2	0.30	0.15	
全体	3	2.75		

(71)

$R^2=0.891$

次に4番目のデータを省いて考える.

$$(X'(4)X(4))^{-1}=\begin{pmatrix}\frac{1}{3} & 0\\ 0 & \frac{1}{2}\end{pmatrix} \tag{72}$$

$$\hat{\beta}(4)=(\hat{\beta}_0(4),\hat{\beta}_1(4))'=(2,1)' \tag{73}$$

$$\hat{y}(4)=(1\quad 2\quad 3\quad 4)' \tag{74}$$

$$\varepsilon(4)=(0\quad 0\quad 0\quad 1)' \tag{75}$$

	D. F.	平方和	平均平方和	F
回帰	2	1	0.5	0.5
誤差	1	1	1	
全体	3	2		

(76)

$R^2=0.500$

(67), (68), (73), (76)より,

$$\tilde{\beta}_1=(\hat{\beta}_1-\hat{\beta}_1(4))/\sqrt{S^2(4)(X'X)^{22}} \tag{77}$$
$$=(0.7-1)/\sqrt{1\cdot 0.2}$$
$$=-0.671$$ ◢

第2の尺度は,次式で示す予測値に対する影響である.

$$\tilde{y}_i=(\hat{y}_i-\hat{y}_i(i))/\sqrt{X_i(X'X)^{-1}X_i'S^2(i)} \tag{78}$$

[例] (67), (69), (74), (76)より,

$$\tilde{y}_4=(\hat{y}_4-\hat{y}_4(4))/\sqrt{X_4(X'X)^{-1}X_4'S^2(4)} \tag{79}$$
$$=(3.3-4)/\sqrt{0.7\cdot 1}$$
$$=-0.837$$ ◢

第3の診断尺度は,データ空間の次元が主として1つのデータに支えられているなら,それを省いた場合の $X(i)'X(i)$ は非正則に近くなる.すなわち $\det((X(i)'X(i))^{-1})$ は大きくなる.次の Covratio 統計量は,i 番目の観測値を削除した結果,$\hat{\beta}$ の共分散行列の行列式の変化率を示す.

$$\text{Covratio}=\det(\text{Cov}(\hat{\beta}(i)))/\det(\text{Cov}(\hat{\beta})) \tag{80}$$
$$=\det(S^2(i)(X(i)'X(i))^{-1})/\det(S^2(X'X)^{-1})$$

この値は,$\det(X'X)/\det(X(i)'X(i))$ で近似できる.

[例] データ(34)でモデル $y=\beta_0+\beta_1x_1+\varepsilon$ を考える.

$$\det(X'X)=\begin{vmatrix}4 & 2\\ 2 & 6\end{vmatrix}=24-4=20$$

$$\det(X(4)'X(4))=\begin{vmatrix}3 & 0\\ 0 & 2\end{vmatrix}=6$$

$\text{Covratio}\fallingdotseq 20/6$ ◢

[例] 以上述べたスチューデント化された誤差と各観測値の影響を計る3尺度は次のようになる.

obs.	ε_i	ε_i^s	$\varepsilon_i^s(i)$	$\tilde{y}_i$	Covratio
1	−0.230	−0.297	−0.215	−0.225	53.430
2	−0.066	−0.484	−0.364	−2.980	1166.160
3	1.242	1.414	127.475	100.899	0.000
4	0.003	0.004	0.003	0.004	77.534
5	0.051	0.274	0.197	1.169	955.484
6	−0.951	−1.340	−2.973	−3.632	0.001
7	−0.049	−0.195	−0.139	−0.599	569.047

obs.	$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$	$\tilde{\beta}_4$
1	0.138	−0.101	−0.016	−0.105	−0.058
2	−0.745	2.338	−0.052	−0.527	1.426
3	36.877	47.777	−63.741	−52.641	−32.896
4	−0.003	0.001	1E−4	0.003	0.002
5	0.408	−0.287	0.925	−0.332	−0.274
6	−1.891	−1.645	2.274	2.486	1.488
7	0.004	0.045	−0.005	0.004	−0.350

この結果から[1],次の点が指摘される.

3番目のデータと6番目のデータのスチューデント化された誤差 ε_i^s の絶対値に大差はないが,$\varepsilon_i^s(i)$ では3番目のデータのものが極端に大きくなっている.3番目のデータを省くことにより回帰係数の値が大きく変化することが第1の尺度 $\tilde{\beta}_0,\tilde{\beta}_1,\tilde{\beta}_2,\tilde{\beta}_3,\tilde{\beta}_4$ から読みとれる.データ数が多ければこのような大きな変化を生じないものと思われる.また,4番目のデータの各 $\tilde{\beta}_i$ の値が小さいのは,このデータの各説明変数の値が平均値に最も近いことから納得できる.第1の尺度は,当然のことながら第2の尺度 $\tilde{y}_i$ ともよく対応している.Covratioは,データ数が多い場合にはその多くが1に近い値をとる.本結果では,2番目のデータが空間 $X'X$ の退化に一番大きな影響をもっていることを示す.

1) このケースは解説用の問題なのでデータ数が少ないので,以下の議論は実は無理な点もあるが,勉強のためにこれを行なう.

11. モデルの決定と検定

11.1 フルモデルと縮小モデル

式(5)の行列 X の列数を h とする．通常の回帰分析では定数項を他の説明変数と区別しているため，回帰モデルの自由度が $(p+1)$ というように煩わしい1が表われる．そこで，定数項も変数とみなし $h=p+1$ と置き換えて考える．この時，回帰モデルの修正前の自由度は h，誤差の自由度は $(n-h)$ で表わされる．この回帰モデルを，考慮すべきすべての説明変数を含むという意味でフルモデル (FM_h) とよぶことにする．h はモデルの自由度または次元である．

一度フルモデルを設定した後は，われわれの研究対象を，このフルモデルに含まれる h 個の説明変数の部分集合による回帰モデルに限定して考える．フルモデルに対比して，自由度 k の部分モデルを縮小モデル (RM_k) とよぶことにする．縮小モデルは全部で 2^h 個考えられるが，重回帰モデルとして定数項を必ず含むことにすれば，2^{h-1} 個の縮小モデルが得られる．特別の場合として，RM_h はフルモデルを，RM_1 は定数項モデルを表わす．

11.2 F 検定

モデルの検定統計量としては，モデルの誤差平方和 (SSE) を用いた次の F 検定量が一般的である．

$$F_{n-h}{}^{h-k}=\frac{[SSE(RM_k)-SSE(FM_h)]/(h-k)}{SSE(FM_h)/(n-h)} \tag{81}$$

分母はフルモデルの平均誤差平方和を表わす．分子は，フルモデルの誤差平方和に対する縮小モデルの誤差平方和の増分を，その両モデルの自由度の差で割ったものに等しい．

縮小モデルとして RM_1 すなわちモデル式 $y=\bar{y}+\varepsilon$ を考える．この時，次の修正済み分散分析表が得られる．ただし，SS は平方和を表わす．

	D. F.	平方和
回帰	0	0
誤差	$n-1$	$SSE(RM_1)=SS(FM_h)+SSE(FM_h)$
全体	$n-1$	$\boldsymbol{y'y}-n\bar{y}^2=SS(FM_h)+SSE(FM_h)$

(82)

すなわち，式(81)は次式に変形される．ただし $h=p+1$，$k=1$ である．

$$F_{n-h}{}^{h-1}=\frac{[SSE(RM_1)-SSE(FM_h)]/(h-1)}{SSE(FM_h)/(n-h)} \tag{83}$$

$$=\frac{SS(FM_{p+1})/p}{SSE(FM_{p+1})/(n-p-1)}$$

分母はフルモデルの平均誤差平方和を，分子は平均平方和を表わしている．この値は分散分析表(26)の通常の F 検定になる．また，この検定は定数項 β_0 以外の回帰係数が零という次の帰無仮説に対応する．

$$H_0:\beta_1=\beta_2=\cdots=\beta_p=0 \tag{84}$$

次に，フルモデルから説明変数を1個省いた縮小モデル RM_{h-1} を考える．このモデルの修正済み分散分析表は次のようになる．

	D. F.	平方和
回帰	$p-1$	$SS(RM_{h-1})$
誤差	$n-p$	$SSE(RM_{h-1})$
全体	$n-1$	$\boldsymbol{y'y}-n\bar{y}^2$

(85)

式(81)は，$h=p+1, k=p$ より次式になる．

$$F_{n-h}{}'=\frac{[SSE(RM_{h-1})-SSE(FM_h)]}{SSE(FM_h)/(n-h)} \tag{86}$$

この検定は，フルモデルから省かれた回帰係数 β_k の帰無仮説に対応するが，縮小モデルに x_k を追加した場合，またはフルモデルから x_k を削除した場合の検定量になる．

$$H_0:\beta_k=0 \tag{87}$$

同様にして，フルモデル FM_h から任意の l 個の説明変数 $x_1, \cdots, x_l$ を省いて得られる縮小モデル RM_{h-l} を考える．式(81)の F 検定を行なうことは，次の帰無仮説の検定に等しい．

$$H_0:\beta_1=\beta_2=\cdots=\beta_l=0 \tag{88}$$

1度に複数個の説明変数を省くことは，固有技術等の助けなくして行なうことはむずかしい．そこで $l=1$ の場合に限定した使用法が多く，後述の逐次変数選択法と関係してくる．

［例］データ(1)で，$y=\beta_0+\sum_{i=1}^{4}\beta_i x_i+\varepsilon$ をフルモデル FM_5 とする分散分析表は(22′)である．

縮小モデルとして次の3モデルを考え，その分散分析表を示す．

$RM_1: y=\bar{y}+\varepsilon$ の場合

	D. F.	平方和	平均平方和	F
回帰	0	0	0	0
誤差	6	19.728	3.288	
全体	6	19.728		

(82′)

$RM_2: y=\beta_0+\beta_3 x_3+\varepsilon$ の場合

	D. F.	平方和	平均平方和	F
回帰	1	15.347	15.347	17.520**
誤差	5	4.381	0.876	
全体	6	19.728		

$RM_4: y=\beta_0+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\varepsilon$ の場合

	D. F.	平方和	平均平方和	F
回帰	3	17.212	5.737	6.840**
誤差	3	2.516	0.839	
全体	6	19.728		

(85′)

以上から，FM_5 に対する RM_1 の帰無仮説と F 検定は次のとおりになる．

$H_0: \beta_1=\beta_2=\beta_3=\beta_4=0$

$$F_2^4=\frac{(19.728-2.508)/4}{2.508/2}=3.433<F_2^4(0.05)=19.25$$

FM_5 に対する RM_2 の帰無仮説と F 検定は次のとおりになる．

$H_0': \beta_1=\beta_2=\beta_4=0$

$$F_2^3=\frac{(4.381-2.508)/3}{2.508/2}=0.062<F_2^3(0.05)=19.00$$

FM_5 に対する RM_4 の帰無仮説と F 検定は次のとおりになる．

$H_0'': \beta_1=0$

$$F_2^1=\frac{(2.516-2.508)/1}{2.508/2}=0.006<F_2^1(0.05)=18.51$$

◢

11.3 AIC と C_p 基準

モデルの検定統計量として F 検定が一般的であるが，以下に述べる AIC(Akaike Information Criterion, 赤池情報量規準)(文献[12][13]) や Mallows の C_p 基準(文献[3])を用いればモデル決定がより容易になる．

AIC は，Kullback-Leibler 情報量の漸近的不偏推定量として導かれる，式(89)で定義される．

$$\text{AIC}=-2\times(\text{モデルの最大対数尤度})+2\times(\text{モデルの自由パラメータ数}) \quad (89)$$

$$=n\log 2\pi+n\log\left(\frac{1}{n}SSE\right)+n+2(h+1) \quad (89')$$

$$=n\log(SSE)+2(h+1)+C \quad (89'')$$

重回帰モデルに限定すれば式 (89′) になる．回帰係数 $\beta_0,\cdots,\beta_p$ と分散 σ^2 の $h+1$ 個の自由パラメータをもつ．これをデータ件数の同じモデルに限定すれば式 (89″) になり，本講座では定数 C を省いたものを用いることにする．

この AIC を最小にするモデルを選択する方式を MAIC (minimum AIC) 方式という．この方式は，評価尺度が同程度なら，次元の小さなモデルのほうを良しとする“ケチの原理「Principal of parsimony」”(文献[7] p. 17)や“オッカムのかみそり”(文献[14] p. 90)と一脈相通じるものがある．

AIC 利用の注意事項 (文献[12] pp.63-64) として，次の点が指摘されている．

1) $h+1<2\sqrt{n}$

2) AICの値の差が 1 ～ 2 程度以上なら，AICの値の差は有意と考えられ，AIC の値の小さなモデルがよい．しかし，その差が 1 以下なら，どちらのモデルも大同小異である．

3) MAIC 方式により選ばれたモデルの次元が高い時は，再検討が必要である．

[例] 分散分析表(85′)より，RM_4 の AIC は，

$$\text{AIC}=7\times\log(2.516)+2\times5+C$$

$$=7\times(0.923)+10+C=16.459+C$$

モデルの比較には定数 C を省く．◢

一方，RM_p の C_p 統計量は式(90)で定義される．

$$C_p=SSERM_p/\hat{w}^2+(2p-n) \quad (90)$$

$\hat{w}^2$ としては，「最も複雑なモデル」すなわち FM_h の誤差分散の推定値に $SSE(FM_h)/(n-h)$ をもってくればよい．モデル決定には，縦軸に C_p 値，そして横軸に p 値をロットしたものを利用する．すなわち $C_p=p$ の直線の近傍にあるモデルが片寄りの少ないよいモデルなので，この中で原点に近いモデルを選べばよい．

AIC は，漸近的には C_p 基準と同等になる．小標本の場合，C_p 基準のほうが，より一層パラメータ節約的である (文献[15]p. 155)．

[例] 分散分析表 (85′) で表わされる RM_4 の C_p 基準は，分散分析表(22′)とから次のようになる．

$$C_p=2.516/1.254+(8-7)=3.006$$ ◢

これらの基準は，漸近的に F 検定の棄却限界として有意水準に無関係に2という値を用いることと同値になる．しかし，モデル決定の目安として実用上便利であり，多くの適用例のフィルターを通して有効性の検証が必要となろう．

11.4 総当り法

本解説で使った数値例(データ(1)) に対して総当り法を適用した結果を示す．

説明変数	R^2	SSE	F	C_p	AIC	p
x_2	6.6E-4	19.715	4.574(*)	12.722	26.870	
x_1	0.178	16.207	3.641(*)	9.924	25.498	
x_5	0.549	8.899				2
x_4	0.641	7.089	1.218(*)	2.653	19.710	
x_3	0.778	4.381	0.498	0.494	16.341	
x_1x_2	0.223	15.322	5.109(*)	11.219	27.105	
x_1x_5	0.620	7.499				
x_1x_4	0.642	7.072	1.820(*)	4.640	21.693	
x_2x_5	0.663	6.649				
x_4x_5	0.667	6.562				
x_2x_4	0.669	6.525	1.602(*)	4.203	21.129	3
x_2x_3	0.778	4.371	0.743	2.486	18.325	
x_1x_3	0.781	4.314	0.720	2.440	16.233	
x_3x_5	0.854	2.876				
x_3x_4	0.862	2.731	0.089	1.178	15.033	
$x_1x_2x_5$	0.667	6.570				
$x_1x_4x_5$	0.670	6.508				4
$x_1x_2x_4$	0.672	6.471	3.160(*)	6.160	23.071	

$x_2x_4x_5$	0.689	6.197				
$x_1x_2x_3$	0.785	4.237	1.379	4.379	20.107	
$x_1x_3x_5$	0.859	2.785				
$x_1x_3x_4$	0.866	2.641	0.106	3.106	16.798	4
$x_2x_3x_4$	0.872	2.516	0.006	3.007	16.459	
$x_3x_4x_5$	0.873	2.505				
$x_2x_3x_5$	0.874	2.490				
$x_1x_2x_4x_5$	0.689	6.139				
$x_1x_2x_3x_4$	0.873	2.508		5.000	18.436	
$x_1x_3x_4x_5$	0.873	2.499				5
$x_1x_2x_3x_4$	0.874	2.483				
$x_2x_3x_4x_5$	0.878	2.400				
$x_1x_2x_3x_4x_5$	0.879	2.397				6

表中の変数は，重回帰モデルに用いられた説明変数を示す．同一次元のモデルでは，R^2 値の小さいもの順に並べた．p は C_p で用いられるモデルの次元 p を表わす．

11.5 逐次変数選択法

(1) アルゴリズム

逐次変数選択法のアルゴリズムを，総当り法の結果を用いて説明する．

変数増加法は，説明変数が1個のモデルの中で R^2 値の最大な $\{x_3\}$ を選ぶことから出発する．次のステップは，このモデルに残りの説明変数 $\{x_1x_2x_4\}$ の中から1個を選んでできる3組のモデル $\{x_3x_1\}$, $\{x_3x_2\}$, $\{x_3x_4\}$ の中で R^2 値最大の $\{x_3x_4\}$ を選ぶ．以下同様にして，$\{x_3x_4x_5\}$, $\{x_2x_3x_4x_5\}$, $\{x_1x_2x_3x_4x_5\}$ が選ばれる．

プログラムでは，各ステップで元のモデルと新しく得られたモデルを式(86)により逐次 F 検定を行ない前もって決められた有意水準(F_{in} 水準)により，帰無仮説($\beta_k=0$)が棄却されない場合停止する．

変数減少法はフルモデル $\{x_1x_2x_3x_4x_5\}$ から出発する．次のステップでは，このモデルから1変数を省いた5個のモデルを検討し，R^2 値最大の $\{x_2x_3x_4x_5\}$ を選ぶ．以下のステップも同様に繰り返す．現在選ばれているモデルと新しく選ばれたモデルを式(86)により逐次 F 検定を行ない，前もって決められた有意水準(F_{out})による帰無仮説($\beta_k=0$)が棄却された時，この β_k をモデルから省くことができないので停止する．

変数増減法は F_{in} 水準により停止するまでは変数増加法と同じであり，その後変数減少法に切り換わり F_{out} 水準で停止する．

変数減増法は F_{out} 水準により停止するまでは変数減少法と同じであり，その後変数増加法に切り換わり F_{in} 水準で停止する．

以上が逐次変数選択法の代表的手法であるが，有名な統計解析システム SAS (文献[6]) には MAXR 法と MINR 法も提案されている．

MAXR 法は，モデル $\{x_3x_4\}$ からモデル $\{x_3x_4x_5\}$ が選ばれる過程は変数増加法と同じである．この後，現モデル $\{x_3x_4x_5\}$ の各1変数をモデル外の変数 $\{x_1x_2\}$ の1変数と置き換えた6組のモデルを考え，最も成績のよいモデル $\{x_2x_3x_5\}$ を選ぶ．次にモデル $\{x_2x_3x_5\}$ の1変数をモデル外の $\{x_1x_4\}$ の1変数と置き換えた6組のモデルを考えるが，モデル $\{x_2x_3x_5\}$ が最大の R^2 値をもつので改良ステップを停止する．モデル $\{x_2x_3x_5\}$ から $\{x_2x_3x_4x_5\}$ へは変数増加法と同様であり，改良ステップではモデルの1変数を $\{x_1\}$ と置き換えた4組のモデルを検討し現モデルの R^2 値が最大であるので改良ステップを停止する．このアルゴリズムは，R^2 値が増加しなければ停止するが，さもなければフルモデルを選んで停止する．

MINR 法は，改良ステップで R^2 値最大のモデルを選ぶのではなく，現モデルより R^2 値の大きい改良モデルの中で，R^2 値最小のモデルを選ぶ．これにより探索されるモデル数が増加するので，一般的に言って他の手法よりよいモデルが選ばれる可能性が大きい．

(2) 問題点

逐次変数選択法には次の問題点がある．

① どの逐次変数選択法を用いても，各次元で最大の R^2 値を与えてくれる最良モデルの系列を確実に選ぶ保証はない．すなわち，次元 p が13程度ならば総当り法[1]を実施したほうが全ての点が明らかになり，逐次変数選択法の結果をあれこれと検討することに比べ思考の節約になる．

優れた統計学書の多くは，コンピュータの未発達な時代に書かれているため，総当り法を馬鹿げた手法とする傾向が強い．また逐次変数選択法の優劣にかなりの頁をさいたものが多い．この優劣論は多分に経験にもとづいているのに対し，フルモデルに対し許容できる縮小モデルを探すという立場にたてばフルモデルから出発する変数減少法や変数減増法をよしとすべきだと考える．

再度成績の優劣の立場にたてば，これら代表的な逐次変数選択法よりも MAXR 法と MINR 法のほうが一般的にいってよい結果を与える．しかし，これらの手法でも十分ではない．1変数の置き換えによる改良ステップが停止した後，2変数さらには3変数の置き換えステップを追加すればさらによいモデルを選ぶことができる．しかし計算時間が増大し総当り法と変らなくなる．

---

1) 計算機の発達と掃き出し法によるアルゴリズムの改良により IBM 4341 程度の中型機で CPU10 秒程度で実行できる．

② バッチプログラムに事前に F_{in} と F_{out} 水準を組み込んでモデル決定することには問題[1]がある．すなわち，有意水準の決定は各分野の固有知識にもとづいて後天的に決定する場合も多い．また，事前に決めた有意水準により逐次変数選択法を停止することによって得られる**計算時間の節約**は，それを行なわないですべての次元にわたって得られるモデル系列のもたらす**情報**よりも重要とは考えられない．すなわち，バッチプログラムでは逐次F検定による停止規則を無効化しすべての次元にわたってモデルを求め，その結果を解析者が試行錯誤して最終モデルの決定を行なったほうがよい．

(3) 逐次変数選択法の利用分野

以上の議論は総当り法が実行可能な範囲では，逐次変数選択法よりも総当り法を用いたほうをよしとする筆者の意見である．大筋において読者の賛同が得られることと思う．しかし，総当り法が実用上実施不可能な範囲での対応策は議論がわかれる．これに対しては私見であるが，変数増加法と変数減少法を用いて全次元にわたってモデルを求め，そのモデルのAIC，C_p，F値により適切と考えられる次元を決定し，次にその次元の前後でのみ総当り法を実施するのが実際的ではないかと考える．

(4) 多重共線性の影響

フルモデルとして5個の説明変数 $\{x_1x_2x_3x_4x_5\}$ を考えた場合，変数増加法では順次モデル $\{x_3\}$, $\{x_3x_4\}$, $\{x_3x_4x_5\}$, $\{x_2x_3x_4x_5\}$, $\{x_1x_2x_3x_4x_5\}$ が選ばれる．変数減少法ではフルモデルから出発して，順次モデル $\{x_2x_3x_4x_5\}$, $\{x_2x_3x_5\}$, $\{x_3x_5\}$, $\{x_3\}$ が選ばれる．この結果，説明変数が2個と3個の場合，両手法の選ぶモデルが異なっていることがわかる．しかし，フルモデルとして x_5 を省いて多重共線性を解消したものを考えれば，両手法の選ぶモデルは $\{x_3\}$, $\{x_3x_4\}$, $\{x_2x_3x_4\}$, $\{x_1x_2x_3x_4\}$ と一致する．このことは，多重共線性の影響を省けば両手法の選ぶモデル系列が一致し，しかもそれが各次元で最高の R^2 値をもつモデルになる可能性が高いことを示唆している．

モデル決定において，逐次変数選択法で選んだモデルが各次元で最良のモデルであれば，モデル決定をこの系列上に限定でき，問題が単純化される．

11.6 最終モデルの決定

ここでは，多重共線性等が解消された後の一応妥当と考えられるフルモデルを仮定する．解析者にとって，与えられた説明変数が全宇宙であるから，すべての基準または出発点をこのモデルに置くべきと考える．すなわち，重回帰分析における最終のモデル決定を次のように定式化したい．

(モデル決定の指針)

フルモデルのモデル適合度のよさを表わす尺度—R^2 値，回帰平方和，AIC 規準，C_p 基準—の ε 近傍にある縮小モデルを満足モデルとよぶことにする．この中で，"ケチの原理"にしたがい最小の自由度をもち，選ばれた説明変数が他の満足モデルの説明変数の多くと共通部分をもつようなモデルを選べばよい．この基準にもとづいて決められたモデルは，固有知識の立場からも支持されることが望ましい．

[例] 総当り法の結果を用いて，フルモデル $\{x_1x_2x_3x_4\}$ から最終モデルを求める過程を述べる．

表中のF欄は，誤差平方和(SSE) を用いた式 (81) によるフルモデルと各縮小モデルとのF検定を示す．今回のデータは作意的なデータであるのですべてのF検定が棄却されない．そこで，かりに $F_{0.05}(3,5)=1.220$, $F_{0.05}(2,4)=1.500$, $F_{0.05}(1,3)=2.000$ とした場合，フルモデルに対して $p=4$ では $\{x_1x_2x_4\}$ のみが棄却される．残りの3モデルは棄却されないのでフルモデルと同等の説明力があると考えられる．この中で一番成績のよい $\{x_2x_3x_4\}$ は最終モデルの候補と考えられる．$p=3$ では，モデル $\{x_3x_4\}$, $\{x_1x_3\}$, $\{x_2x_3\}$ が，$p=2$ ではモデル $\{x_3\}$ が棄却されない．そこで，有意水準を5%に固定して考えるならば，モデル $\{x_3\}$ すなわち $\hat{y}=\hat{\beta}_0+\hat{\beta}_3x_3$ が許容できる最小次元のモデルである．説明変数 x_3 は，他の棄却されないモデルの共通集合でもあるので妥当と考えられる．

C_p 統計量は予測値の平均誤差平方和の合計を標準化した尺度である．モデルが片寄りのないものならば C_p の期待値は p となるので，$C_p=p$ からの片寄りが2以内のものの中から原点に近いモデル $\{x_3\}$ を選ぶ．

AIC では MAIC 方式により，最小値15.033をもつモデル $\{x_3x_4\}$ が選ばれる．

以上から今回のデータでは，$y=-13.216+0.140x_3$ か $y=-6.722+0.100x_3-0.025x_4$ のいずれかに決めればよい．そして，現実のシステムへ適用し有効性の評価を受ける必要があろう．

1) プログラムにおける停止規則の役割は，アルゴリズムが収束しないでコンピュータ資源の浪費をさけることが第1目的である．

参考文献

([1]～[12]は前号参照)

13) 新村秀一，清水憲彦：自己回帰モデルによる汚染質濃度のスペクトル解析について，大気汚染研究12(2)，59／70(1977)

14) 佐和隆光：経済学とは何だろうか，岩波書店，1982

15) 佐和隆光：回帰分析，朝倉書店，1979