重回帰分析におけるモデル決定

新村 秀一

1. はじめに

本講座では，2回にわたり行列表現を用いて重回帰分析の一般論を述べた．前号では，重回帰分析の理解を深め，プログラム作成上必要となる「掃き出し法」について述べた．本号では，最後のしめくくりとして，実際のデータを用いてモデル決定に到る道筋を示す．このようなデータ解析の手順や方法には，筆者の私見が入ることをお許しいただきたい．

2. 使用データ

読者の追試や検証を容易にするため，応用回帰分析（文献[1]，p.341，p.351）掲載のAデータおよびBデータを用いる注)．

Bデータは13件のデータからなり，次の4個の説明変数（x_1～x_4）と目的変数（yまたはx_5）をもつ工場データである（表1）．

x_1＝$3CaO\cdot Al_2O_3$量　　x_3＝$4CaO\cdot Al_2O_3\cdot Fe_2O_3$量

x_2＝$3CaO\cdot SiO_2$量　　x_4＝$2CaO\cdot SiO_2$量

y＝x_5＝セメント1グラム当り発熱量

x_1，x_2，x_3，x_4はセメントの製造されたクリンカーの重量百分率を示すので，合計はほぼ100になる．本講座では，このデータを説明変数の数の少ない場合とみなす．

Aデータは25件のデータからなり，次の9個の説明変数（x_2～x_{10}）と目的変数（yまたはx_1）をもつ工場データである（表2）．

y＝x_1＝月蒸気使用量（ポンド）

x_2＝純脂肪酸各月在庫量（ポンド）

x_3＝製造された粗グリセリン量（ポンド）

x_4＝平均風速（マイル/時）

x_5＝各月日数　　x_8＝平均気温（$°F$）

x_6＝各月稼動日数　　x_9＝(平均風速)^2

x_7＝32$°F$以上の日数　　x_{10}＝始動回数

表1 Bデータ

OBS	x_1	x_2	x_3	x_4	x_5
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
m	7,462	48,154	11,769	30,000	95,423
s	5.882	15.561	6.405	16.738	15.044

本講座では，このデータを説明変数の数の多い場合とみなして扱う．

3. Bデータの解析

3.1 基礎統計量と主成分分析

表1に各変数の平均値と標準偏差を示す．表3の上三角行列は相関係数を示す．$r_{x_1x_3}$と$r_{x_2x_4}$が5％で有意になる．説明変数の間に，$x_1+x_2+x_3+x_4\fallingdotseq 100$ の関係があるので，多重共線性が現われることが予想される．

多重共線性の検出のため，4個の説明変数で主成分分析を行なう．表3の1列目にその結果を示す．上段は固有値を，下段はその百分比を示す．

固有値は対応する主成分軸上でのデータの分散を表わす．第4主成分軸の固有値は2E－3であり，ほぼ零とみなせる．すなわち，第4主成分軸上の各観測値の座標

しんむら しゅういち　住商コンピュータサービス㈱

注) 電話にて森北出版株式会社の許可を得てあります．

表 2　Aデータ

OBS	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	10.98	5.20	0.61	7.4	31	20	22	35.3	54.8	4
2	11.13	5.12	0.64	8.0	29	20	25	29.7	64.0	5
3	12.51	6.19	0.78	7.4	31	23	17	30.8	54.8	4
4	8.40	3.89	0.49	7.5	30	20	22	58.8	56.3	4
5	9.27	6.28	0.84	5.5	31	21	0	61.4	30.3	5
6	8.73	5.76	0.74	8.9	30	22	0	71.3	79.2	4
7	6.36	3.45	0.42	4.1	31	11	0	74.4	16.8	2
8	8.50	6.57	0.87	4.1	31	23	0	76.7	16.8	5
9	7.82	5.69	0.75	4.1	30	21	0	70.7	16.8	4
10	9.14	6.14	0.76	4.5	31	20	0	57.5	20.3	5
11	8.24	4.84	0.65	10.3	30	20	11	46.4	106.1	4
12	12.19	4.88	0.62	6.9	31	21	12	28.9	47.6	4
13	11.88	6.03	0.79	6.6	31	21	25	28.1	43.6	5
14	9.57	4.55	0.60	7.3	28	19	18	39.1	53.3	5
15	10.94	5.71	0.70	8.1	31	23	5	46.8	65.6	4
16	9.58	5.67	0.74	8.4	30	20	7	48.5	70.6	4
17	10.09	6.72	0.85	6.1	31	22	0	59.3	37.2	6
18	8.11	4.95	0.67	4.9	30	22	0	70.0	24.0	4
19	6.83	4.62	0.45	4.6	31	11	0	70.0	21.2	3
20	8.88	6.60	0.95	3.7	31	23	0	74.5	13.7	4
21	7.68	5.01	0.64	4.7	30	20	0	72.1	22.1	4
22	8.47	5.68	0.75	5.3	31	21	1	58.1	28.1	6
23	8.86	5.28	0.70	6.2	30	20	14	44.6	38.4	4
24	10.36	5.36	0.67	6.8	31	20	22	33.4	46.2	4
25	11.08	5.87	0.70	7.5	31	22	28	28.6	56.3	5
m	9.424	5.442	0.695	6.356	30.480	20.240	9.160	52.600	43.364	4.320
s	1.631	0.817	0.126	1.754	0.770	3.018	10.282	17.266	23.199	0.852

はほぼ零となり，次の関係が成立する．

$$0.010x_1+0.026x_2+0.011x_3+0.027x_4 \fallingdotseq 0 \quad (1)$$

すべての因子負荷量（係数値）の絶対値に大差がないことから，多重共線性はすべての変数と関係していると考えられる．説明変数の合計が 100という制約以外の特定変数間の強い多重共線性は認められないようである．

主成分分析を用いての多重共線性の検出基準の設定は難しい．固有値の絶対値が極端に小さい場合は問題はないが，一応固有値の百分比が 1 %未満(基準 1)を目度としたい．今回は，第 3 主成分軸の固有値の百分比が 4.7 %であるが，第 2 主成分軸の39.4%に比べて極端に小さいので参考に検討することにする．第 3 主成分軸の表わす多重共線性は次式になる．

$$0.292x_1-0.136x_2+0.275x_3-0.084x_4 \fallingdotseq 0 \quad (2)$$

各主成分軸で一番大きな係数の20%以上の値(基準 2)をもつ変数間に多重共線関係を認めることにすれば，式(1)(2)からわかるとおり 4 変数とも多重共線関係にあることになる．

これらの基準の水準の設定は非常に恣意的であり，経験的な性格をもつ．特に説明変数の多い場合には，これらの水準を変えることにより，結果があ易に異なることが明らかで容る．

3.2　フルモデルの重回帰分析

表 4 はフルモデル（$y=a_0+a_1x_1+a_2x_2+a_3x_3+a_4x_4+\varepsilon$）の分散分析表である．使用したプログラムは，汎用統計解析システムＳＡＳ（文献[2]～[4]）のＲＥＧプロセジャーである．ＩＢＭ 4341でＣＰＵ0.60秒である．表 5 は，回帰係数の推定値,推定値の標準誤差 t値，規準化された推定値，tolerance（ＶＩＦの逆数)を示す．トレランスから，x_2とx_4の多重共線性がx_1とx_3よりも強いことがわかる．x_2とx_4，x_1とx_3の相関係数が高く，他の相関係数の値は低い．このことと併せて，式(1)(2)の示す事実と少し異なり，x_2 と x_4 の間に多重共線関係が，x_1とx_3の間にそれよりも弱い多重共線関係があることをうかがわせる．

3.3　総当り法

表 6 は総当り法の結果を示す．RSQUAREプロセジャーでＣＰＵ0.41秒かかった．ＲＥＧよりも出力結果が少ないのでこのような実績値になった．

表中の統計量は，決定係数（R^2），回帰平方和（SS），誤差平方和(SSE)，平均誤差平方和(s^2)，自由度修正決定係数（$\bar{R}^2$），C_p 統計量，ＡＩＣ規準である．モデル決定に際してはC_p統計量とＡＩＣ規準が特に参考になる．

表 3　相関行列と主成分分析の結果（Bデータ）

		x_1	x_2	x_3	x_4
第 1 主成分	2.236	1.000	0.229	−0.824*	−0.245
	0.559				
第 2 主成分	1.576		1.000	−0.139	−0.973**
	0.394				
第 3 主成分	0.187			1.000	0.030
	0.047				
第 4 主成分	0.002				1.000
	0.000				

上段：固有値　下段：固有値の百分比

表 4　フルモデル $(x_5=a_0+\sum_{i=1}^{4}a_ix_i+\varepsilon)$ の分散分析表（Bデータ）

	$d.f.$	平方和	平均平方和	F	prob.$>F$
モデル	4	2667.899	666.975	111.479	0.0001
誤　差	8	47.864	5.983		
全　体	12	2715.763			

$R^2=0.982$

表 5　フルモデルの各種統計量（Bデータ）

変数	推定値	標準誤差	t値	規準化推定値	tolerance
定数項	62.405	70.071	0.891	0.0	
x_1	1.551	0.745	2.083	0.607	0.026
x_2	0.510	0.724	0.705	0.528	0.004
x_3	0.102	0.755	0.135	0.043	0.021
x_4	−0.144	0.709	−0.203	−0.160	0.004

ＡＩＣ規準にしたがえば，その値の小さなモデルを選べばよいので，値の小さい順に順位をつけた．すなわちモデル$(x_1x_2x_4)$，$(x_1x_2x_3)$，(x_1x_2)，$(x_1x_3x_4)$が第1順位から第4順位に対応する．

C_p統計量によるモデル決定は，その値が〈説明変数の数＋1〉の近傍にあるものの中から絶対値の小さなものを選べばよい．表には，C_p値の小さなもの順に一応順位をつけた．第4順位までのモデルはＡＩＣの選んだモデルと一致しており，絶対値が4以下で〈説明変数の数＋1〉との偏差が1以内であることがわかる．

s^2, $\bar{R}^2$にも一応順位づけを行なった．偶然ではあるが，第4順位までのモデルは，ＡＩＣ規準とC_p統計量のそれと一致した．

さらに，フルモデルのR^2，SS，SSEに比べて，第4順位までのモデルに対応するこれらの値には遜色がない．これらのことを総合して，この4モデルのいずれかを採用すればよい．

3.4　逐次変数選択法と多重共線性

表6から，変数増加法はモデル系列$(x_4)\rightarrow(x_1x_4)\rightarrow(x_1x_2x_4)\rightarrow(x_1x_2x_3x_4)$を選ぶことがわかる．逐次$F$検定による停止規則を考えないので，増加基本系列とよぶことにする．同様に，変数減少法はモデル系列$(x_1x_2x_3x_4)\rightarrow(x_1x_2x_4)\rightarrow(x_1x_2)\rightarrow(x_2)$を選ぶことがわかる．これを減少基本系列とよぶことにする．一方，各次数で最大のR^2値をもつモデル系列$(x_4)\rightarrow(x_1x_2)\rightarrow(x_1x_2x_4)\rightarrow(x_1x_2x_3x_4)$を最良系列とよぶことにする．

両基本系列が一致する場合，それが最良系列である可能性が大きいことは，経験および常識的に納得できる．この場合，対象領域の固有知識の助けをかりないとすれば，モデル決定を最良系列上のモデルに限定して考えても大きな間違いをおこさず思考の節約になる．

両基本系列の不一致は多重共線性の影響によることが大きいと言われる．そこで，x_2またはx_4のいずれかをフルモデルから省くことにする．x_2を省いた場合，両基本

表 6　総当り法によるBデータの各種統計量

変　数	R^2	SS	SSE	s^2	$\bar{R}^2$	C_p	ＡＩＣ
定数項	0	0	2715.763	226.314	0	442.917	106.789
3	.286	776.363	1939.400	176.309(15)	.221(15)	315.154(15)	104.412(15)
1	.534	1450.076	1265.687	115.062(13)	.492(13)	202.549(14)	98.864(13)
2	.666	1809.427	906.336	82.394(11)	.636(11)	142.486(12)	94.522(11)
4	.675	1831.896	883.867	80.352(10)	.645(10)	138.731(11)	94.196(10)
1　3	.548	1488.691	1227.072	122.707(14)	.458(14)	198.093(13)	100.461(14)
2　4	.680	1846.883	806.880	86.888(12)	.616(12)	138.226(10)	95.974(12)
2　3	.847	2300.320	415.443	41.544(9)	.816(9)	62.438(9)	86.381(9)
3　4	.935	2540.025	175.738	17.574(8)	.922(8)	22.373(8)	75.197(8)
1　4	.973	2641.001	74.762	7.476(6)	.967(6)	5.496(6)	64.086(6)
1　2	.979	2657.859	57.904	5.790(4)	.974(4)	2.678(1)	60.764(3)
2　3　4	.973	2641.949	73.815	8.202(7)	.964(7)	7.337(7)	65.920(7)
1　3　4	.981	2664.927	50.836	5.648(3)	.975(3)	3.497(4)	61.072(4)
1　2　3	.982	2667.652	48.111	5.346(2)	.976(2)	3.041(3)	60.356(2)
1　2　4	.982	2667.790	47.973	5.330(1)	.976(1)	3.018(2)	60.318(1)
1　～　4	.983	2667.899	47.864	5.983(5)	.974(5)	5.000(5)	62.289(5)

ＡＩＣ$=n\log(SSE)+2(h+1)+C$（Cを省いて計算した）

系列は$(x_4)\rightarrow(x_1x_4)\rightarrow(x_1x_3x_4)$と一致する．$x_4$を省いた場合も同様に，両基本系列は$(x_2)\rightarrow(x_1x_2)\rightarrow(x_1x_2x_3)$と一致する．しかし，$x_4$を省いたほうがＡＩＣ規準の第2と第3順位モデルが残るのに対し，x_2を省いた場合には第4順位モデルしか残らないので，x_4を省くべきであろう．次にx_1とx_3のいずれかを省くことも考えられるが，両系列が一致したことと，C_p 統計量やＡＩＣ規準等を考慮した情況判断によりやめたほうがよいと考える．

3.5 モデルの決定

Ｂデータの解析から次のことがわかった．

① 相関分析では，$r_{x_2x_4}$，$r_{x_1x_3}$ が5％で有意となった．

② 主成分分析で，固有値の百分比が1％未満の主成分軸上で，最大の係数値の20％以上の値をもつ変数間に多重共線関係があるものと考えた．この基準にしたがえば，第4主成分に対応する(1)のような多重共線関係を認めることになる．

③ フルモデルの重回帰分析では，x_2とx_4のトレランスがx_1とx_3に比べ1桁小さかった．すなわち，x_2とx_4のほうがx_1とx_3に比べ強い多重共線性があり，①よりこの2組の変数間はほぼ独立と考えられる．

④ 総当り法では，ＡＩＣ規準，C_p統計量，$\bar{R}^2$，s^2の選んだ上位4モデルは，偶然一致していた．最終モデルはこの中から選んでよいと考えられる．

⑤ フルモデルからx_2またはx_4を省いてそれぞれ両基本系列を求めなおした．どちらの場合も両基本系列は一致した．x_4を省いた場合，ＡＩＣ規準の第2，第3順位モデルがこの基本系列上にある．x_2を省いた場合，第4順位モデルのみが基本系列と一致した．

以上から，モデル$(x_1x_2x_3)$または(x_1x_2)のいずれかを最終のモデルとすべきと考える．この両モデルの係数の推定値等を表7に示す．表5と比べて良好なことがわかる．

表 7　最終候補モデルの各種統計量（Ｂデータ）

変数	推定値	標準誤差	t値	規準化推定値	tolerance
定数項	48.194	3.913	12.315***	0.0	
x_1	1.696	0.205	8.290***	0.663	0.308
x_2	0.657	0.044	14.851***	0.679	0.940
x_3	0.250	0.185	1.354	0.106	0.318
定数項	52.577	2.286	22.998***	0.0	
x_1	1.468	0.121	12.105***	0.574	0.948
x_2	0.662	0.046	14.442***	0.685	0.648

4. Ａデータの解析

4.1 基礎統計量と主成分分析

表2に各変数の平均値と標準偏差を示す．表8の上三角行列は相関係数を示す．11個の相関係数が1％で有意となった．

多重共線性の検出のため，9個の説明変数で主成分分析を行なう．表8の1列目にその結果を示す．固有値の百分比が1％未満のものは第8主成分（C_8）と第9主成分（C_9）であるが，参考に第7主成分（C_7）も考える．係数の最大値の20％以下のものを無視して次の多重共線関係が得られる．

$$C_9：0.047x_4-0.043x_9\fallingdotseq 0 \quad (3)$$

$$C_8：-0.120x_2+0.129x_3-0.030x_6\fallingdotseq 0 \quad (4)$$

$$C_7：0.046x_2-0.065x_6+0.215x_7+0.224x_8\fallingdotseq 0 \quad (5)$$

多重共線性の解消に際しては，最初にx_4またはx_9のいずれかを次にx_2, x_3, x_6のいずれかをフルモデルから除くことが考えられる(階層性)．それらの結果をみて，x_2，x_6，x_7，x_8の取扱いを決めるべきと考える．

4.2 フルモデルの重回帰分析

表9はフルモデル$(x_1=a_0+\sum_{i=2}^{10}a_ix_i+\varepsilon)$の分散分析表である．ＲＥＧプロセジャーでＣＰＵ1.07秒かかった．

表10は，回帰係数の推定値，推定値の標準誤差，t値，

表 8　相関行列と主成分分析の結果（Ａデータ）

		x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
C_1	3.295(0.366)	1.000	0.944**	−0.126	0.382	0.685**	−0.191	−0.002	−0.131	0.616**
C_2	3.155(0.351)		1.000	−0.144	0.248	0.764**	−0.226	0.068	−0.134	0.601**
C_3	1.020(0.113)			1.000	−0.317	0.231	0.558**	−0.616**	0.990**	0.074
C_4	0.809(0.090)				1.000	0.020	−0.205	0.077	−0.321	−0.053
C_5	0.363(0.040)					1.000	0.117	−0.210	0.212	0.601**
C_6	0.215(0.024)						1.000	−0.858**	0.492	0.118
C_7	0.106(0.012)							1.000	−0.541**	−0.237
C_8	0.033(0.004)								1.000	0.028
C_9	0.004(0.000)									1.000

固有値（固有値の百分比）

表 9　フルモデル $(x_1=a_0+\sum_{i=2}^{10}a_ix_i+\varepsilon)$ の分散分析表
（Aデータ）

	$d.f.$	平方和	平均平方和	F	prob.$>F$
モデル	9	58.947	6.550	20.177	0.0001
誤　差	15	4.869	0.325		
全　体	24	63.816			

$R^2=0.924$

規準化された推定値，tolerance（ＶＩＦの逆数）を示す．トレランスから，x_4とx_9が小数点３桁，x_2とx_3が小数点２桁というように，式(3)(4)で示された主成分分析の結果を補足説明している．逆に，トレランスの方を重視すれば，主成分分析の基準１と２で得られる多重共線関係は範囲を広げすぎていると言える．

4.3　総当り法

表11は総当り法の結果の一部を示す．RSQUARE プロセジャーでＣＰＵ3.36秒である．総当り法は説明変数の増加に伴ないＣＰＵ時間は急激に増加する．説明変数の数が13個程度でＣＰＵ１～２分前後であるのでこのあたりが限度になるが，解析者の置かれたマシン環境によっても異なってくる．そこで，本講座では仮にＢデータを総当り法の可能なデータ，Ａデータを総当り法の実施が困難なデータとみなそう．すなわち，ここで得られた総当り法の結果は，検証にのみ使用する．

4.4　逐次変数選択法

説明変数の数が少ない場合には総当り法の実施が有効であることを示した．Ａデータのように数が多い場合には，代替案として変数増加法と変数減少法の２手法を対として逐次Ｆ検定による停止規則を考えない基本系列の検討を考えたい．

表11に基本系列を示す．増加基本系列はＦ印で示され，$(x_8)\to(x_2x_8)\to(x_2x_6x_8)\to(x_2x_5x_6x_8)\to(x_2x_5x_6x_8x_{10})\to(x_2x_3x_5x_6x_8x_{10})\to(x_2x_3x_5x_6x_8x_9x_{10})\to(x_2\sim x_6, x_8\sim x_{10})\to(x_2\sim x_{10})$になる．減少基本系列はフルモデル$(x_2\sim x_{10})$より始まって，順次Ｂ印で表わされるモデルになる．逐次変数選択法の実施でわかることは，変数欄で示されたモデルの説明変数と，その決定係数および両基本系列の優劣のみである．順位は総当り法を実施することによりはじめて，モデルの自由度の等しいグループ内での順位がわかる．これを順位(1)欄に示す．モデルの自由度が４から８までで両系列は一致していない．また両系列が一致しているモデル$(x_2x_6x_8)$は自由度３の最良モデルではないが，第２順位とそれほど悪くない成績になっている．

4.5　多重共線性の解消

相関係数およびフルモデルのトレランスの検討から，

表10　フルモデルの各種統計量（Ａデータ）

変数	推定値	標準誤差	t値	規準化推定値	tolerance
定数項	1.894	6.996	0.271	0.0	
x_2	0.705	0.565	1.249	0.353	0.064
x_3	−1.894	4.146	−0.457	−0.146	0.050
x_4	1.134	0.746	1.520	1.220	0.008
x_5	0.119	0.205	0.580	0.056	0.544
x_6	0.179	0.081	2.216*	0.332	0.227
x_7	−0.018	0.025	−0.742	−0.115	0.213
x_8	−0.077	0.017	−4.666***	−0.820	0.165
x_9	−0.086	0.052	−1.651	−1.221	0.009
x_{10}	−0.345	0.211	−1.637	−0.180	0.419

(x_4, x_9)と(x_2, x_3, x_8)の各組から変数を省いて多重共線性の解消を計ることが示唆される．

どの変数をフルモデルから省くかの基準として，基本系列上において低次のモデルに入っている変数は，それより高次のモデルではじめて入ってくる変数よりも重要であると仮定する．すなわち，増加基本系列のモデル$(x_2x_3x_5x_6x_8x_9x_{10})$では，$x_9$が入っており$x_4$が入っていないので，$x_4$をフルモデルから省くことにする．減少基本

表11　総当り法によるＡデータの順位および各種統計量

変　数	R^2	順位(1)	順位(2)	C_p	ＡＩＣ
8	0.714	1 FB	1 FB	29.707	−25.329
2 8	0.860	1 FB	1 F	5.852	−41.153
6 8	0.849	2	2 B	7.792	−39.278
5 6 8	0.885	1	1 B	3.435	−44.049
2 6 8	0.880	2 FB	2 F	4.383	−42.923
2 5 6 8	0.893	1 F	1 B	3.927	−43.967
2 6 8 10	0.891	2 B	2 F	4.291	−43.502
2 3 6 8 10	0.900	1			
2 5 6 8 10	0.899	2 F	1 FB	4.972	−43.266
2 6 8 9 10	0.897	5 B			
2 4 6 8 9 10	0.917	1 B			
2 3 5 6 8 10	0.904	5 F			
2 5 6 8 9 10	0.902	9	1 FB	6.392	−42.095
2 4 5 6 8 9 10	0.920	1 B			
2 3 5 6 8 9 10	0.909	7 F			
2 5 6 7 8 9 10	0.904	15	1 FB	8.000	−40.663
2 4-10	0.923	1 B	ＡＩＣ$=n\log(1-R^2)$		
2-4 6-10	0.922	2	$+2(h+1)+C$		
2-6 8-10	0.921	3 F			
2-10	0.924	1 FB			

系列のモデル（$x_2x_6x_8x_9x_{10}$）でも同様に，x_4をフルモデルから省けばよいことになる．同じ論法でx_2，x_3，x_6に関してもx_8を省けばよいことになる．

この基準の妥当性を示す1つの証拠として，モデルの自由度7で次の検討を行なう．すなわち，（x_4x_9）と（$x_2x_3x_6$）の各組から1個ずつ変数を取る組合せは6組できる．この2個の変数をフルモデルから省いたモデルの自由度7での順位を総当り法で検討する．

（x_4x_3）を省いたモデル（x_2，x_5～x_{10}）が第15順位である．これに対し，（x_4x_2）が第21順位，（x_4x_6）が第29順位，（x_9x_2）が第24順位，（x_9x_3）が第17順位，（x_9x_6）が第27順位になっている．これらの結果から，フルモデルから多重共線性解消のため2個変数を省くとすれば，本データでは（x_4x_3）を省けばよいことがわかる．

x_4とx_3を省いた残り7変数（x_2，x_5～x_{10}）に主成分分析を行なった結果，第7主成分の固有値は0.104，固有値の百分比は1.5%であった．またトレランスの最小値はx_8の0.199であった．これらの事実より顕著な多重共線性は解消されたものと考えられる．

この7変数を新しいフルモデルの説明変数として基本系列を求めた．表11の順位(2)がこの結果を示す．モデルの自由度2～4で両基本系列は一致していないが，元の順位(1)に比べ改善されたと言える．検証のため総当り法で得られる順位を示す．

4·6 モデルの決定

最終モデルの決定のため，基本系列上のC_p統計量とＡＩＣ規準を表11に示す．両方とも，モデル（$x_5x_6x_8$）を最良としている．しかし，両系列が一致していない難点があることと，このモデルと以下のモデル（$x_2x_6x_8$），（$x_2x_5x_6x_8$），（$x_2x_6x_8x_{10}$），（$x_2x_5x_6x_8x_{10}$）のC_p統計量とＡＩＣ規準に大差がないことから，これらを併用するか固有領域の判断をまたねばならない．表12に，これらのモデルのうち3モデルの各種統計量を示す，いずれのモデルも表10で示されたモデルに比べてよいことがわかる．

5. 考　察

本講座では，重回帰分析のモデル決定で必要最低限検討しておく事項を説明変数の多い場合と少ない場合に分けて論じた．

共通の事前作業としては次のことを行ないたい．

① 主成分分析により多重共線性の大枠の把握．

② 相関分析とフルモデルのトレランスより，変数間の強い多重共線性を検出し，それが主成分分析の結果に矛盾しないことを確認する．

これらの事前作業の後，説明変数の少ない場合に次の手順をとる．

表12 最終候補モデルの各種統計量（Ａデータ）

変数	推定値	標準誤差	t値	規準化推定値	tolelance
定数項	−2.968	4.833	−0.614	0.0	
x_5	0.402	0.157	2.556*	0.190	0.993
x_6	0.199	0.041	4.859***	0.368	0.955
x_8	−0.074	0.007	−10.302***	−0.783	0.949
定数項	0.099	5.350	0.018	0.0	
x_2	0.298	0.236	1.262	0.149	0.382
x_5	0.289	0.179	1.611	0.136	0.744
x_6	0.142	0.060	2.358*	0.263	0.427
x_8	−0.076	0.007	−10.502***	−0.800	0.918
定数項	2.143	5.723	0.375	0.0	
x_2	0.433	0.271	1.594	0.217	0.288
x_5	0.225	0.190	1.181	0.106	0.660
x_6	0.150	0.061	2.467*	0.278	0.420
x_8	−0.077	0.007	−10.420***	−0.820	0.860
x_{10}	−0.204	0.204	−1.005	−0.107	0.471

③ 総当り法を実施する．

④ ②で得られた多重共線関係を考慮しながら，基本系列の動向およびC_p統計量とＡＩＣ規準を参考にして最終モデルの候補を決定する．

一方，説明変数の数が多くて総当り法を実施できない場合として，次の手順をとる．

③′ フルモデルに対する両基本系列を求める．

④′ 両基本系列から得られた情報をもとに，②で得られた多重共線性を示す変数の組の中からフルモデルから省く変数を決める．

⑤′ 上記で決められた変数をフルモデルから省き，主成分分析とトレランスにより多重共線性が解消されたことを確認の上，両基本系列を計算しなおす．元の基本系列に比べ何らかの改善点があることを確認する．

⑥′ 両基本系列上でC_p統計量とＡＩＣ規準を計算し，これらの統計量のよいモデルの一群を最終候補として決定する．

6. おわりに

重回帰分析のモデル決定は，逐次変数選択法を慎重に適用しただけでは解決のつかない問題と考える．しかるに，プログラムにデェフォルトとして組み込まれた停止規則により自動的に求まったモデルを無考察に最終結果として報告している事例報告も多い．

本講座では，私見であるが，できるだけ簡単な手順で客観的な立場から統計の玄人でない解析者のために，必要最低限の作業手順を述べた．統計の専門家の中には，このような方法は馬鹿げておりフルモデルまたは逐次変

数選択法で得られた統計量を慎重に考察すればよいとする意見もあるかと思う．しかし，実際問題に直面した多くの半玄人の解析者にとってわかりやすい手順の提案が必要と考える．

参考文献

[1] N・ドレイパー他：応用回帰分析，森北出版，1968

[2] SASユーザーズガイド，SAS Inc., 1982

[3] 新村秀一：統計解析システムSASの紹介—SAS言語を中心として—，情報処理学会医療情報学研究会資料 6-4，1/7 (1980)

[4] 新村秀一：アメリカから吹き寄せる新しい高級言語の風—SASについて，第8回日本MUG学術大会講演報告集，1/6 (1981)

行列表現による重回帰分析 (1)〔9月号〕の訂正と補足説明

3章下5行目「決定論的変数」は「実験計画法の因子」を考えればよい．

式(18)の「$x_i \in R^n$」は削除する．

式(24)を「$y_i=\beta_0+\varepsilon_i$」に変更．

式(25)から「$\beta_0=\bar{y}$と$\beta_0=\hat{\beta}_0$」を削除する．モデルや仮説に推定値を入れないほうがよい．

6章下2行目「$\hat{\beta}$は確定的でなく，多重共線性をもつ」は，「$\hat{\beta}$は正確に求められなくなる．このような状況を多重共線性が強いという」に変更．

6.2節下4行の主成分の説明は簡便法であり，一般には分布の制約はないので注意．

p.445右上2行以降「バイアスの……考えられる」は，「バイアスの影響である」に変える．

行列表現による重回帰分析 (2)〔10月号〕の訂正と補足説明

7章の平均予測値（定着した日本語訳はない）と，式(49)上2行目と下2行目の$\hat{y}_i$は，$\eta=X_i\beta$を意味する．

式(48)下1行目「標準偏差」は「標準誤差」の誤り．その下の「t統計量」は「有意点」の誤り．統計量は確率変数である．

9章の(注)の「y自身の射影子」という表現は正しくないが，式(21)の平方和分解と対応づけるための便宜的説明として用いた．

10章，式(60)上2行を「回帰分析における誤差中の系列相関を検出するための統計量で」に変更．式(61)の分子の自乗が抜けている．同頁右上7行目の「ρ」は「$\hat{\rho}$」の誤り．その下7行目「nが14以下のものは検定できない」は「作表されたものがない」の誤り．

③で大きな誤差をもつデータに対してダミー変数を導入する代りに除いても同じである．

11章式(90)下3行目で，$C_p=p$の近傍にあるモデルがよいとしているが，現実の多くの実例では傾き2の直線のまわりにC_p値がくることが多く判断に困ることが多い．

11.6節3行目「全宇宙」は，説明変数の組を決めたらそれを用いて最善の努力をすべきであるという意図である．結果が悪ければ，データを追加するか他の説明変数を探すかして新しい宇宙で再出発すべきである．しかし，TQC等で工場実験データ等の解析をしておられる方は，何回も対象データを改変されているのでこの言葉は奇異に感じるかもしれない．

重回帰分析における掃き出し演算子〔11月号〕の訂正

式(2.1)下1行目「n元n次」と2.2節下1行目「2元2次」を，「n元1次」と「2元1次」に変える．

式(2.2)下2行目と3.1節下1行目の「変数」を「未知数」に変える．

3.3節下1行目「パラメータβ」を「配置行列X」に変える．

4章1行上「各モデルの」を「各モデルの下での反応変数の」に変える．誤差平方和に関する記述のすべてを同様に変更する．

式(4.7)下1行目以降の「を考えれば，……説明変数とし」を，「に対してのみADJUSTを実行すれば，X_1を配置行として」に変える．

式(6.8)下7行目の部分モデルの誤差平方和の「3」を「8」に変える．

7章上4と5行目．x_1のタイプII平方和は「$(3/8)^2/(1/8)=9/8$」，x_2のタイプII平方和は「$(-3/8)^2/(3/8)=3/8$」の誤りです．

「謝辞」 本訂正と補足説明に関して，熱心な読者の方の意見を参考としました．