多変量解析入門

高井省三（筑波大学体育科学系）

１．多変量解析（Multivariate statistical analysis）とはなにか？

大学入学試験を考えてみよう。いくつかの科目について単純に得点を合計した点数の多いものから順に合格させる，という方法が一般的なようだ。しかし，理学部であれば数学，理科にウェイトを，文学部であれば国語，社会にウェイトをおいた合計点があっても良い。多変量解析とは，（１）この重みづけ（係数）を多種の観測値から成る変数の相互の関連を考えながら最適な重みを与え，（２）その結果として得られる合成得点（総合特性値）を多次元空間に位置づけて，（３）客観的な規準を与える，手法といえる。

２．多変量解析の手法

多くの変数が２つの変数群に分けられるとき，xの変動が原因となってyの変動が結果として生ずるような場合を考える。このときxを説明変数（独立変数，内的基準），yを規準変数（従属変数，外的基準）を呼ぶ。しかし，説明変数の中には直接観察できないものもあろう。このときは規準変数の変動パターンから間接的に知るしかない。このような変数を潜在変数という。これらの変数に基づいて多変量解析は次の表のように分類できる。外的基準のある場合は「予測の問題」，外的基準が潜在変数で与えられている場合あるいは外的基準がない場合が「内的構造分析」の方法である。これらのうちのいくつかをJMP 4.0を使って解説していく。

		説明変数		規準変数		潜在変数	
		名義尺度	間隔尺度	名義尺度	間隔尺度	名義尺度	間隔尺度
外的基準のある場合	重回帰分析	-	多数	-	1	-	-
	多変量回帰分析	-	多数	-	多数	-	-
	正準相関分析	-	多数	-	多数	-	-
	重判別分析	-	多数	多数	-	-	-
	判別分析	-	多数	2	-	-	-
	分散分析	多数	-	-	1	-	-
	共分散分析	多数	-	-	1	-	-
	数量化第一類	多数	-	-	1	-	-
	数量化第二類	多数	-	多数	-	-	-
外的基準が潜在変数である場合	因子分析	-	（多数）	-	多数	-	多数
	潜在構造分析	多数	-	-	-	多数	-
外的基準のない場合	主成分分析	-	多数	-	（多数）	-	-
	数量化第三・四類	多数	-	-	-	-	-
	クラスター分析	-	-	-	多数	多数	-

３．重回帰分析（Multiple regression analysis）

数年前から，文部科学省の学校保健統計調査では身体測定項目から胸囲が削除されて，身長，体重，座高の3項目となった。女子の胸囲のデータが必要となったが，手もとには上の3項目しかない。この3項目から胸囲を予測してみよう。

分析メニューから**モデルのあてはめ**サブメニューを選択する。

胸囲をYに割り当てる。shiftキーを押しながら身長，体重，座高を選んで追加ボタンを押す（Xに割り当てる）。性を選んでByボタンを押す。女子の結果についてだけながめれば良い。

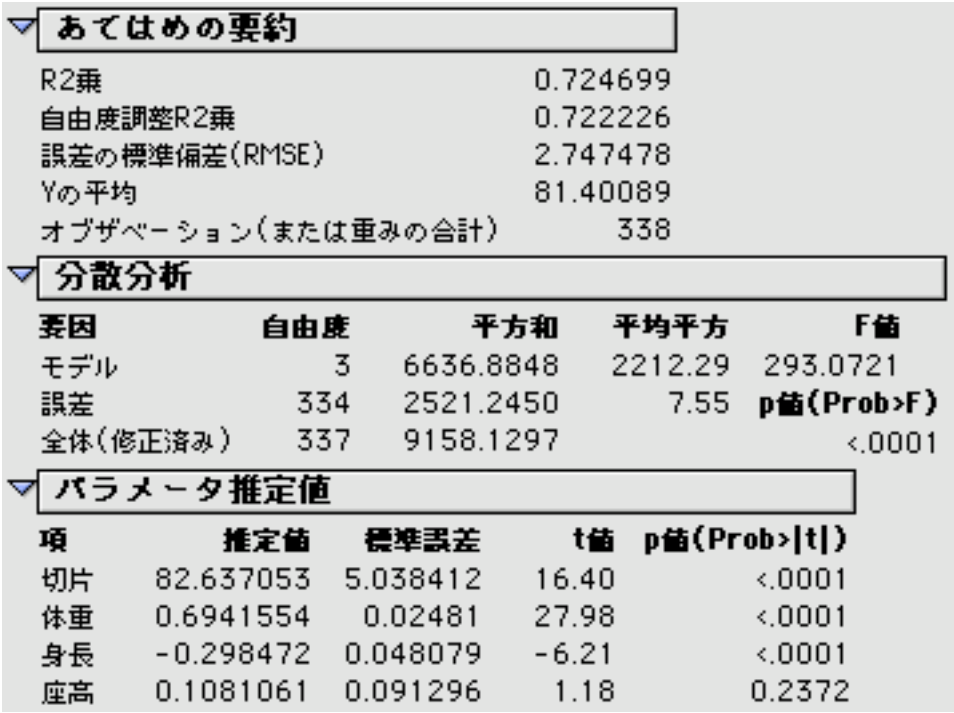
モデルの実行ボタンを押して，分析を実行する。

あてはめの要約欄を見てみよう。**R2乗**は予測値（身長，体重，座高からの）と実測値の相関係数（重相関係数）の２乗で，決定係数とか寄与率と呼ぶ。予測値の変動の72.5%が身長，体重，座高から説明できる，ということを表す。**誤差の標準偏差（RMSE）**は標準誤差で一変数の標準偏差に相当する。予測胸囲の標準偏差が2.75cmということだ。

分散分析の欄をみよう。これは切片（Intercept）以外のすべての回帰係数が０であるという仮説を検定するものである。**p値（Prob > F）**の値は5%よりも小さな値（< 0.001）だから，仮説が棄却できる。すなわち，この回帰式は統計的に有意であるということだ。

パラメータ推定値欄には，重回帰式の係数と定数（切片）が示されている。右端のtテストの値が0.5より小さければ，このパラメータが０であるという帰無仮説を棄却できる。このケースでは座高の値は０と見なせることが示されている。つまり，座高という項目は身長，体重があれば，胸囲を予測するには不必要である，という

ことを示している。



あてはめの要約

R2乗	0.724699
自由度調整R2乗	0.722226
誤差の標準偏差(RMSE)	2.747478
Yの平均	81.40089
オブザベーション(または重みの合計)	338

分散分析

要因	自由度	平方和	平均平方	F値
モデル	3	6636.8848	2212.29	293.0721
誤差	334	2521.2450	7.55	p値(Prob>F)
全体(修正済み)	337	9158.1297		<.0001

パラメータ推定値

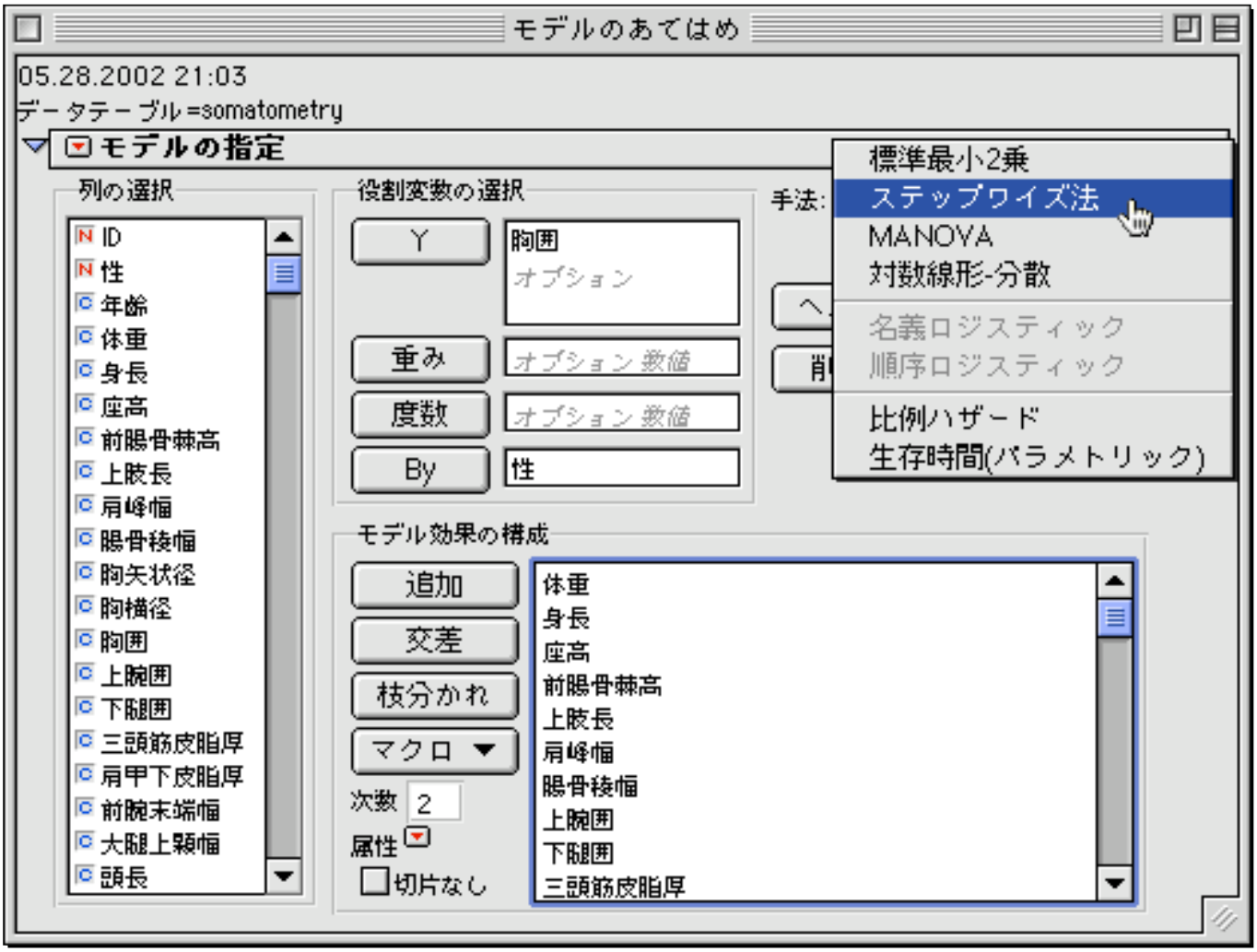
項	推定値	標準誤差	t値	p値(Prob>|t|)
切片	82.637053	5.038412	16.40	<.0001
体重	0.6941554	0.02481	27.98	<.0001
身長	-0.298472	0.048079	-6.21	<.0001
座高	0.1081061	0.091296	1.18	0.2372

こんどは，somatometoryファイルにある全ての項目を使って胸囲を予測する重回帰式を考えてみよう。ただし，重回帰分析では役に立たない項目（さきほどの回帰係数の検定で有意でなかった座高）であっても，式を構成する変数に追加すると必ず重相関係数は高くなっていく。そして，項目数と例数が等しくなると重回帰係数は1.0になる。つまり，完全に予測できるということだ。これは何かおかしい。そんなにたくさんの項目を測ったりすることはできないし，逆にそんなに少ない例数で重回帰分析をおこなうこと自体に問題がある。そこで，登場するのが「ケチの原理」とか「オッカムの剃刀」とかいわれる考えだ。最小努力の最大効果，ということを考えるわけだ。この指標がAICとかMallows C(p)というものだ。

まず，ID，性，年齢そして胸矢状径，胸横径（これらが測れるくらいなら胸囲は測れるはずだから）を除いた項目一つ一つをダブルクリックしていけば**モデル効果の構成**の欄に追加されていく。つぎに，**手法**ボタンを押して**ステップワイズ法**を選ぶ。

このステップワイズ法は，ある規準にしたがって説明変数を追加したり，削除したリをくりかえしながら，説明変数のサブセットを選択しようとするものだ。

ほんとうは，全ての変数の組み合わせについての重回帰式を求め，その中で最適な式を選ぶというのがベストな方法だが，現実にはそんなことはできそうもない（説明変数の数が増えると，計算時間は幾何級数的に増えていく）。ステップワイズ法による説明変数の選択は次善の策であることを憶えておこう。



モデルの実行ボタンを押すが，こんどはすぐには計算結果は表示されない。下図のようなステップワイズ回帰分析のコントロールパネルが表示される。

ここで変数選択の方向を決める。追加していくのか，削除していくのか，それとも追加／削除のくみあわせか。ここでは，追加／削除の組み合わせを指示するために**方向**欄で**変数増減**を選ぶ。

それから，**実行**ボタンを押すと，一気に（よく見るとステップワイズがおこなわれているので画面は刻々に変わっていくが）計算が始まる。この例では，変数は取り込まれる方向だけで，いったん取り込まれた変数が削除される（別の変数が追加されたために）ということはなかったようだ。このことが**ステップ履歴**の欄に示されている。

女子の胸囲を予測する重回帰式は，体重，身長，座高，上肢長，肩峰幅，上腕囲，肩甲下皮脂厚，前腕末端幅，大腿上顆幅で構成される説明変数をセットにもつものが，精度がよいということになったわけだ。このセットの内容は，別な項目を新たに追加すると変わってくる可能性がある。説明変数の選択はこのように重回帰分析

でおこなう前に，固有技術的見地（測りやすい，計測精度が良い，コントロールしやすい，生物学的に意味がある）からおこなうべきである。

現在の推定値

SSE	DFE	MSE	R2乗	自由度調整R2乗	Cp	AIC
2180.5744	326	6.6888786	0.7614	0.7548	5.7127727	648.3981

ロック	追加	パラメータ	推定値	自由度の数	平方和	"F値"	"p値(Prob>F)"
☑	☑	切片	68.8916971	1	0	0.000	1.0000
☐	☑	体重	0.56461151	1	1148.77	171.743	0.0000
☐	☑	身長	-0.3419656	1	170.2943	25.459	0.0000
☐	☑	座高	0.20169747	1	33.21591	4.966	0.0265
☐	☐	前腸骨棘高	.	1	0.551958	0.082	0.7744
☐	☑	上肢長	0.17142248	1	28.31554	4.233	0.0404
☐	☑	肩峰幅	0.3073026	1	30.321	4.533	0.0340
☐	☐	腸骨稜幅	.	1	3.425853	0.511	0.4750
☐	☑	上腕囲	0.16317273	1	17.27097	2.582	0.1091
☐	☐	下腿囲	.	1	2.811989	0.420	0.5176
☐	☐	三頭筋皮脂厚	.	1	4.641803	0.693	0.4057
☐	☑	肩甲下皮脂厚	0.1723899	1	170.0539	25.423	0.0000
☐	☑	前腕末端幅	-0.9966205	1	12.74797	1.906	0.1684
☐	☑	大腿上顆幅	-0.5610244	1	11.99902	1.794	0.1814
☐	☐	頭長	.	1	5.340558	0.798	0.3724
☐	☐	頭幅	.	1	0.031016	0.005	0.9458
☐	☐	下顎角幅	.	1	0.226712	0.034	0.8543

ステップ履歴

ステップ	パラメータ	アクション	"有意確率"	Seq SS	R2乗	Cp	p
1	体重	追加	0.0000	6150.116	0.6731	108.75	2
2	肩甲下皮脂厚	追加	0.0000	531.3959	0.7312	32.349	3
3	身長	追加	0.0000	163.7239	0.7491	10.194	4
4	肩峰幅	追加	0.0183	38.26405	0.7533	6.5489	5
5	座高	追加	0.1324	15.43243	0.7550	6.2721	6
6	上肢長	追加	0.1105	17.29145	0.7569	5.721	7
7	前腕末端幅	追加	0.1200	16.33893	0.7587	5.3104	8
8	上腕囲	追加	0.1750	12.38567	0.7600	5.4831	9
9	大腿上顆幅	追加	0.1814	11.99902	0.7614	5.7128	10

さて，下の欄ははじめの身長，体重，座高から予測する胸囲の重回帰式をステップワイズ法で実行した結果の一部だ。

現在の推定値

SSE	DFE	MSE	R2乗	自由度調整R2乗	Cp	AIC
2521.245	334	7.5486377	0.7247	0.7222	4	687.1982

AICの値を見ると，3項目の重回帰式は687で９項目からの重回帰式では648である。AICの値が小さいモデルの方がより良い式をあらわす。9項目を測るというめんどうに見合うだけの，予測精度の向上があるということだ。説明変数のセットによっては，項目を追加したにも関わらず精度が上がらないこともあるだろう。

ところで重回帰係数とは何を表しているのだろうか？下の3項目での重回帰分析では，体重は他の項目の影響を除いたときに（身長，座高がおなじである場合）0.694の影響を胸囲に及ぼしている，ということを意味する。

パラメータ推定値

項	推定値	標準誤差	t値	p値(Prob>|t|)
切片	82.637053	5.038412	16.40	<.0001
体重	0.6941554	0.02481	27.98	<.0001
身長	-0.298472	0.048079	-6.21	<.0001
座高	0.1081061	0.091296	1.18	0.2372

４．主成分分析（Principal component analysis）

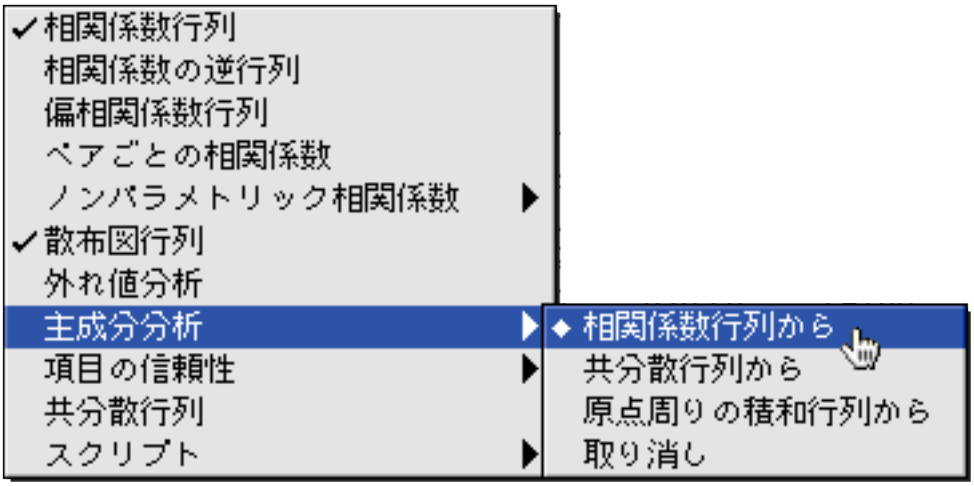
さまざまな身体計測項目の相関関係を分析して，総合された特性値を考えよう。つまり重みづき合計得点で体格をあらわしてみよう。p項目の変数からはp個の総合特性値を求めることができる。そのかたちは，

$$z_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$
$$z_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$
$$\ldots$$
$$z_p = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p$$

のようになる。実際には全体の情報の80%以上を説明するm個の総合特性値を採用するのが普通である。

男子のsomatometryデータ（ID，性，年齢を除く19項目）を使って体格の主成分分析（PCA）をためしてみよう。相関関係を分析することから出発するので，**分析**メニューの**多変量の相関**を選ぼう。Ｙには体重から下顎幅までの項目を指定する。

相関行列と散布図行列が表示される。**多変量**タイトルバー左端のチェックマークボタンを押して，**主成分分析⇒相関係数行列から**を選ぶ。このsomatometryデータは，長さの単位のデータの他にkg単位で計測している体重が含まれるので相関行列から出発する。相関行列はデータを平均０，分散１に基準化するので単位による数値の質が均等化される。



数学的には，主成分分析は相関行列の固有値（Eigenvalue）・固有ベクトル（Eigenvector）を求めることに他ならない。固有値は情報量をあらわしていると考え，19個の固有値の合計に対する割合すなわち寄与率（%）を考える。累積寄与率が80%を越えるまでを考えると7個の主成分が抽出できる。

主成分分析: 相関係数行列から

固有値	7.6704	3.0742	1.6383	1.0404	0.9698	0.7929	0.6578
寄与率	40.3705	16.1801	8.6226	5.4760	5.1041	4.1730	3.4621
累積寄与率	40.3705	56.5506	65.1732	70.6492	75.7533	79.9262	83.3883
固有ベクトル							

それぞれの主成分（Z_i）の重み係数が固有ベクトルだ。固有ベクトル欄の表はz_1～z_{19}のこれらの数値が第2列以降に示されている。

固有ベクトル							
体重	0.34526	-0.06613	0.01363	0.01602	-0.08361	0.02757	-0.07759
身長	0.20143	0.43786	-0.04794	0.07793	-0.02408	-0.00598	-0.01988
座高	0.17566	0.28934	-0.00413	-0.22223	-0.18691	-0.39065	0.19100
前腸骨棘高	0.15999	0.41042	-0.14074	0.18777	0.24263	0.29411	-0.14075
上肢長	0.16314	0.41857	0.06737	0.22913	0.25347	0.11093	-0.13899
肩峰幅	0.25214	0.04940	-0.20309	-0.36470	-0.01357	0.22978	0.23482
腸骨稜幅	0.23928	0.12327	-0.10087	-0.22091	0.33192	-0.01627	0.42734
胸矢状径	0.27253	-0.11183	0.01485	0.10279	0.04069	0.07745	-0.54170
胸横径	0.28579	-0.11792	-0.09985	-0.21308	0.06696	0.08920	0.09623
胸囲	0.30583	-0.14991	-0.11300	-0.14834	0.00111	0.21537	-0.23209
上腕囲	0.25146	-0.21891	-0.12554	-0.12813	-0.20846	-0.05753	-0.18490
下腿囲	0.29518	-0.16296	0.00159	0.01456	-0.25980	0.02626	-0.06762
三頭筋皮脂厚	0.18157	-0.27617	0.08454	0.35719	0.45476	-0.19857	0.15967
肩甲下皮脂厚	0.17440	-0.37204	0.05726	0.17298	0.34953	-0.01200	0.18443
前腕末端幅	0.20374	0.13461	0.25926	0.19912	-0.11830	-0.49528	-0.11245
大腿上顆幅	0.28602	0.03044	-0.02757	-0.00745	-0.06688	-0.38382	0.03022
頭長	0.16179	-0.01923	0.00942	0.54345	-0.49158	0.35598	0.41441
頭幅	0.03056	0.03550	0.66538	-0.28632	0.09156	0.23409	-0.08591
下顎角幅	0.15057	0.01330	0.60148	-0.07476	-0.09721	0.12916	0.18081

つまり，第1主成分Z_1は，

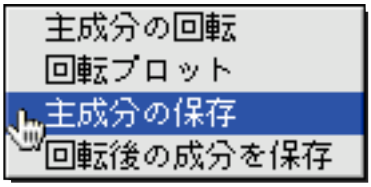
$$Z_1 = 0.345 \cdot 体重 + 0.201 \cdot 身長 + 0.176 \cdot 座高 + \ldots + 0.031 \cdot 頭幅 + 0.151 \cdot 下顎角幅$$

と表される。ただし，体重，身長などの数値は（体重－体重の平均値）／体重の標準偏差という標準化されたものである。

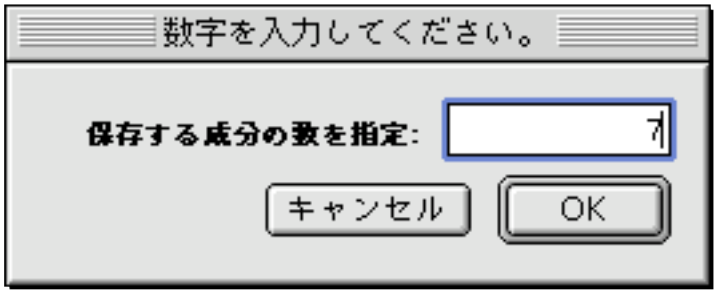
ここで，第1主成分の固有ベクトルの要素の符号はすべて正であることに注意しよう。すなわち，第1主成分は

からだ全体のサイズをあらわす総合特性値ということになる。第2主成分に注意すると，体重，胸矢状径，胸横径，胸囲，上腕囲，下腿囲，三頭筋皮脂厚，肩甲下皮脂厚，頭長の符号は負である。頭長を除けばいずれも肥っている体型に関係する項目であることが分かるだろう。だから第2主成分は「ずんぐりむっくり」か「ひょろなが」かをあらわす総合特性値であることが分かる。以下おなじようにして，生物人類学の固有知識を基にして総合特性値の内容を解釈していく。ただし，解釈がうまく行っても情報量（寄与率）が小さいときは，実用上はあまり意味がないだろう。その逆に情報量が多くても解釈がうまく行かないときは，これも実用上の意味がない。

それぞれの主成分のスコア（総合特性値）は**主成分/因子分析**バーの赤い▽ボタンを押して，**主成分の保存**を選ぶ。



するといくつの主成分のスコアを求めるか聞いてくるので数値（この場合は7）で答えると，データシートの最後のカラムに新しく追加される。

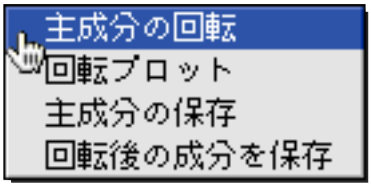


（問題）

上で求めた主成分同士の相関係数を求めよ。その結果は何を表すかを考えよ。

5．因子分析（Factor analysis）

主成分分析は総合特性値を求めることに重点をおいた分析法である。これと似た手法に因子分析がある。こちらは変数（項目）の間の相関関係を分析し，それらの変数の背後にあってそれらのバラツキを規定する潜在因子を探す手法である。歴史的には心理学者が知能の因子を探索するために発展させたといういきさつがある。



主成分分析と同じく，**分析 → 多変量の相関; 多変量 → 主成分分析 → 相関係数行列から**の手続きをとる。ついで，**主成分／因子分析**バーの赤い▽アイコンをクリックしてこんどは**主成分の回転**を選ぶ。すると回転させる因子数を聞いてくるので，累積寄与率が80%を越える主成分の個数（ここでは７）を答える。このようにJMPでは，因子分析を主成分分析の主成分を回転（Varimax回転）させるということで実行している。因子分析の手法は，因子の回転をふくめて，多種にわたっているがこの，主成分分析 → 主成分の回転，手法が一般的だろう。

因子回転

回転因子のパターン

体重	0.7985035	0.2287870	0.1407714	0.2075209	0.2378116	-0.290042	0.2149042
身長	0.1466099	0.7880016	0.0249109	0.1241801	-0.19287	-0.401282	0.2342139
座高	0.1376511	0.2942535	0.0476414	0.0038474	-0.231257	-0.641738	0.3888695
前腸骨棘高	0.1099028	0.9330724	-0.060135	0.0617218	-0.064992	-0.030796	0.1634755
上肢長	0.0299776	0.9023428	0.1363957	0.0326548	0.0224645	-0.222101	0.0702665
肩峰幅	0.5549963	0.2142560	0.0050609	0.0745873	-0.068627	-0.037693	0.6432561
腸骨稜幅	0.2636971	0.3607908	0.0440192	-0.04122	0.2422954	-0.193764	0.6934742
胸矢状径	0.8047959	0.2595891	0.0938970	-0.008661	0.2514118	-0.104463	-0.144165
胸横径	0.6933420	0.0799085	0.0580849	0.0488993	0.2262213	-0.094158	0.4434227
胸囲	0.8910140	0.1402527	0.0523923	0.0450206	0.1734336	-0.001209	0.2326615
上腕囲	0.8078074	-0.134438	-0.056269	0.0839202	0.1154926	-0.17805	0.1288507
下腿囲	0.7900122	-0.005906	0.0990419	0.2871149	0.1659265	-0.243825	0.1258376
三頭筋皮脂厚	0.2948644	-0.016009	0.0273401	0.0595926	0.8810504	-0.108182	0.0305376
肩甲下皮脂厚	0.4185372	-0.199896	0.0732691	0.0790775	0.7682625	0.0693752	0.1203572
前腕末端幅	0.2369467	0.2716955	0.2259972	0.0878568	0.1610130	-0.718585	-0.103898
大腿上顆幅	0.5296144	0.1912663	-0.008608	0.0664633	0.1966548	-0.57735	0.2445760
頭長	0.2478209	0.1355240	0.0160107	0.9291513	0.0917430	-0.072971	0.0055487
頭幅	-0.007411	0.0273751	0.9253288	-0.140057	-0.021601	0.0168069	-0.0263
下顎角幅	0.1600880	0.0469978	0.8193323	0.2135437	0.1145554	-0.227636	0.0847140

因子分析では主成分（Z_i）に相当する因子（f_i）の係数ベクトルを因子負荷量（Factor loading）といい，

0.0～±1.0の値をとる。この因子負荷量の大きいものに注目して因子の性格を考えていく。このあたりは，主成分分析と同じように，固有の技術，知識といった知恵を働かせなければならない。因子負荷量（絶対値）の大きいものをワクで囲ってみた。第1因子の因子負荷量は体重，胸矢状径，胸横径，胸囲，上腕囲，下腿囲で大きい。だから，これを体幹，四肢の太さの因子と考える。第2因子は身長，前腸骨棘高（下肢長），上肢長に大きい因子負荷量があらわれている。だから，これをからだ・四肢の長さの因子と考える。第3因子は頭幅，下顎角幅に大きい因子負荷量をもつ。これは頭部の幅の因子と考える。同様にして第4因子は頭部の長さの因子，第5因子は脂肪太りの因子，第6因子は肘，膝の関節の太さ（骨太）の因子，第７因子は肩幅，腰幅の因子負荷量が高いので体幹の幅の因子と考えていく。

先ほどの趣旨分分析の結果との違いはどうだろうか？太さと脂肪太りの因子が分かれたり，頭部の因子が分かれたりしている。さて，この因子に対応するような遺伝子があるのだろうか？