College Analysis 総合マニュアル

― 多変量解析１ ―

目次

１．実験計画法

1.1　実験計画法とは

2 群間の量的データの比較検定では、対応がない場合、t 検定、Welch の t 検定、Wilcoxon の順位和検定が利用され、対応がある場合、対応のある t 検定、Wilcoxon の符号付き順位和検定が利用されるが、3 群以上（2 群のとき実行しても問題はない）の比較の問題を取り扱うのが 1 元配置実験計画法である。1 元配置の問題は、2 群間の比較の拡張であるが、どの群間に差があるかまでを問題にする場合、以下で述べる多重比較の問題が生じる。

実験計画法にはこれ以外に 2 元配置以上の手法がある。これは 1 つの分類（変数）でのデータの差だけでなく、2 つ以上の分類間の互いの影響（交互作用という）も検討する手法である。1 元配置分散分析や 2 元配置分散分析の問題を 1 元比較の問題、2 元比較の問題ということもある。

1)　多重比較

まず多重比較について少し詳しく説明する。n 種の群間を比較する場合、以下の回数の比較が必要になる。

$$ {}_nC_2 = \frac{n(n-1)}{2} $$

これは例えば$n=5$の場合、${}_5C_2=10$回、$n=10$の場合、${}_{10}C_2=45$回の比較になる。前者の場合、差がある確率を5%として、10 回の比較をしたら、その中で差が偶然出る確率は50%になってしまう。そのため、通常の比較を繰り返すことは偶然の差を生み出す危険性を含んでいる。このような問題を多重比較の問題と言い、解決のために Fisher の LSD 法、Bonferroni の方法、それを修正した Holm-Bonferroni の方法、Turkey の方法、Sheffe の方法等、様々な方法が考えられている。

その中で最も簡単なものは、有意水準を$\alpha \div {}_nC_2$として検定を行う Bonferroni の方法である。この方法は非常に簡単でいつでも使える方法であるが、差を見つけにくいという弱点がある。これに対して、比較的差を見つけやすい方法として Fisher の LSD 法がある。これらの代表的な手法はソフトに組み込まれている。

2)　対応のない１元配置実験計画法

対応のない 1 元配置実験計画法及び Fisher の LSD 法について説明する。これは図 1 の手順で検定が行われる。

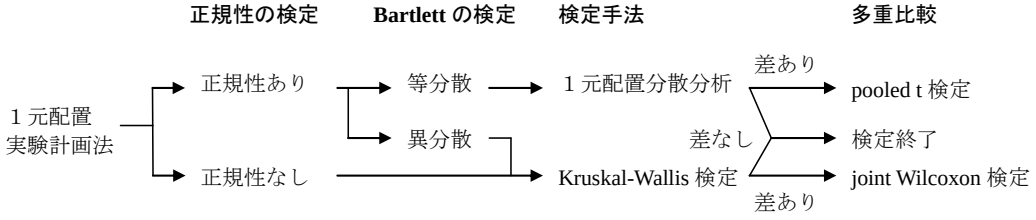


図１　対応のない１元配置実験計画法の構造

まず各変数について正規性の検定を行い、すべての変数に正規性が認められる場合は、多

変数の等分散の検定である Bartlett の検定に進む。これで等分散性が認められる場合は検定手法は 1 元配置分散分析となり、等分散性が認められない場合は順位検定の一種である Kruskal-Wallis 検定を利用する。正規性の検定で、正規性が認められない場合も同じく Kruskal-Wallis 検定となる。1 元配置分散分析及び Kruskal-Wallis 検定で差が見られない場合は検定を終了し、差が見られる場合は、それぞれ pooled t 検定及び joint Wilcoxon 検定でどの変数間に差があるか調べる。

3) 対応がある場合の 1 元配置実験計画法

対応がある場合は、データの正規性を調べ、正規性が認められる場合は、繰り返しのない 2 元配置分散分析、正規性が認められない場合は Freedman 検定を行う。

繰り返しのない 2 元配置分散分析はブロック（レコード）間の変動、1 つの変数内の群（水準）間の変動を分けて、群（水準）やブロック間の差を調べるものである。これは対応のある 1 元配置分散分析とも呼ばれる。

対応のある 1 元比較（繰返しのない 1 元比較）でブロック差が大きい場合や誤差の正規性に問題がある場合は、Friedman の順位検定を用いる。これはブロック毎にデータに順位を付け、群（水準）毎の順位和を用いて検定を行なうものである。

4) 2 元配置分散分析

繰り返しのある 2 元配置分散分析では 2 つの変数内の群（水準）間と変数間の交互作用を同時に検定する。変数の群（水準）の特別な組み合わせに意味がある場合に有効である。2 元配置分散分析は、正規性が認められ、各群（水準）やブロック間で分散が等しい場合に有効である。

1.2 プログラムの利用法

1.2.1 実験計画法

メニュー［分析－多変量解析他－実験計画法－実験計画法］で示される分析実行画面を図 2 に示す。

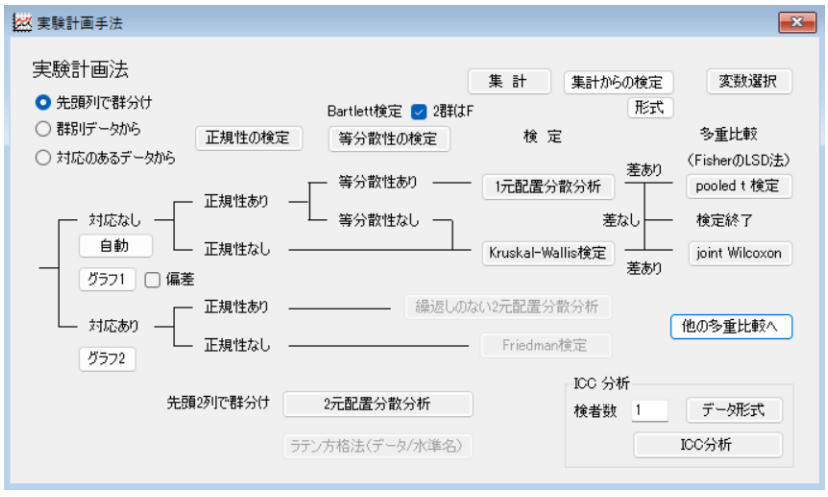


図 2　実験計画法分析実行画面

画面は基本統計の量的データの検定メニューのように､分析選択手順を図式化したものになっている。データは先頭列で群分けする場合と既に群別になっている場合と 2 通りから選択できる。コマンドボタン「集計」は群（水準）毎の基本統計量を出力する。図 3 に「等分散の検定」の出力画面を示す。特に 2 群の場合、F 検定を使うか Bartlett 検定を使うかは､「2 群は F」チェックボックスで選択できる。

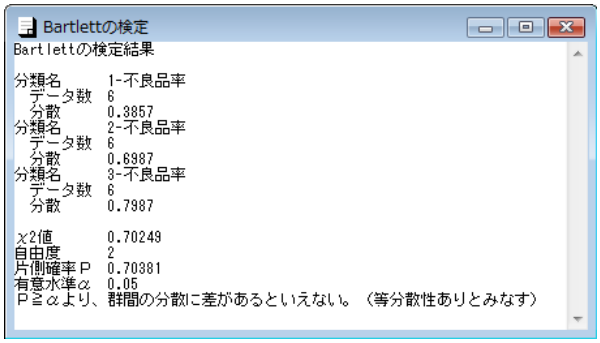


図 3　等分散の検定出力画面

図 4a と図 4b に「1 元配置分散分析」の検定結果と分散分析表の出力画面を示す。

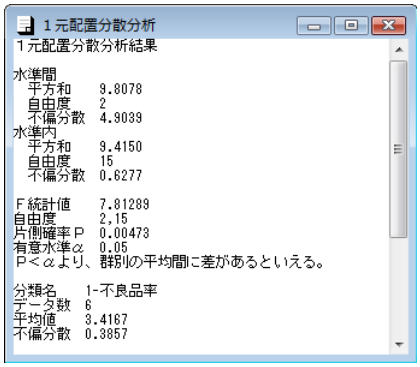


図 4a　1 元配置分散分析出力画面

図 4b　1 元配置分散分析表

また、図 5 に「Kruskal-Wallis 検定」の検定結果の出力画面を示す。

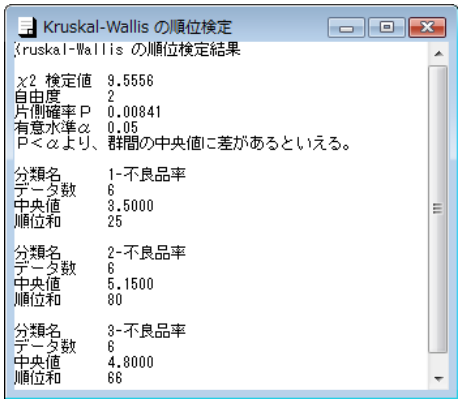
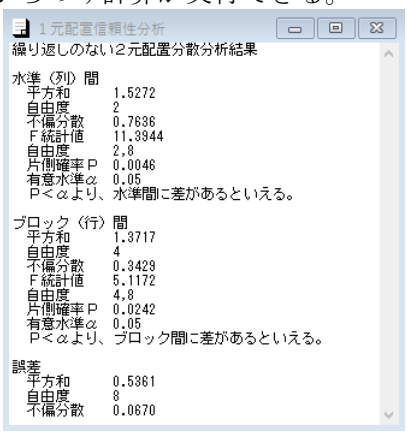


図 5　Kruskal-Wallis 検定出力画面

「繰返しのない 2 元配置分散分析」は、対応のある 1 元配置分散分析とも呼ばれる。「繰り返しのない 2 元配置分散分析」の出力結果と分散分析表をそれぞれ図 6a と図 6b に示す。この場合はブロックと群（水準）の交点に 1 つだけデータがある形式で、群分けされたデー

タからのみ計算が実行できる。



	平方和	自由度	不偏分散	F値	確率値
全変動	3.435	14			
水準(列)間	1.527	2	0.764	11.3944	0.0046
ブロック(行)間	1.372	4	0.343	5.1172	0.0242
誤差	0.536	8	0.067		

図 6a　2 元配置分散分析（繰り返しなし）　　図 6b　分析表（繰り返しなし）

対応のある 1 元比較の問題（繰返しのない 2 元比較の問題）で正規性に疑いがある場合やブロック間の平均の差が大きい場合、Friedman 検定を行なう。出力画面を図 7 に示す。

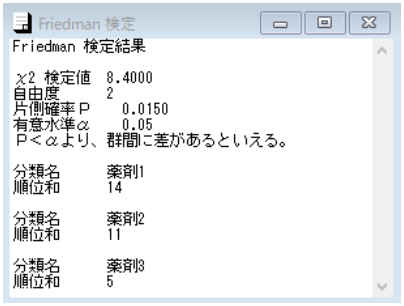
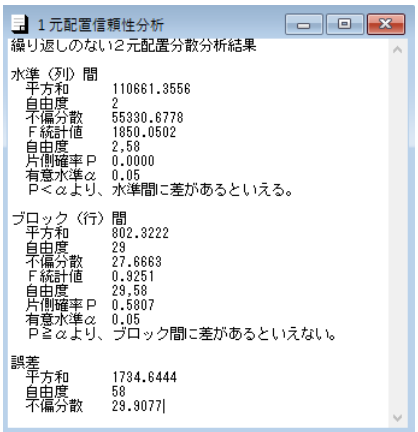
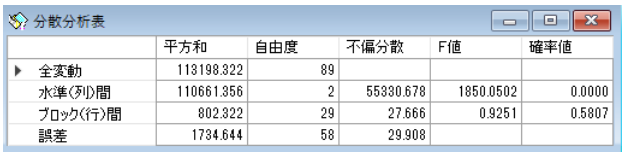


図 7 Friedman 検定出力画面

繰り返しがある場合の「2 元配置分散分析」の出力結果と分散分析表をそれぞれ図 8a と図 8b に示す。この場合、データは先頭 2 列で群分けされたものだけが利用できる。





	平方和	自由度	不偏分散	F値	確率値
全変動	113198.322	89			
水準(列)間	110661.356	2	55330.678	1850.0502	0.0000
ブロック(行)間	802.322	29	27.666	0.9251	0.5807
誤差	1734.644	58	29.908		

図 8a　2 元配置分散分析（繰り返しあり）　　図 8b　分散分析表（繰り返しあり）

データの処理順序の差も検出したい場合、ラテン方格法を利用する。これには処理順序を入力しておく必要があるため、データに加えて順序を「データ/順序」のように / で区切って入力する。このデータ形式の例を図 9 に示す。出力は群（水準）、ブロック、配置間の差

を検定した結果を、図 10a と図 10b のようにテキストと分散分析表の 2 種類で表示する。

図 9　ラテン方格法データ例

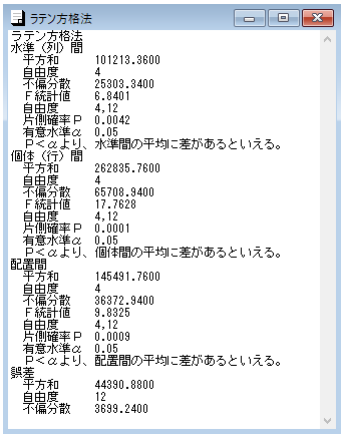


図 10a　ラテン方格法

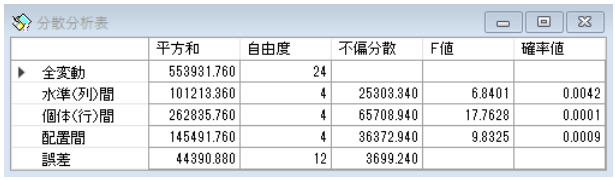


図 10b　分散分析表

多重比較については、正規性が認められる場合と認められない場合について、結合された不偏分散による t 検定(pooled　t 検定)と結合された順位による Wilcoxon の順位和検定(joint Wilcoxon 検定）の出力結果をそれぞれ図 10 と図 11 に示す。２群（水準）間の差の検定確率は各表の下に示される。

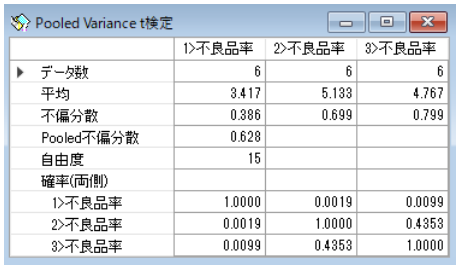


図 10　pooled t 検定出力結果

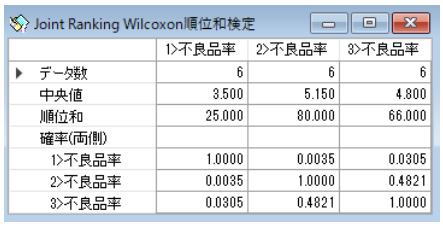


図 11　pooled Wilcoxon 検定出力結果

最後に、2 群の検定では集計データを元にした t 検定が組み込まれていたが、1 元配置分散分析ではこれまでできなかった。今回新しく、2 群の場合も含み、複数の変数に対して一括で処理するプログラムを作成した。そのデータ形式を図 12 に示す。群は縦に、標本数、平均値、標準偏差の順に複数群入力する。群の数は同一でなくてもよい（この場合最後の変数は 3 群)。データの形式は「群別データから」である。


データ編集 分散分析（集計から）.txt

区分	身長	体重	握力	上体起こし	長座体前屈	反復横とび	20mシャトル…	急歩	立ち幅とび	合計点
標本数	442	437	453	454	458	451	375	70	454	426
平均値	172.26	66.11	47.64	30.07	45.7	56.99	76.96	657.53	233.34	44.77
標準偏差	5.42	8.92	7.38	5.17	9.47	7.13	23.37	95.66	21.92	6.4
標本数	850	842	874	877	874	872	711	151	867	832
平均値	171.68	65.36	45.84	28.94	44.84	54.78	70.63	685.41	224.65	41.92
標準偏差	5.61	9.13	7.57	5.76	10.41	8.17	26.08	79.5	24.93	7.76
標本数										385
平均値										43.22
標準偏差										7.85

1/1 (1,1)　分析　備考

図 12　集計データからの分散分析データ

必要な変数を選んで、実行画面最上段の「集計からの検定」ボタンをクリックすると、図 13 のような結果が示される。


1元配置分散分析

	自由度1	自由度2	F値	P値
身長	1	1290	3.181	0.0748
体重	1	1277	1.972	0.1605
握力	1	1325	17.159	0.0000
上体起こし	1	1329	12.330	0.0005
長座体前屈	1	1330	2.180	0.1400
反復横とび	1	1321	23.673	0.0000
20mシャトルラン	1	1084	15.518	0.0001
急歩	1	219	5.155	0.0242
立ち幅とび	1	1319	39.265	0.0000
合計点	2	1640	20.893	0.0000

図 13　集計からの検定結果

ここには詳しい検定結果は表示されないが、1 つだけ変数を選んで、例えば「合格点」などを選んで「集計からの検定」ボタンをクリックすると、図 14 のように pooled t 検定まで含めた詳細な結果が表示される。


1元配置分散分析

	平方和	自由度	不偏分散	F値
全変動	93433.380	1642	56.902	20.8935
水準間	2321.515	2	1160.757	P値
水準内	91111.866	1640	55.556	0.0000
Pooled t 検定				
分散分析有意なら	群1	群2	群3	
群1	1.0000	0.0000	0.0031	
群2	0.0000	1.0000	0.0047	
群3	0.0031	0.0047	1.0000	

図 14　集計からの検定詳細結果

この形式での処理は、政府の調査資料などを検定する場合に使うと便利である。

1.2.2　多重比較

実験計画法の画面以外の多重比較については、別メニューで利用が可能である。図 2 の分析実行画面で「他の多重比較へ」ボタンをクリックすると図 15 のような多重比較実行メニューが表示される。

図 15　多重比較実行画面

ここでは、正規性（等分散性も含めた）がある場合とない場合に分けた、対応がない場合とある場合の Bonferroni の方法及びそれを修正した Holm-Bonferroni の方法、よく用いられる Turkey の方法が含まれている。ここで、Turkey の方法には、正規性と等分散性が必要である。また正規性または等分散性の条件が満たされない場合に用いられる手法として Scheffe の方法も含まれている。これには最初に Kruskal-Wallis の検定が必要である。

これらの検定の実行結果を分散分析 1.txt のデータを用いて以下に与えておく。Bonferroni の方法は 2 群の検定確率に比較確率を掛けて確率を与えてある。また、Holm-Bonferroni の方法は予め分析実行画面で有意水準を与えておき、それが満たされるかどうかで結果を表示している。以下の例の場合、有意水準を 0.05 に設定している。

最初に、正規性がある場合の pooled t 検定を用いた Bonferroni と Holm-Bonferroni の方法の結果を図 16 に与える。

pooled t Bonferroniの方法

	工場1	工場2	工場3
データ数	6	6	6
平均	3.417	5.133	4.767
不偏分散	0.386	0.699	0.799
Pooled不偏分散	0.628		
自由度	15		
確率*回数(両側)	工場1	工場2	工場3
工場1	1.0000	0.0058	0.0297
工場2	0.0058	1.0000	1.0000
工場3	0.0297	1.0000	1.0000

pooled t Holm-Bonferroniの方法

	工場1	工場2	工場3
データ数	6	6	6
平均	3.417	5.133	4.767
不偏分散	0.386	0.699	0.799
Pooled不偏分散	0.628		
自由度	15		
確率判定(両側)	工場1	工場2	工場3
工場1		p<0.05	p<0.05
工場2	p<0.05		n.s.
工場3	p<0.05	n.s.	

図 16　pooled t 検定の Bonferroni の方法と同 Holm-Bonferroni の方法

次に、正規性がある場合の対応のある t 検定を用いた Bonferroni と Holm-Bonferroni の方法の結果を図 17 に与える。

対応のあるt Bonferroniの方法

	工場1	工場2	工場3
データ数	6	6	6
平均	3.417	5.133	4.767
不偏分散	0.386	0.699	0.799
標準偏差	0.621	0.836	0.894
自由度	5	5	5
確率*回数(両側)	工場1	工場2	工場3
工場1	1.0000	0.0009	0.1455
工場2	0.0009	1.0000	1.0000
工場3	0.1455	1.0000	1.0000

対応のあるt Holm-Bonferroniの方法

	工場1	工場2	工場3
データ数	6	6	6
平均	3.417	5.133	4.767
不偏分散	0.386	0.699	0.799
標準偏差	0.621	0.836	0.894
自由度	5	5	5
確率判定(両側)	工場1	工場2	工場3
工場1		p<0.05	n.s.
工場2	p<0.05		n.s.
工場3	n.s.	n.s.	

図 17　対応のある t 検定の Bonferroni と Holm-Bonferroni の方法

次に、正規性がない場合の joint rank Wilcoxon 検定を用いた Bonferroni と Holm-Bonferroni の方法の結果を図 18 に与える。

Joint rank Bonferroniの方法

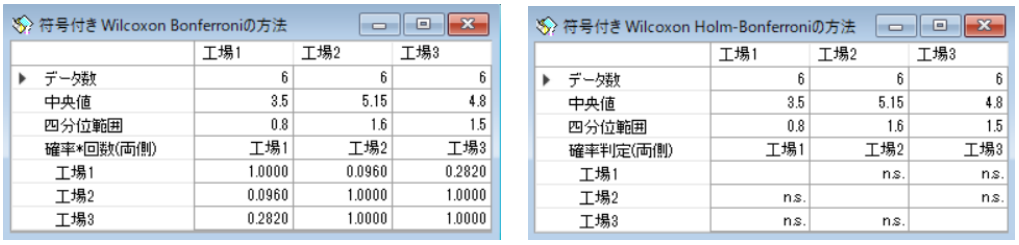
	工場1	工場2	工場3
データ数	6	6	6
中央値	3.5	5.15	4.8
順位和	25	80	66
確率*回数(両側)	工場1	工場2	工場3
工場1	1.0000	0.0104	0.0911
工場2	0.0104	1.0000	1.0000
工場3	0.0911	1.0000	1.0000

Joint rank Holm-Bonferroniの方法

	工場1	工場2	工場3
データ数	6	6	6
中央値	3.5	5.15	4.8
順位和	25	80	66
確率判定(両側)	工場1	工場2	工場3
工場1		p<0.05	n.s.
工場2	p<0.05		n.s.
工場3	n.s.	n.s.	

図 18　joint rank Wilcoxon 検定の Bonferroni と Holm-Bonferroni の方法

最後に、正規性がない場合の Wilcoxon 符号付き順位和検定を用いた Bonferroni と Holm-Bonferroni の方法の結果を図 19 に与える。



符号付き Wilcoxon Bonferroniの方法

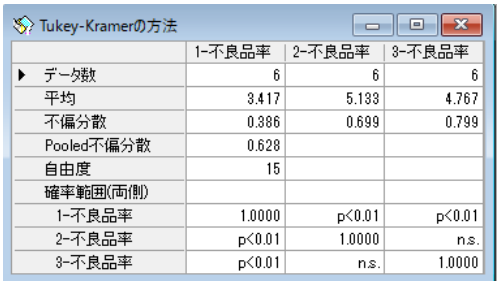
	工場1	工場2	工場3
データ数	6	6	6
中央値	3.5	5.15	4.8
四分位範囲	0.8	1.6	1.5
確率*回数(両側)	工場1	工場2	工場3
工場1	1.0000	0.0960	0.2820
工場2	0.0960	1.0000	1.0000
工場3	0.2820	1.0000	1.0000

符号付き Wilcoxon Holm-Bonferroniの方法

	工場1	工場2	工場3
データ数	6	6	6
中央値	3.5	5.15	4.8
四分位範囲	0.8	1.6	1.5
確率判定(両側)	工場1	工場2	工場3
工場1		n.s.	n.s.
工場2	n.s.		n.s.
工場3	n.s.	n.s.	

図 19　Wilcoxon 符号付き順位和検定の Bonferroni と Holm-Bonferroni の方法

よく利用される Turkey の方法と Scheffe の方法の分析結果を図 20 と図 21 に示す。Turkey の方法は数表を使うため結果の確率は定まった数値で表されていない。



Tukey-Kramerの方法

	1-不良品率	2-不良品率	3-不良品率
データ数	6	6	6
平均	3.417	5.133	4.767
不偏分散	0.386	0.699	0.799
Pooled不偏分散	0.628		
自由度	15		
確率範囲(両側)			
1-不良品率	1.0000	p<0.01	p<0.01
2-不良品率	p<0.01	1.0000	n.s.
3-不良品率	p<0.01	n.s.	1.0000

図 20　Turkey の方法検定結果

Scheffeの方法

	1-不良品率	2-不良品率	3-不良品率
データ数	6	6	6
平均	3.417	5.133	4.767
不偏分散	0.386	0.699	0.799
Pooled不偏分散	0.628		
自由度	15		
確率(片側)			
1-不良品率	1.0000	0.0070	0.0323
2-不良品率	0.0070	1.0000	0.7301
3-不良品率	0.0323	0.7301	1.0000

図 21　Scheffe の方法

例１　１元配置分散分析

３つの条件である商品の売上を調査したところ、Samples¥分散分析 ex.txt の結果を得た。各群に差があるといえるか、実験計画法を用いて有意水準 5%で判定せよ。

正規性の検定　　正規分布と［みなす・いえない］

等分散性の検定　　検定確率［ 0.9260 ］　等分散と［みなす・いえない］

検定名［ 1 元配置分散分析 ］　検定確率［ 0.0028 ］

判定　条件間に差があると［いえる・いえない］

差があるとするとどの条件間に差があるか。差がある条件同士を条件２＜条件３（これは実際の結果とは関係ない）のように不等号で表せ。

検定名［ pooled t 検定 ］

結果［ 条件１＜条件２，条件３＜条件２ ］

例２　対応がある場合の 1 元配置分散分析

３つの条件である商品の売上を調査したところ、分散分析 ex.txt の結果を得た。各データに対応があるとして差があるか検定せよ。

条件１　115, 110, 108, 114, 120, 116, 108, 112, 115, 122

条件２　121, 118, 124, 117, 119, 130, 121, 115, 118, 119

条件３　116, 112, 120, 111, 112, 108, 114, 119, 104, 113

注）正規性の有無により、（繰り返しのない）２元配置分散分析か Friedman 検定を利用する。これは repeated measured 1 元配置分散分析、repeated measured Kruskal-Wallis 検定とも呼ばれている。

例３　2 元配置分散分析

分散分析 4（2 元配置）.txt は作物の品種と肥料の組み合わせによる収穫量を表したデータである。2 元配置分散分析を用いて判定せよ。

注）2 元配置分散分析では、分類データの組み合わせによる交互作用が分かる。

品種水準間

検定確率［ 0.6049 ］　水準間に差があると　［いえる・いえない］

肥料水準間

検定確率［ 0.3556 ］　水準間に差があると　［いえる・いえない］

交互作用

検定確率［ 0.0184 ］　交互作用に差があると［いえる・いえない］

問題１（分散分析 1.txt）

分散分析 1.txt は３つの工場群の不良品率を与えたものである。各群に差があるといえる

か、実験計画法を用いて有意水準 5%で検討せよ。

正規性の検定　　正規分布と［みなす・いえない］

等分散性の検定　　検定確率［　　］　等分散と［みなす・いえない］

検定名［　　］　検定確率［　　］

判定　工場群間の不良品率に差があると［いえる・いえない］

差があるとするとどの条件間に差があるか。差がある条件同士を工場２＜工場３（これは実際の結果とは関係ない）のように不等号で表せ。

検定名［　　］

結果［　　］

問題２（分散分析 2.txt）

分散分析 2.txt は４つの群のデータであるが、各群に差があるといえるか、実験計画法を用いて有意水準 5%で検討せよ。

正規性の検定　　正規分布と［みなす・いえない］

等分散性の検定　　検定確率［　　］　等分散と［みなす・いえない］

検定名［　　］　検定確率［　　］

判定　群間に差があると［いえる・いえない］

差があるとするとどの群間に差があるか。差がある群同士を群２＜群３（これは実際の結果とは関係ない）のように不等号で表せ。

検定名［　　］

結果［　　］

問題３（分散分析 3.txt）

分散分析 3.txt は３群のデータであるが、各群に差があるといえるか、実験計画法を用いて有意水準 5%で検討せよ。

正規性の検定　　正規分布と［みなす・いえない］

等分散性の検定　　検定確率［　　］　等分散と［みなす・いえない］

検定名［　　］　検定確率［　　］

判定　群間に差があると［いえる・いえない］

差があるとするとどの群間に差があるか。差がある群同士を群２＜群３（これは実際の結果とは関係ない）のように不等号で表せ。

検定名［　　］

結果［　　］

演習（多変量演習 1.txt）

ある４つの中学について英語・数学・国語の試験結果を調べた。多変量演習 1.txt のデータを読み込んで、以下の質問に答えよ。但し、検定は有意水準 5%とすること。

１．中学　　1）A 中学　2）B 中学　3）C 中学　4）D 中学

２．英語点数

３．数学点数

４．国語点数

1）数学について、各中学の平均（中央）値に差があるといえるか。

検定名［　　］　検定確率［　　］

判定　平均（中央）値に差があると［いえる・いえない］。

2）数学について各中学の平均（中央）値に差があるとすると、A, B, C, D どの中学の間に差があるか調べて A<B のように不等号で表せ。（差がある場合のみ）

検定名［　　］

結果［　　］

3）国語について、各中学間の平均（中央）値に差があるといえるか。

検定名［　　］　検定確率［　　］

判定　平均（中央）値に差があると［いえる・いえない］。

4）国語について、各中学間の平均（中央）値に差があるとすると、A, B, C, D どの中学の間に差があるか調べて A<B のように不等号で表せ。（差がある場合のみ）

検定名［　　］

結果［　　］

5）３教科の平均（中央）値に差があるといえるか。対応は考えないものとせよ。

検定名［　　］　検定確率［　　］

判定　平均（中央）値に差があると［いえる・いえない］。

6）３教科の平均（中央）値に差があるとすれば、どの教科の間に差があるか調べて英語＜数学のように不等号で表せ。（差がある場合のみ）

検定名［　　］　結果［　　］

問題１解答

分散分析 1.txt は３つの工場群の不良品率を与えたものである。各群に差があるといえるか、実験計画法を用いて有意水準 5%で検討せよ。

正規性の検定　　正規分布と［みなす・いえない］

等分散性の検定　　検定確率［ 0.7038 ］　等分散と［みなす・いえない］

検定名［ １元配置分散分析 ］　検定確率［ 0.0047 ］

判定　工場群間の不良品率に差があると［いえる・いえない］

差があるとするとどの条件間に差があるか。

検定名［ pooled t 検定 ］

結果［ 工場１＜工場２，工場１＜工場３ ］

問題２解答

分散分析 2.txt は４つの群のデータであるが、各群に差があるといえるか、実験計画法を用いて有意水準 5%で検討せよ。

正規性の検定　　正規分布と［みなす・いえない］
等分散性の検定　　検定確率［ 0.0046 ］　等分散と［みなす・いえない］
検定名［ Kruskal-Wallis 検定 ］　検定確率［ 0.2371 ］
判定　群間に差があると［いえる・いえない］
差があるとするとどの群間に差があるか。
検定名［　　］　← やらない（書かない）
結果［　　］

問題３解答

分散分析 3.txt は３群のデータであるが、各群に差があるといえるか、実験計画法を用いて有意水準 5%で検討せよ。

正規性の検定　　正規分布と［みなす・いえない］
等分散性の検定　　検定確率［　　］　等分散と［みなす・いえない］
検定名［ Kruskal-Wallis 検定 ］　検定確率［ 0.0213 ］
判定　群間に差があると［いえる・いえない］
差があるとするとどの群間に差があるか。
検定名［ joint Wilcoxon 検定 ］
結果［ 群２＜群３ ］

演習解答（多変量演習 1.txt）

１）数学について、各中学の平均（中央）値に差があるといえるか。
　検定名［ １元配置分散分析 ］　検定確率［ 0.0000 ］
　判定　平均（中央）値に差があると［いえる・いえない］。

２）数学について各中学の平均（中央）値に差があるとすると、A, B, C, D どの中学の間に差があるか調べて A<B のように不等号で表せ。（差がある場合のみ）
　検定名［ pooled t 検定 ］
　結果［ A<B，C<B，D<B，D<A，D<C ］

３）国語について、各中学間の平均（中央）値に差があるといえるか。
　検定名［ Kruskal-Wallis 検定 ］　検定確率［ 0.0745 ］
　判定　平均（中央）値に差があると［いえる・いえない］。

４）国語について、各中学間の平均（中央）値に差があるとすると、A, B, C, D どの中学の間に差があるか調べて A<B のように不等号で表せ。（差がある場合のみ）
　検定名［　　］　←やらない（書かない）
　結果［　　］

５）３教科の平均（中央）値に差があるといえるか。対応は考えないものとせよ。
　検定名［ Kruskal-Wallis 検定 ］　検定確率［ 0.0009 ］
　判定　平均（中央）値に差があると［いえる・いえない］。

６）３教科の平均（中央）値に差があるとすれば、どの教科の間に差があるか調べて英語＜数学のように不等号で表せ。（差がある場合のみ）
　検定名［ joint Wilcoxon 検定 ］　結果［ 英語＜国語，数学＜国語 ］

1.3　実験計画法の理論

実験計画法は、異なるいくつかの条件下でデータを求め、その間に差があるかどうか検討する手法の総称である。それぞれの検定について方法を解説しておこう。

1)　１元配置分散分析

１元比較の場合、データは表１の形で与えられる。ここに水準数は p、水準 i のデータ数は n_i で与えられ、データは一般に $x_{i\lambda}$ で表わされる。

表１　１元比較のデータ

水準１	水準２	…	水準 p
x_{11}	x_{21}	…	x_{p1}
x_{12}	x_{22}	…	x_{p2}
⋮	⋮		⋮
x_{1n_1}	x_{2n_2}	…	x_{pn_p}

位置母数の比較は正規性と等分散性の有無によって１元配置分散分析か、Kruskal-Wallis 検定かに分かれる。正規性が認められ、多群間の等分散性が認められる場合には、１元配置分散分析が利用できる。この等分散性の検定には Bartlett 検定を利用することができる。

１元配置分散分析のデータ $x_{i\lambda}$ は、水準 i に固有な値 α_i と誤差 $\varepsilon_{i\lambda}$ を用いて以下のように表わされると考える。

$x_{i\lambda} = \mu + \alpha_i + \varepsilon_{i\lambda}$,　　$\varepsilon_{i\lambda} \sim N(0,\sigma^2)$ 分布［異なる i,λ について独立］

データの全変動 S は、水準内変動 S_E 及び水準間変動 S_P を用いて以下のように表わされる。

$$S=\sum_{i=1}^{p}\sum_{\lambda=1}^{n_i}(x_{i\lambda}-\bar{x})^2=\sum_{i=1}^{p}\sum_{\lambda=1}^{n_i}(x_{i\lambda}-\bar{x}_i)^2+\sum_{i=1}^{p}n_i(\bar{x}_i-\bar{x})^2=S_E+S_P$$

誤差 $\varepsilon_{i\lambda}$ の正規性から、それぞれの変動は以下の分布に従うことが分かる。

$S/\sigma^2 \sim \chi^2_{n-1}$ 分布，　$S_E/\sigma^2 \sim \chi^2_{n-p}$ 分布，　$S_P/\sigma^2 \sim \chi^2_{p-1}$ 分布

１元配置分散分析は、$\alpha_i = 0$ として、以下の性質を利用する。

$$F=\frac{S_P/(p-1)}{S_E/(n-p)} \sim F_{p-1,n-p} \text{ 分布}$$

2) Kruskal-Wallis の順位検定

Kruskal-Wallis の順位検定は、データの分布型によらず、p 種類の水準の中央値に差があるかどうか判定する手法である。まず、全データの小さい順に順位 $r_{i\lambda}$ を付け、水準ごとの順位和 w_i を求める。但し、同じ大きさのデータにはそれらに順番があるものとした場合の順位の平均値を与える。検定には各水準の中央値が等しいとして以下の性質を利用する。ここで τ_a は小さい方から a 番目の同順位データの数である。

$$H=\frac{12}{N(N+1)}\sum_{i=1}^{p}n_i\left(\frac{w_i}{n_i}-\frac{N+1}{2}\right)^2=\frac{12}{N(N+1)}\sum_{i=1}^{p}\frac{\left[w_i-n_i(N+1)/2\right]^2}{n_i}\sim\chi^2_{p-1}$$

$$\rightarrow\frac{12}{N(N+1)}\sum_{i=1}^{p}n_i\left(\left|\frac{w_i}{n_i}-\frac{N+1}{2}\right|-\frac{1}{2n_i}\right)^2\left[1-\frac{1}{N(N^2-1)}\sum_{a=1}^{e}\tau_a(\tau_a^2-1)\right]^{-1}$$

$$=\frac{12}{N(N+1)}\sum_{i=1}^{p}\frac{\left[|w_i-n_i(N+1)/2|-1/2\right]^2}{n_i}\left[1-\frac{1}{N(N^2-1)}\sum_{a=1}^{e}\tau_a(\tau_a^2-1)\right]^{-1}$$

上が補正なし、下が Yates の連続補正と同順位の補正を加えたものである。

3) Bartlett の検定

多群間の等分散の検定である Bartlett の検定は、各水準の母分散が等しいとして以下の性質を利用する。

$$\chi^2=\frac{1}{C}\left[(n-p)\log V_E-\sum_{i=1}^{p}(n_i-1)\log V_i\right]\sim\chi^2_{p-1}\text{ 分布}$$

ここに、V_E，V_i，C は n を全データ数として以下のように与えられる。

$$V_E=\frac{1}{n-p}\sum_{i=1}^{p}\sum_{\lambda=1}^{n_i}(x_{i\lambda}-\bar{x}_i)^2,\qquad V_i=\frac{1}{n_i-1}\sum_{\lambda=1}^{n_i}(x_{i\lambda}-\bar{x}_i)^2,$$

$$C=1+\frac{1}{3(p-1)}\left[\sum_{j=1}^{p}\frac{1}{n_j-1}-\frac{1}{n-p}\right]$$

4) ２元配置分散分析

２元比較の場合、２つの水準間または水準とブロック間の差を同時に検定する。前者は２つの水準の交点に複数のデータを含んだデータ構造であり、繰り返しのある場合とも言われる。後者は水準とブロックの交点に完備乱塊法によって得た１つのデータが含まれ、繰り返しのない場合とも言われる[8]。２元配置分散分析は、正規性が認められ、各水準やブロック間で分散が等しい場合にのみ有効である。以下２つの場合に分けて分析法について説明する。

表２　２元配置分散分析（繰り返しあり）

	水準 Q_1	…	水準 Q_s
水準 P_1	x_{111} : $x_{11n_{11}}$	… … …	x_{1s1} : $x_{1sn_{1s}}$
:	:	:	:
水準 P_r	x_{r11} : $x_{r1n_{r1}}$	… … …	x_{rs1} : $x_{rsn_{rs}}$

まず繰り返しがある場合を考える。データは表 2 の形式で与えられる。各データは水準 P_i に固有の量を α_i、水準 Q_j に固有の量を β_j、水準 P_i と水準 Q_j の相互作用を γ_{ij}、誤差を $\varepsilon_{ij\lambda}$ として、以下のように表わせると考える。

$x_{ij\lambda} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij\lambda}$，　$\varepsilon_{ij\lambda} \sim N(0, \sigma^2)$ 分布［異なる i, j, λ に対して独立］

但し、各パラメータには以下の条件を付ける。

$$\sum_{i=1}^{r} n_{i\bullet}\alpha_i = 0,\ \sum_{j=1}^{s} n_{\bullet j}\beta_j = 0,\ \sum_{i=1}^{r} n_{ij}\gamma_{ij} = 0,\ \sum_{j=1}^{s} n_{ij}\gamma_{ij} = 0$$

ここにデータ数に関しては以下の記法を用いている。

$$n_{i\bullet} = \sum_{j=1}^{s} n_{ij},\ \ n_{\bullet j} = \sum_{i=1}^{r} n_{ij},\ \ n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$$

各水準及び全体のデータ平均を $\bar{x}_{ij}$，$\bar{x}_{i\bullet}$，$\bar{x}_{\bullet j}$，$\bar{x}$ として、全変動 S、水準 P 間の変動 S_P、水準 Q 間の変動 S_Q、相互作用の変動 S_I、水準内変動 S_E を以下で与えると、

$$S = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{\lambda=1}^{n_{ij}} (x_{ij\lambda} - \bar{x})^2,\ \ S_P = \sum_{i=1}^{r} n_{i\bullet}(\bar{x}_{i\bullet} - \bar{x})^2,\ \ S_Q = \sum_{j=1}^{s} n_{\bullet j}(\bar{x}_{\bullet j} - \bar{x})^2,$$

$$S_I = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(\bar{x}_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x})^2,\ \ S_E = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{\lambda=1}^{n_{ij}} (x_{ij\lambda} - \bar{x}_{ij})^2,$$

全変動 S はその他の変動を用いて以下のように表わされる。

$S = S_P + S_Q + S_I + S_E (+ S_{Cross})$

ここに S_{Cross} は繰り返し数が一定のとき 0 になる。しかし、繰り返し数が一定でない場合、この値は 0 にならず、分散分析の分離が正確には行えない。この場合、我々のプログラムでは、重回帰分析の考え方を用いた方法で Type3 と呼ばれる手法を使って計算を行っている。詳しくは節を改めて説明する。

水準間の差や相互作用の有無を検定するためには、以下の性質を利用する。

$\alpha_i = 0$ のとき　　$F_P = \dfrac{S_P/(r-1)}{S_E/(n-rs)} \sim F_{r-1,\,n-rs}$ 分布　　（水準 P 間の差）

$\beta_j = 0$ のとき　　$F_Q = \dfrac{S_Q/(s-1)}{S_E/(n-rs)} \sim F_{s-1,\,n-rs}$ 分布　　（水準 Q 間の差）

$\gamma_{ij} = 0$ のとき　　$F_I = \dfrac{S_I/(r-1)(s-1)}{S_E/(n-rs)} \sim F_{(r-1)(s-1),\,n-rs}$ 分布　　（相互作用）

もう１つの２元配置分散分析はブロック毎に無作為化されたデータを用いて、水準やブロック間の差を調べるもので、繰り返しのない場合と呼ばれている。これは対応のある 1 元配置分散分析とも呼ばれ、データは表 3 のようにブロックと水準の交点に１つだけ値が入る。

表３　２元配置分散分析（繰り返しなし）

	水準１	水準２	…	水準 s
ブロック１	x_{11}	x_{12}	…	x_{1s}
ブロック２	x_{21}	x_{22}	…	x_{2s}
:	:	:		:
ブロック r	x_{r1}	x_{r2}	…	x_{rs}

水準 j に固有な量を α_j、ブロック i に固有な量を β_i、誤差を ε_{ij} として、データ x_{ij} を以下のように表わす。

$x_{ij} = \mu + \alpha_j + \beta_i + \varepsilon_{ij}$,　　$\varepsilon_{ij} \sim N(0,\sigma^2)$ 分布［異なる i, j に対して独立］

但し、パラメータ α_j, β_i には以下の条件を付ける。

$$\sum_{j=1}^{s}\alpha_j = 0,\ \sum_{i=1}^{r}\beta_i = 0$$

水準、ブロック及び全体の平均を、$\bar{x}_{\bullet j}$, $\bar{x}_{i\bullet}$, $\bar{x}$ として、全変動 S、水準間の変動 S_P、ブロック間の変動 S_B、誤差変動 S_E を以下で与えると、

$$S = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij}-\bar{x})^2,\ \ S_P = \sum_{i=1}^{r}\sum_{j=1}^{s}(\bar{x}_{\bullet j}-\bar{x})^2,\ \ S_B = \sum_{i=1}^{r}\sum_{j=1}^{s}(\bar{x}_{i\bullet}-\bar{x})^2,$$

$$S_E = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij}-\bar{x}_{i\bullet}-\bar{x}_{\bullet j}+\bar{x})^2,$$

全変動 S はその他の変動を用いて以下のように表わされる。

$S = S_P + S_B + S_E$

水準間やブロック間の差を検定するためには、以下の性質を利用する。

$\alpha_j = 0$ のとき　　$F_P = \dfrac{S_P/(s-1)}{S_E/(r-1)(s-1)} \sim F_{s-1,(r-1)(s-1)}$ 分布　（水準間の差）

$\beta_i = 0$ のとき　　$F_B = \dfrac{S_B/(r-1)}{S_E/(r-1)(s-1)} \sim F_{r-1,(r-1)(s-1)}$ 分布　（ブロック間の差）

5）Friedman の順位検定

対応のある 1 元比較（繰返しのない 2 元比較）でブロック差が大きい場合や誤差の正規性に問題がある場合は、Friedman の順位検定を用いる。これは各ブロック毎にデータに順位を付け、水準毎の順位和を用いて検定を行なうものである。今、水準 j の順位和を w_j とし、水準間に差がないことを仮定して、以下の性質を用いる。

$$D = \frac{12}{rs(s+1)}\sum_{j=1}^{s}\left[w_j - r(s+1)/2\right]^2 = \frac{12}{rs(s+1)}\sum_{j=1}^{s}w_j^2 - 3r(s+1) \sim \chi^2_{s-1}$$

$$\to \frac{12}{rs(s+1)}\sum_{j=1}^{s}\left[\left|w_j - r(s+1)/2\right| - 1/2\right]^2 \left[1 - \frac{1}{rs(s^2-1)}\sum_{i=1}^{r}\sum_{j=1}^{e_i}\tau_{ij}(\tau_{ij}^2-1)\right]^{-1}$$

一般に Friedman 検定は対応のある場合の Wilcoxon の符号付順位和検定の拡張のように

考えられがちだが、群間で順位を付ける理論構成から、むしろ McNemar 検定の拡張と言ってもよい。

6）ラテン方格法

実験順序によって結果に影響が出るような場合、それぞれの個体に対する処理（水準と呼ぶ）を順序を変えて１回ずつ施す方法がラテン方格法である。表 4 にデータとその処理順序（配置と呼ぶ）の例を示す。

表 4 ラテン方格法のデータと処理順序の例

	水準１	水準２	水準３	水準４
個体１	$x_{11(1)}$	$x_{12(2)}$	$x_{13(3)}$	$x_{14(4)}$
個体２	$x_{21(2)}$	$x_{22(3)}$	$x_{23(4)}$	$x_{24(1)}$
個体３	$x_{31(3)}$	$x_{32(4)}$	$x_{33(1)}$	$x_{34(2)}$
個体４	$x_{41(4)}$	$x_{42(1)}$	$x_{43(2)}$	$x_{44(3)}$

配置は、データの添え字に付いた括弧内の数字で表わすが、配置 k は各水準と各個体に一度だけ現れ、水準 j と個体 i による関数とみなすことができる。データ $x_{ij(k)}$ は、水準 j に固有な量を α_j、個体 i に固有な量を β_i、配置差に固有な量を γ_k として、以下のように表わせるものとする。

$x_{ij(k)} = \mu + \alpha_j + \beta_i + \gamma_k + \varepsilon_{ijk}$，　$\varepsilon_{ijk} \sim N(0, \sigma^2)$ 分布［異なる i, j, k に対して独立］

但し、パラメータ α_j，β_i，γ_k には以下の条件を付ける。

$$\sum_{j=1}^{r} \alpha_j = 0,\ \sum_{i=1}^{r} \beta_i = 0,\ \sum_{k=1}^{r} \gamma_k = 0$$

今後の計算のために、水準別合計 $T_{\bullet j}$，個体別合計 $T_{i\bullet}$，全合計 T を以下のように与える。

$$T_{\bullet j} = \sum_{i=1}^{r} x_{ij(k)},\ T_{i\bullet} = \sum_{j=1}^{r} x_{ij(k)},\ T = \sum_{i=1}^{r}\sum_{j=1}^{r} x_{ij(k)}$$

また、順序 k が付いたデータの合計 T_k も求めておく。さて $C = T^2/r^2$ とおいて、全変動 S、水準間の変動 S_P、個体間の変動 S_B、配置による変動 S_R を以下で与える。

$$S = \sum_{i=1}^{r}\sum_{j=1}^{r} X_{ij(k)}^2 - C,\ S_P = \frac{1}{r}\sum_{j=1}^{r} T_{\bullet j}^2 - C,\ S_B = \frac{1}{r}\sum_{i=1}^{r} T_{i\bullet}^2 - C,\ S_R = \frac{1}{r}\sum_{k=1}^{r} T_k^2 - C$$

これらの変動から誤差変動 S_E を以下のように定義する。

$$S_E = S - S_P - S_B - S_R$$

水準間の差や個体間の差及び配置による差の検定は、それぞれ以下の性質を利用する。

$\alpha_j = 0$ のとき、　$F_P = \dfrac{S_P/(r-1)}{S_E/(r-1)(r-2)} \sim F_{r-1,(r-1)(r-2)}$ 分布

$\beta_i = 0$ のとき、　　$F_B = \dfrac{S_B/(r-1)}{S_E/(r-1)(r-2)} \sim F_{r-1,(r-1)(r-2)}$ 分布

$\gamma_k = 0$ のとき、　　$F_R = \dfrac{S_R/(r-1)}{S_E/(r-1)(r-2)} \sim F_{r-1,(r-1)(r-2)}$ 分布

7）多重比較

１元比較の場合、１元配置分散分析も Kruskal-Wallis の順位検定も水準間に差があることは分かってもどこに差があるのか判定することはできない。また、p 個の水準から２つの水準を選んで 2 群間の差の検定を行なうことはできるが、${}_pC_2$ 回の検定を行なうことによる有意水準の解釈には問題がある。このような多重比較の場合にどのような検定を行なうかについて、Bonferroni の方法、Tukey の方法、Dunnet の方法等様々な検定方法が考えられてきたが、ここではその中で比較的有効と考えられる結合された（pooled）不偏分散による t 検定及び結合された順位による Wilcoxon の順位和検定をプログラム化した。実際の検定では Fisher の LSD 法を用いて、それぞれ 1 元配置分散分析や Kruskal-Wallis の順位検定と併用する。

結合された不偏分散による t 検定

データは表 1 の形式であり、水準 i のデータ数を n_i、平均を $\bar{x}_i$、不偏分散を s_i^2 として、水準 i, j の差について考える。結合された不偏分散 s^2 は以下のように与えられる。

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{p}(n_i - 1)s_i^2$$

ここに全データ数を n としている。検定には以下の性質を利用する。

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{s\sqrt{\dfrac{1}{n_i} + \dfrac{1}{n_j}}} \sim t_{n-p} \text{ 分布}$$

結合された順位による Wilcoxon の順位和検定

データは上と同様に表 1 の形式であるが、全データの小さい順に順位を付ける。水準 i の順位合計を w_i とし、データ数が十分多いとして以下の性質を利用する。

$$Z_{ij} = \frac{w_i/n_i - w_j/n_j}{\sqrt{N(N+1)\left(1/n_i + 1/n_j\right)/12}} \sim N(0,1)$$

$$\rightarrow \frac{\left|w_i/n_i - w_j/n_j\right| - \dfrac{1}{2}\left(1/n_i + 1/n_j\right)}{\sqrt{N(N+1)\left(1/n_i + 1/n_j\right)/12}}\left[1 - \frac{1}{N(N^2-1)}\sum_{i=1}^{e}\tau_i(\tau_i^2 - 1)\right]^{-1/2}$$

上が補正なしの場合、下が Yates の連続補正と同順位の補正を加えた場合である。

ICC（Intraclass correlation coefficient）分析について

ICC 分析については、作ってはいるが著者の理解が不十分であるため、以下の参考文献を参照してもらいたい。

今井樹，潮見泰藏，理学療法研究における“評価の信頼性”の検査法，理学療法科学 19(3):261-265,2004

1.4　繰り返し数の一定でない２元配置分散分析について

２元配置分散分析のデータは水準 P_i に固有の量 α_i、水準 Q_j に固有の量 β_j、水準 P_i と水準 Q_j の相互作用に固有の量 γ_{ij}、誤差を $\varepsilon_{ij\lambda}$ として、以下のように表わせると考える。

$$x_{ij\lambda} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij\lambda}, \qquad \varepsilon_{ij\lambda} \sim N(0,\sigma^2) \text{ 分布［異なる } i,j,\lambda \text{ に対して独立］}$$

但し、各パラメータには以下の条件を付ける。

$$\sum_{i=1}^{r} n_{i\cdot}\alpha_i = 0,\ \sum_{j=1}^{s} n_{\cdot j}\beta_j = 0,\ \sum_{i=1}^{r} n_{ij}\gamma_{ij} = 0,\ \sum_{j=1}^{s} n_{ij}\gamma_{ij} = 0$$

ここにデータ数に関しては以下の記法を用いている。

$$n_{i\cdot} = \sum_{j=1}^{s} n_{ij},\ n_{\cdot j} = \sum_{i=1}^{r} n_{ij},\ n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$$

各水準及び全体のデータ平均を $\bar{x}_{ij}$，$\bar{x}_{i\cdot}$，$\bar{x}_{\cdot j}$，$\bar{x}$ として、全変動 S、水準 P 間の変動 S_P、水準 Q 間の変動 S_Q、相互作用の変動 S_I、水準内変動 S_E を以下で与えると、

$$S = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{\lambda=1}^{n_{ij}}(x_{ij\lambda}-\bar{x})^2,\ S_P = \sum_{i=1}^{r} n_{i\cdot}(\bar{x}_{i\cdot}-\bar{x})^2,\ S_Q = \sum_{j=1}^{s} n_{\cdot j}(\bar{x}_{\cdot j}-\bar{x})^2,$$

$$S_I = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(\bar{x}_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})^2,\ S_E = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{\lambda=1}^{n_{ij}}(x_{ij\lambda}-\bar{x}_{ij})^2$$

全変動 S はその他の変動 S_{Cross} を加えて以下のように表わされる。

$$\begin{aligned}
S &= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{\lambda=1}^{n_{ij}}(x_{ij\lambda}-\bar{x})^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{\lambda=1}^{n_{ij}}\left[(x_{ij\lambda}-\bar{x}_{ij})+(\bar{x}_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})+(\bar{x}_{i\cdot}-\bar{x})+(\bar{x}_{\cdot j}-\bar{x})\right]^2 \\
&= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{\lambda=1}^{n_{ij}}(x_{ij\lambda}-\bar{x}_{ij})^2 + \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(\bar{x}_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})^2 \\
&\quad + \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(\bar{x}_{i\cdot}-\bar{x})^2 + \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(\bar{x}_{\cdot j}-\bar{x})^2 + 2\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(\bar{x}_{i\cdot}-\bar{x})(\bar{x}_{\cdot j}-\bar{x}) \\
&\quad + 2\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(\bar{x}_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})(\bar{x}_{i\cdot}-\bar{x}) + 2\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(\bar{x}_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})(\bar{x}_{\cdot j}-\bar{x}) \\
&= S_P + S_Q + S_I + S_E + S_{Cross}
\end{aligned}$$

ここで、最初に与えた関係とデータの対応は以下である。

$$\varepsilon_{ij\lambda} = x_{ij\lambda}-\bar{x}_{ij},\ \gamma_{ij} = \bar{x}_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x},\ \alpha_i = \bar{x}_{i\cdot}-\bar{x},\ \beta_j = \bar{x}_{\cdot j}-\bar{x}$$

ここでもし、$n_{ij} = n = const.$ ならば、$S_{Cross} = 0$ となるが、一般には $S_{Cross} \neq 0$ である。これが繰り返し数の異なる場合の 2 元配置分散分析の計算を困難にする理由である。ここではこのような問題に対して我々が採用している重回帰分析の考え方を用いた Type3 と呼ばれる手法を紹介する。

この問題を扱う前に分散分析と重回帰分析の関係について見ておく。今、図 1 のようなデータを考えるものとする。


データ編集 分散分析と重回帰分析.txt

	工場	不良品率	工場_1	工場_2	工場_3	工場_1b	工場_2b
1	1	3.1	1	0	0	1	0
2	1	4.1	1	0	0	1	0
3	1	3.3	1	0	0	1	0
4	1	3.9	1	0	0	1	0
5	1	3.7	1	0	0	1	0
6	1	2.4	1	0	0	1	0
7	2	4.7	0	1	0	0	1
8	2	5.6	0	1	0	0	1
9	2	4.3	0	1	0	0	1
10	2	5.9	0	1	0	0	1
11	2	6.1	0	1	0	0	1
12	2	4.2	0	1	0	0	1
13	3	5.1	0	0	1	-1	-1
14	3	3.7	0	0	1	-1	-1
15	3	4.5	0	0	1	-1	-1
16	3	6.0	0	0	1	-1	-1

1/2 (1,1) 分析 備考

図 1　1 元配置分散分析と重回帰分析のデータ

1 元配置分散分析で利用するデータは「工場」と「不良品率」のデータで、分析結果を図 2 に示す。

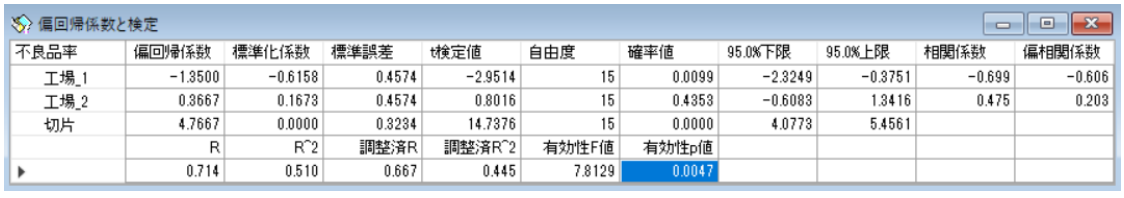

分散分析表

	平方和	自由度	不偏分散	F値
全変動	19.223	17	1.131	7.8129
水準間	9.808	2	4.904	P値
水準内	9.415	15	0.628	0.0047

図 2　1 元配置分散分析結果

ここで注意すべきは F 値と P 値である。

次にこの問題を重回帰分析を利用して解いてみよう。この場合、目的変数を「不良品率」説明変数を「工場」から作られたダミー(0/1)変数「工場_1，工場_2，工場_3」とするが、多重共線性を除くためにどれか 1 つの変数を省く。ここでは、「工場_3」を省いてみよう。結果を図 3 に示す。


偏回帰係数と検定

不良品率	偏回帰係数	標準化係数	標準誤差	t検定値	自由度	確率値	95.0%下限	95.0%上限	相関係数	偏相関係数
工場_1	-1.3500	-0.6158	0.4574	-2.9514	15	0.0099	-2.3249	-0.3751	-0.699	-0.606
工場_2	0.3667	0.1673	0.4574	0.8016	15	0.4353	-0.6083	1.3416	0.475	0.203
切片	4.7667	0.0000	0.3234	14.7376	15	0.0000	4.0773	5.4561		
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値				
	0.714	0.510	0.667	0.445	7.8129	0.0047				

図 3　「工場_1，工場_2」を利用した重回帰分析結果

ここで、「有効性 F 値」と「有効性 p 値」に注目すると、上の 1 元配置分散分析の結果に一

致する。有効性 p 値の値を与える検定は 2 つの変数の係数が同時に 0 になる結合仮説の検定と呼ばれる検定の結果である。

同様にもう 1 つの方法で重回帰分析を実行してみよう。「工場_1b，工場_2b」を説明変数に使う方法である。この変数は、工場_1、工場_2 の場合はそのまま、工場_3 の場合は、上で与えた制約に類似の、$\alpha_1+\alpha_2+\alpha_3=0$ という制約を利用して、$\alpha_3=-\alpha_1-\alpha_2$ の条件を付けたものである。すなわち、「工場_3」が選択された場合は、「工場_1」と「工場_2」のところに -1 を代入して、「工場_1b，工場_2b」としている。この説明変数を用いた場合の重回帰分析の結果を図 4 に示す。

偏回帰係数と検定

不良品率	偏回帰係数	標準化係数	標準誤差	t検定値	自由度	確率値	95.0%下限	95.0%上限	相関係数	偏相関係数
工場_1b	-1.0222	-0.8077	0.2641	-3.8708	15	0.0015	-1.5851	-0.4593	-0.533	-0.707
工場_2b	0.6944	0.5487	0.2641	2.6296	15	0.0189	0.1316	1.2573	0.145	0.562
切片	4.4389	0.0000	0.1867	23.7709	15	0.0000	4.0409	4.8369		
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値				
▶	0.714	0.510	0.667	0.445	7.8129	0.0047				

図 4　「工場_1b，工場_2b」を利用した重回帰分析結果

やはりこれも 1 元配置分散分析と同じ結果を与える。この重回帰分析の考え方を用いるのが Type3 と呼ばれる手法である。

2 元配置分散分析のデータを図 5 に示す。右の方には「条件 1」と「条件 2」をダミー(0/1)変数にしたものを加えている。

データ編集 分散分析（2元配置・繰り返し数不一致）.txt

	条件1	条件2	データ	条件1_1	条件1_2	条件1_3	条件2_1	条件2_2	条件2_3
▶ 1	2	1	72	0	1	0	1	0	0
2	1	3	73	1	0	0	0	0	1
3	3	2	76	0	0	1	0	1	0
4	1	2	76	1	0	0	0	1	0
5	1	3	91	1	0	0	0	0	1
6	2	2	67	0	1	0	0	1	0
7	3	1	64	0	0	1	1	0	0
8	2	2	100	0	1	0	0	1	0
9	1	2	69	1	0	0	0	1	0
10	2	1	61	0	1	0	1	0	0
11	2	2	69	0	1	0	0	1	0

1/1 (1,1)　分析　備考

図 5　2 元配置分散分析のデータ

2 元配置分散分析では交絡因子があることが 1 元配置と大きく異なる点である。交絡因子はどのようなデータとして与えたらよいだろうか。交絡因子は 2 つの変数の交わるところであるから、簡単な計算法として独立な成分「条件 1_1，条件 1_2」と「条件 2_1，条件 2_2」の積として与えてみよう。しかし、このままの形式では「条件 1_3」または「条件 2_3」が選ばれたデータは掛け算がすべて 0 になり、どちらの変数が 3 なのか区別がつかない。そこで利用されるのが図 1 で与えられた「工場_1b、工場_2b」の形式である。この変数形式だと「条件 1_3」または「条件 2_3」のどちらかが選ばれたとしても表 1 のように区別がつく。

表１　交絡変数の区別

		α1	α2	β1	β2	γ11	γ12	γ21	γ22
v1	v2	v1a	v1b	v2a	v2b	1a*2a	1a*2b	1b*2a	1b*2b
1	1	1	0	1	0	1	0	0	0
1	2	1	0	0	1	0	1	0	0
1	3	1	0	-1	-1	-1	-1	0	0
2	1	0	1	1	0	0	0	1	0
2	2	0	1	0	1	0	0	0	1
2	3	0	1	-1	-1	0	0	-1	-1
3	1	-1	-1	1	0	-1	0	-1	0
3	2	-1	-1	0	1	0	-1	0	-1
3	3	-1	-1	-1	-1	1	1	1	1

この場合は、２つの変数がそれぞれ 3 つのカテゴリに分かれているとしている。また、このデータで単独因子の部分は各変数 3-1=2 成分、交絡因子の部分は(3-1)(3-1)=4 成分になっている。

結局 2 元配置分散分析は、このデータ形式を利用して重回帰分析を実行し、単独因子と交絡因子についてそれぞれ結合仮説の検定を行い検定確率を求めることになる。これによってそれぞれの変動の和は一般に全変動には一致しないが、変数による推測の誤差は最も小さくなっている。

参考文献

[1] 田中豊他, パソコン統計解析ハンドブック<V 多変量分散分析・線形モデル編>, 共立出版, 1989.

２．重回帰分析

2.1　重回帰分析とは

重回帰分析は、説明変数が１つの回帰分析の拡張で、複数の説明変数の線形結合による１つの量的な目的変数の予測手法である。これは多変量解析の基礎的で代表的な分析手法で、これを応用して、非線形最小２乗法、局所重回帰分析など、様々な手法が考えられている。

重回帰分析では、実測値 y_λ を以下のような１次式と正規分布する誤差 ε_λ で与えられるものと考える。

$$y_\lambda = \sum_{i=1}^{p} b_i x_{i\lambda} + b_0 + \varepsilon_\lambda , \qquad \varepsilon_\lambda \sim N(0,\sigma^2) \text{ 分布［異なる } \lambda \text{ について独立］}$$

線形回帰式は偏回帰係数 b_i ， b_0 を用いて、以下の形で与えられる。

$$Y_\lambda = \sum_{i=1}^{p} b_i x_{i\lambda} + b_0$$

これらの偏回帰係数は実測値と予測値のずれの 2 乗和 EV が最小になるように決定される。

$$EV = \sum_{\lambda=1}^{n} (y_\lambda - Y_\lambda)^2 = \sum_{\lambda=1}^{n} \left(y_\lambda - \sum_{i=1}^{p} b_i x_{i\lambda} - b_0 \right)^2 \qquad \text{最小化}$$

この式の最小化はパラメータで微分した式を 0 とおいて、パラメータについての連立方程式を作り、その方程式を解くことで得られるが、その解は解析的に与えられる。

重回帰分析で利用される指標とその意味を以下にまとめておく。まず、標準化偏回帰係数は、すべての変数を標準化して重回帰分析を行ったときの係数であり、それぞれの変数の大きさや単位に依存しないその変数の真の重要性を表す。重相関係数は、目的変数と重回帰式で求められた予測値との相関係数で、重回帰式がどれだけ目的変数の予測をしているかを表す１つの指標である。重相関係数は説明変数の数が多くなるとそれに連れて大きくなっていくが、それを修正して説明変数の数に応じた値になるようにしたものが、自由度調整済み重相関係数である。寄与率は目的変数の変動（データと平均との差の 2 乗和）と重回帰式による変動の比で与えられ、目的変数の変動の何%が重回帰式で説明できるかを表している。また寄与率は決定係数とも呼ばれ、重相関係数の 2 乗としても計算可能である。

重回帰分析の検定では、式の有効性の検定と各偏回帰係数の検定とが別に扱われる。式の重要性の検定では重回帰式の変動が誤差変動に比べて十分大きいかどうかを検定する。偏回帰係数の検定では、各偏回帰係数の値が 0 と異なるかどうかの検定を行う。これらの有効性の検定や偏回帰係数の検定では残差の正規性が必要である。

以下に例を用いて、重回帰分析を説明する。

例

以下の表１のデータ（重回帰分析 1.txt）をもとに体重を身長と胸囲の１次関数で予測する。

表１　重回帰分析データ

体重	身長	胸囲	体重	身長	胸囲
61.0	167.0	84.0	49.5	164.7	78.0
55.5	167.5	87.0	61.0	171.0	90.0
57.0	168.4	86.0	59.5	162.6	88.0
57.0	172.0	85.0	58.4	164.8	87.0
50.0	155.3	82.0	53.5	163.3	82.0
50.0	151.4	87.0	54.0	167.6	84.0
66.5	163.0	92.0	60.0	169.2	86.0
65.0	174.0	94.0	58.8	168.0	83.0
60.5	168.0	88.0	54.0	167.4	85.2
49.5	160.4	84.9	56.0	172.0	82.0

この問題について上の指標を用いて答えた結果は以下となる。実際には書くことはないが、ここでは確率の値も入れておく。

目的変数を体重に、説明変数を身長と胸囲にして、重回帰分析を行ったところ、以下の重回帰式を得た。

体重 = 0.386*身長+0.858*胸囲-80.7

予測体重と実測体重の相関である重相関係数は 0.841 で、重回帰式の寄与率は 0.707 となった。これから体重変動の約 71%が説明できることが分かる。各変数の予測における重要性を示す標準化偏回帰係数は、身長が 0.433、胸囲が 0.640 と胸囲が少し上回っている。

回帰式の有効性の検定を行ったところ $p<0.001$（$p=0.0000$）となり、有効性が有意に示された。また、各偏回帰係数が 0 と異なることを示す検定では、身長が $p<0.01$（$p=0.0049$）、胸囲が $p<0.001$（$p=0.0002$）、切片は $p<0.01$（$p=0.0023$）となり、各係数とも有意に 0 と異なっている。

以上のことからこの回帰式は予測モデルとして、かなり良いモデルになっている。

重回帰分析では、偏回帰係数の検定によって不要と思われる説明変数を除き、寄与率をある程度保ち、よりコンパクトな式にモデルをまとめることがよく行われる。多くのソフトにはこのための変数自動選択の機能が付いている。変数選択の方法には、変数増加法、変数減少法、変数増減法が利用されるが、手動でこれを行うには、変数減少法が使いやすい。またソフトで自動的に行う場合は変数増減法がよく使われる。これらはいずれも偏回帰係数の検定確率（または検定値）を利用して、変数の選別を行っている。

2.2　プログラムの利用法

メニュー［分析－多変量解析他－予測手法－重回帰分析］を選択すると表示される分析画面を図 1、データを図 2 に示す。この場合は変数選択で全てのデータを選択する。

図１　重回帰分析実行画面

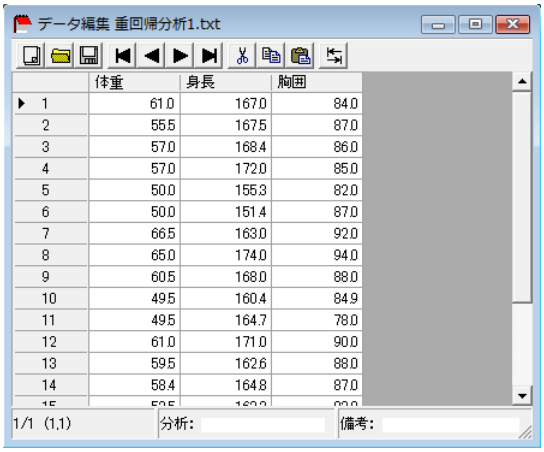


図２　重回帰分析データ

「相関行列」ボタンでは目的変数と説明変数を含んだ相関行列**R**が表示される。その際、相関係数を 0 と比較する検定の確率値も表示される。「重回帰分析」ボタンでは、テキスト画面とグリッド画面の２つのウィンドウが開き、図 3a と図 3b の分析結果が表示される。

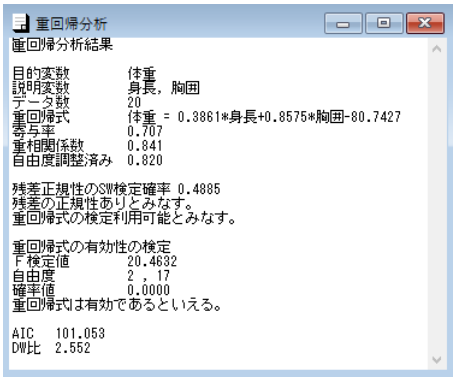


図 3a　重回帰分析出力結果１

図 3b　重回帰分析出力結果２

これらは同じ内容であるが、テキスト出力は初心者を意識した出力である。

次に、「分散分析表」ボタンをクリックすると、図 4 に示す結果が表示される。

図 4　分散分析表画面

「予測値と残差」ボタンでは、図 5 のように各レコード毎の実測値、予測値、残差、標準

化残差、てこ比が表示される。

予測値と残差

	実測値	予測値	残差	標準化残差	てこ比<0.375
1	61.0	55.762	5.238	1.798	0.067
2	55.5	58.528	-3.028	-1.040	0.059
3	57.0	58.018	-1.018	-0.350	0.061
4	57.0	58.550	-1.550	-0.532	0.124
5	50.0	49.530	0.470	0.161	0.265
6	50.0	52.312	-2.312	-0.794	0.450
7	66.5	61.078	5.422	1.862	0.241
8	65.0	67.040	-2.040	-0.701	0.372
9	60.5	59.579	0.921	0.316	0.073
10	49.5	53.986	-4.486	-1.540	0.102

図５　予測値と残差

また、「実測／予測値の散布図」ボタンでは、図６のように実測値と予測値の散布図が描かれる。図のラベルはグラフメニューの［設定ーデータラベル］で表示を ON にしている。

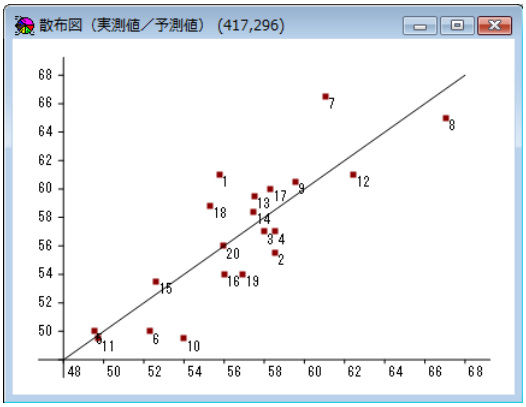


図６　実測値と予測値の散布図

次に変数の自動選択について、図７のデータを用いて説明する。

データ編集 重回帰分析2.txt

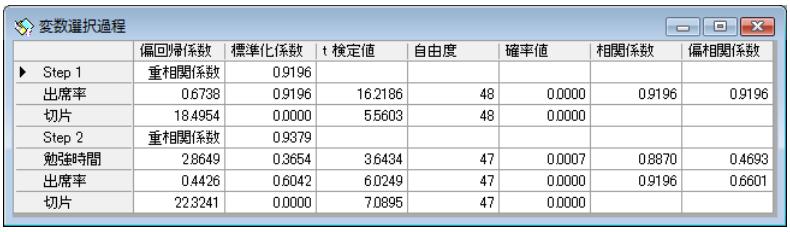
	卒業試験	入試点数	内申点数	勉強時間	出席率
	83	67	3.0	7.4	100
	90	71	3.7	8.0	100
	80	57	3.9	6.5	78
	39	43	2.8	1.8	38
	81	63	3.6	6.1	100
	47	51	3.7	2.7	55
	92	72	4.1	7.9	100
	75	62	3.8	4.6	90

1/1 (1,1)　分析:　備考:

図７　変数自動選択のデータ

最初に全ての変数を選択して分析を実行する。変数の追加と削除の基準は、追加と削除の変数の係数についての検定確率または F 検定値のどちらかで与えられる。「Pin」左側のラジオボタンをチェックすると検定確率で指定し「Fin」左側のラジオボタンをチェックすると F 検定値で指定することになる。デフォルトは検定確率になっている。

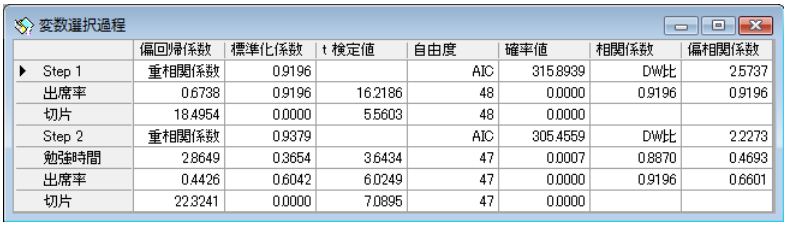
変数の選択法として、変数増加法、変数減少法、変数増減法のどれかを選び、「選択」ボタンをクリックすると図８のように選択過程での種々の統計量が表示される。



	偏回帰係数	標準化係数	t 検定値	自由度	確率値	相関係数	偏相関係数
Step 1	重相関係数	0.9196					
出席率	0.6738	0.9196	16.2186	48	0.0000	0.9196	0.9196
切片	18.4954	0.0000	5.5603	48	0.0000		
Step 2	重相関係数	0.9379					
勉強時間	2.8649	0.3654	3.6434	47	0.0007	0.8870	0.4693
出席率	0.4426	0.6042	6.0249	47	0.0000	0.9196	0.6601
切片	22.3241	0.0000	7.0895	47	0.0000		

図 8　変数選択過程表示画面

この場合は、2 段階で変数が 2 つ選択されている。図 1 で「AIC」チェックボックスや「DW 比」チェックボックスにチェックを入れると、各過程での AIC の値やダービン・ワトソン比が図 8 の画面上に図 9 のように追加して表示される。



	偏回帰係数	標準化係数	t 検定値	自由度	確率値	相関係数	偏相関係数
Step 1	重相関係数	0.9196		AIC	315.8939	DW比	2.5737
出席率	0.6738	0.9196	16.2186	48	0.0000	0.9196	0.9196
切片	18.4954	0.0000	5.5603	48	0.0000		
Step 2	重相関係数	0.9379		AIC	305.4559	DW比	2.2273
勉強時間	2.8649	0.3654	3.6434	47	0.0007	0.8870	0.4693
出席率	0.4426	0.6042	6.0249	47	0.0000	0.9196	0.6601
切片	22.3241	0.0000	7.0895	47	0.0000		

図 9　AIC と DW 比を加えた変数選択過程表示画面

ここに AIC の値はモデルの良さを与える指標で、小さな値になるほど良いと判断される。またダービン・ワトソン比は重回帰式の残差がレコードに独立かどうかを調べる指標で、2 に近い値が出れば良いとされる。特に、相関が 0.5 で 1、-0.5 で 3 に近い値となるので、このような値だと注意を要する。

重回帰分析は 1 つの目的変数を複数の説明変数の線形結合で予測するモデルであるが、データによっては、1 つの線形結合として表すのではなく、複数の線形結合の混ざり合ったものとして表す方が良い予測結果を与える場合がある。ここではこの機能について図 10 の例を用いて説明する。変数選択では、最初に群分け用変数、次に目的変数、続けて説明変数を選択する。データの形式は、図 1 の分析メニューで、「先頭列で群分け」ラジオボタンを選択する。

	群	体重	身長	胸囲
17	1	60.0	169.2	86.0
18	1	58.8	168.0	83.0
19	1	54.0	167.4	85.2
20	1	56.0	172.0	82.0
21	2	63.3	167.0	84.0
22	2	67.5	167.5	87.0
23	2	68.3	168.4	86.0
24	2	67.2	172.0	85.0

図 10　群分けした重回帰分析のデータ

「相関行列」ボタンをクリックすると、図 11 のように、「群」変数で群分けした変数間の相関行列が表示される。

相関係数

	体重	身長	胸囲
▶ 群 1			
体重	1.000	0.559	0.725
身長	0.559	1.000	0.197
胸囲	0.725	0.197	1.000
群 2			
体重	1.000	0.667	0.676
身長	0.667	1.000	0.197
胸囲	0.676	0.197	1.000

図 11　群分けした相関行列

また、「重回帰分析」ボタンをクリックすると、図 12a と図 12b のような群分けした結果が表示される。

重回帰分析

```
局所重回帰分析結果

群 1
目的変数          体重
説明変数          身長, 胸囲
データ数          20
重回帰式          体重 = 0.386*身長+0.858*胸囲-80.743
寄与率            0.707
重相関係数        0.841
自由度調整済み    0.820

残差正規性のSW検定確率 0.488
残差の正規性ありとみなす。
重回帰式の検定利用可能とみなす。

重回帰式の有効性の検定
F検定値           20.463
自由度            2 , 17
確率値            0.000
重回帰式は有効であるといえる。

AIC    101.05
DW比   2.552


群 2
目的変数          身長
説明変数          身長, 胸囲
データ数          20
重回帰式          体重 = 0.334*身長+0.511*胸囲-33.542
```

図 12a　群分けした重回帰分析結果１

偏回帰係数と検定

	偏回帰係数	標準化係数	標準誤差	t 検定値	自由度	確率値	95.0%下限	95.0%上限	相関係数	偏相関係数
▶ 群 1										
身長	0.3861	0.4333	0.1194	3.2335	17	0.0049	0.1342	0.6380	0.386	0.617
胸囲	0.8575	0.6401	0.1795	4.7768	17	0.0002	0.4788	1.2363	0.858	0.757
切片	-80.7427	0.0000	22.5785	-3.5761	17	0.0023	-128.3791	-33.1063		
群 2										
身長	0.3335	0.5560	0.0736	4.5292	17	0.0003	0.1782	0.4889	0.334	0.739
胸囲	0.5109	0.5664	0.1107	4.6141	17	0.0002	0.2773	0.7445	0.511	0.746
切片	-33.5423	0.0000	13.9250	-2.4088	17	0.0276	-62.9216	-4.1630		

図 12b　群分けした重回帰分析結果２

ここで、図 12a の画面下方には、群分けした結果の他に、図 12c のような、全体的な指標も表示される。

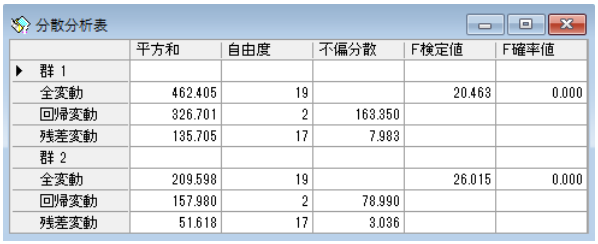
重回帰分析

```
全実測・予測相関係数              0.933
全実測・予測相関係数２乗          0.870
Σ群内回帰変動／Σ群内全変動  0.721
```

図 12c　群分けした重回帰分析結果３

これは、群分けした結果から、予測値を求め、それを元にして全体的な予測の程度を与えたものである。重回帰分析では、実測値と予測値の相関係数（重相関係数）の 2 乗と回帰変動／全変動（寄与率）の結果が一致するが、ここの定義だと異なっている。

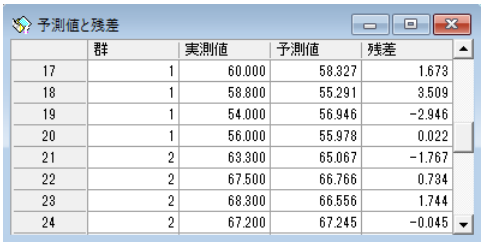
「分散分析表」ボタンをクリックすると、図 13 のように、群別に計算された分散分析表が表示される。


分散分析表

	平方和	自由度	不偏分散	F検定値	F確率値
▶ 群 1					
全変動	462.405	19		20.463	0.000
回帰変動	326.701	2	163.350		
残差変動	135.705	17	7.983		
群 2					
全変動	209.598	19		26.015	0.000
回帰変動	157.980	2	78.990		
残差変動	51.618	17	3.036		

図 13　群分けされた分散分析表

「予測値と残差」ボタンをクリックすると、レコード順に、群別に計算された予測値と残差を図 14 のように表示する。


予測値と残差

	群	実測値	予測値	残差
17	1	60.000	58.327	1.673
18	1	58.800	55.291	3.509
19	1	54.000	56.946	-2.946
20	1	56.000	55.978	0.022
21	2	63.300	65.067	-1.767
22	2	67.500	66.766	0.734
23	2	68.300	66.556	1.744
24	2	67.200	67.245	-0.045

図 14　群分けされた予測値と残差結果

「実測／予測散布図」ボタンをクリックすると、図 15 のように図 14 の実測値と予測値を用いたグラフが表示されるが、このグラフの回帰直線は実測値＝予測値（y=x）の直線を示しており、（当然ながら）重なって表示されている。

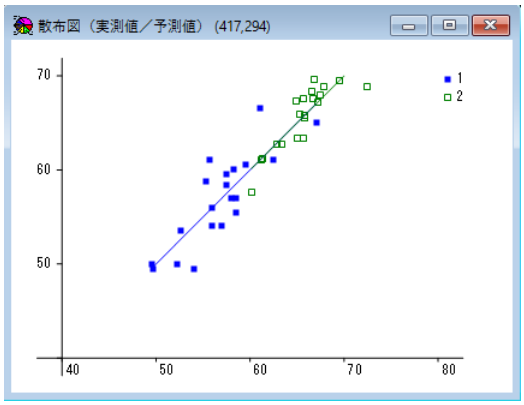


図 15　群分けされた実測値／予測値散布図

予測について

機械学習などでは、訓練データに対して独立なテストデータを用いて目的変数の予測を行う。図 1 の分析実行画面の一番下に「新データ予測」ボタンがあるが、これは一度重回帰分析を実行した後、新たなデータを選択して予測を行うためのボタンである。テストデータに目的変数を含む場合は「目的変数付き」チェックボックスにチェックを入れて、「新データ予測」ボタンをクリックする。テストデータで予測値を計算し、適合性を R^2 の値で示してくれる。これは「群別データ」のとき有効である。

計量経済分析によく利用される機能

計量経済分析には回帰分析が多用される。その目的は、因果関係の導出やデータの予測などであり、標準的な回帰分析から派生した様々な手法、例えばパネルデータ分析や操作変数回帰分析なども広く利用されている。さらに、経済分野では標準的な回帰分析でも他分野と多少異なった使われ方をする場合があり、興味深い。特に経済の分析では数多くのデータを扱うので、中心極限定理が成り立つと仮定する場合が多く、基本的に正規分布を基礎とした検定が行われる。

さて、このような計量経済で利用される様々な回帰分析に類似する手法は他の章で述べるとして、ここでは重回帰分析の中で気をつけるべき事柄を解説する。1 つは重回帰分析の誤差項の不均一性の問題である。基本的な重回帰分析では、回帰の誤差は説明変数の値によらず一定と考える。しかし、経済分野では基本的に回帰の誤差は説明変数に依存すると考えることが多い。そのため誤差項の扱いや検定に利用する分布が異なってくる。我々のプログラムでは図 16 の赤枠の「不均一分散回帰分析」チェックボックスにチェックを入れることで、その処理が行える。

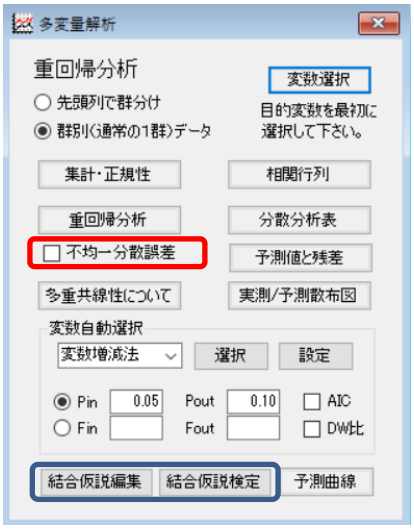
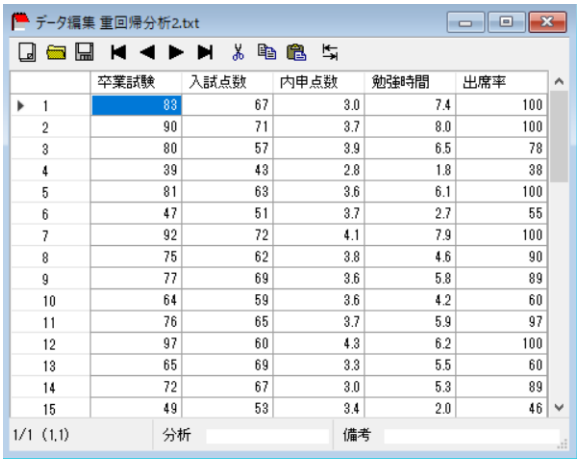


図 16　重回帰分析実行画面（再掲）



	卒業試験	入試点数	内申点数	勉強時間	出席率
1	83	67	3.0	7.4	100
2	90	71	3.7	8.0	100
3	80	57	3.9	6.5	78
4	39	43	2.8	1.8	38
5	81	63	3.6	6.1	100
6	47	51	3.7	2.7	55
7	92	72	4.1	7.9	100
8	75	62	3.8	4.6	90
9	77	69	3.6	5.8	89
10	64	59	3.6	4.2	60
11	76	65	3.7	5.9	97
12	97	60	4.3	6.2	100
13	65	69	3.3	5.5	60
14	72	67	3.0	5.3	89
15	49	53	3.4	2.0	46

図 17　サンプルデータ（重回帰分析 2.txt）

図 17 のデータで処理した結果を図 18 と図 19 に示す。

卒業試験	偏回帰係数	標準化係数	標準誤差	z統計量	確率値	95.0%下限	95.0%上限	相関係数	偏相関係数
入試点数	0.1490	0.0749	0.1092	1.3647	0.1724	-0.0650	0.3629	0.386	0.203
内申点数	2.2233	0.0527	2.7280	0.8150	0.4151	-3.1235	7.5702	0.121	0.152
勉強時間	2.7614	0.3522	0.7032	3.9271	0.0001	1.3832	4.1396	0.887	0.469
出席率	0.4314	0.5888	0.0702	6.1439	0.0000	0.2938	0.5690	0.920	0.664
切片	6.8654	0.0000	8.1304	0.8444	0.3984	-9.0699	22.8007		
	R	R^2	調整済R	調整済R^2	有効性F値	有効性p値			
	0.943	0.889	0.938	0.879	158.4717	0.0000			

図 18　不均一分散の計算結果 1

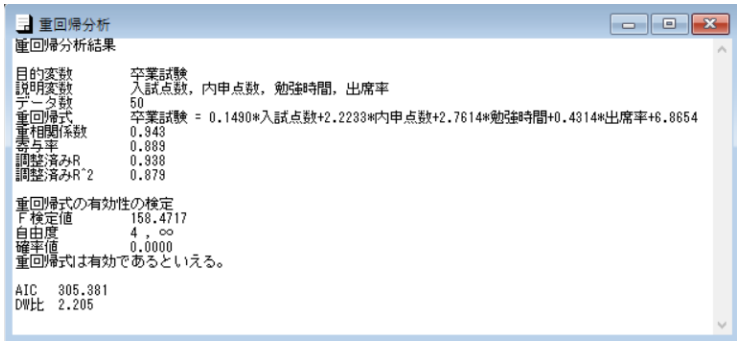


図 19　不均一分散の有効性の検定

ここではまず図 18 で、偏回帰係数の標準誤差を与え、そこから z 統計量を通じて区間推定を行っている。また、図 19 の有効性の検定は $F_{1,\infty}$の検定を行っており、通常の F 検定とは異なる。

さらに、経済分野での重回帰分析では、結合仮説の検定が使われる。この簡単な例はすべての説明変数の係数が 0 であるという帰無仮説を用いる有効性の検定であるが、特定の係数の線形結合がある値を取るという検定も使われることがある。そのため我々は、図 16 の青枠で囲まれた部分に結合仮説検定の機能を追加した。

「結合仮説編集」のボタンをクリックすると選択された変数を用いて図 20 のような結合仮説編集画面が表示される。

図 20　結合仮説編集画面

この画面を表示したまま隣の結合仮説検定ボタンをクリックすると、図 21 のような検定結果が表示される。

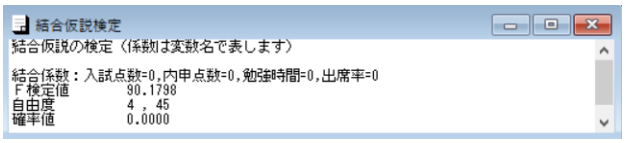


図 21　結合仮説検定結果

この結果は、回帰式の有効性の検定結果に等しい。

次に、あまり現実的ではないが、例えば「入試点数+3*内申点数=1」などという複雑な検定を行ってみよう。結合仮説編集画面で図 22 のように入力する。

図 22　結合仮説編集画面

結果は図 23 のようになる。

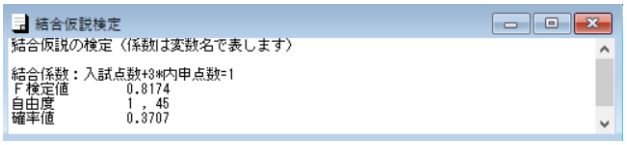


図 23　結合仮説検定結果

ここで「入試点数+3*内心点数=1」になっているのが分かる。ここでは、均一分散の場合を扱ったが、不均一分散の場合でも同様のことを実行することができる。しかし、先頭列で群分けの場合については対応していない。また現在、結合仮説の検定の基本設定の部分までは 2 値ロジスティック分析にも含まれている。これらの拡張については今後検討する。

問題１（重回帰分析 2.txt）

重回帰分析 2.txt について、重回帰分析を行い、以下の問いに答えよ。

１）回帰式を求めよ。

卒業試験＝　［　　］入試点数　＋［　　］内申点数
＋［　　］勉強時間　＋［　　］出席率
＋［　　］

２）この回帰式の寄与率を求めよ。［　　］

３）この場合残差の分布は正規分布といえるか。［正規分布・正規分布でない］

変数増減法を用いて変数を自動選択する。

４）最終的な回帰式はどのようになるか。不要な変数の係数欄は空欄のままでよい。

卒業試験＝　［　　］入試点数　＋［　　］内申点数
＋［　　］勉強時間　＋［　　］出席率
＋［　　］

５）上の回帰式の寄与率を求めよ。［　　］

６）上の回帰式の寄与率はすべての変数を使った場合に比べ大きく下がっているか。
［大きく下がっている・あまり下がっていない］

７）この式を新しい予測モデルとして採用するか。　［採用する・採用しない］

８）新しい予測モデルで、データ中の最初（1 番）の学生について卒業試験の実測値，その予測値，残差（実測値と予測値の差）はいくらか。

実測値［　　］　予測値［　　］　残差［　　］

９）上と同様のモデルで、質問項目の値が入試点数 70、内申点数 3.5、勉強時間 5、出席率 70%の学生の卒業試験はいくらに予測されるか。　［　　］

演習１（多変量演習 3.txt）

１）多変量演習 3.txt で、全変数を使った以下の重回帰式はどのように与えられるか。

試験成績＝［　　］×評定平均＋［　　］×模試１

［　　］×模試２＋［　　］×模試３＋［　　］

２）この重回帰式の寄与率はいくらか。［　　］

変数自動選択で偏回帰係数が有効である回帰モデルを作り、以下の問いに答えよ。

３）重回帰式はどのようになるか。説明変数に含まれないものは空欄のままにすること。

試験成績＝［　　］×評定平均＋［　　］×模試１

［　　］×模試２＋［　　］×模試３＋［　　］

４）寄与率はいくらになったか。［　　］

５）上の重回帰式を新しい予測モデルにして良いと思うか。［思う・思わない］

以後、新しいモデルで答えること。

６）データの中で最初の学生の予測試験成績はいくらか。［　　］

７）新しい重回帰式を利用すると以下の点数の学生の試験成績は何点に予測されるか。

変数名	評定平均	模試１	模試２	模試３
成績	3.5	70	73	75

予測試験成績［　　］

演習２（多変量演習 4.txt）

多変量演習 4.txt のデータは各質問項目について 5 段階評価で、講義ごとに平均を取ったものである。多変量解析の重回帰分析を用いて以下の問いに答えよ。

総合評価を調査数以外のすべての変数で予測する重回帰モデル

１）重回帰式を求めよ。

総合評価＝　［　　］進む速さ　＋［　　］声の大きさ

＋［　　］黒板等　　＋［　　］私語注意

＋［　　］分かり易さ＋［　　］有益さ

＋［　　］受講態度　＋［　　］

２）この回帰式の寄与率を求めよ。［　　］

変数自動選択で変数増減法を用いて、すべての偏回帰係数が有効である回帰モデルを作り、以下の問いに答えよ。

３）最終的な回帰式はどのようになるか。不要な変数の係数欄は空欄のままでよい。

総合評価＝　［　　］進む速さ　＋［　　］声の大きさ

＋［　　］黒板等　　＋［　　］私語注意

＋［　　］分かり易さ＋［　　］有益さ

＋［　　］受講態度　＋［　　］

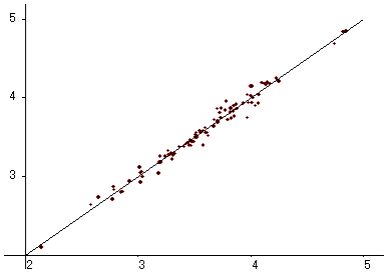
４）上の回帰式の寄与率を求めよ。［　　］

５）上の回帰式の寄与率はすべての変数を使った場合に比べ大きく下がっているか。

［大きく下がっている・あまり下がっていない］

６）この式を新しい予測モデルとして採用するか。
［採用する・採用しない］

７）予測値がどの程度実測値に近いかを見るために、
右のような散布図を描け。



８）総合評価に影響を与える重要な説明変数を２つ挙げよ。
［　　］［　　］

９）データ中の最初（1 番）の授業について、総合評価の実測値，その予測値，残差（実測値と予測値の差）はいくらか。
実測値［　　］　予測値［　　］　残差［　　］

10）すべての質問項目の値が 3.5 の授業の総合評価はいくらに予測されるか。
［　　］

問題１解答（重回帰分析 2.txt）

１）回帰式を求めよ。
卒業試験＝　［ 0.1490 ］入試点数　＋［ 2.2233 ］内申点数
＋［ 2.7614 ］勉強時間　＋［ 0.4314 ］出席率　＋［ 6.8654 ］

２）この回帰式の寄与率を求めよ。［ 0.889 ］

３）この場合残差の分布は正規分布といえるか。［正規分布・正規分布でない］

４）最終的な回帰式はどのようになるか。不要な変数の係数欄は空欄のままでよい。
卒業試験＝　［　　］入試点数　＋［　　］内申点数
＋［ 2.8649 ］勉強時間　＋［ 0.4426 ］出席率　＋［ 22.3241 ］

５）上の回帰式の寄与率を求めよ。［ 0.880 ］

６）上の回帰式の寄与率はすべての変数を使った場合に比べ大きく下がっているか。
［大きく下がっている・あまり下がっていない］

７）この式を新しい予測モデルとして採用するか。　［採用する・採用しない］

８）新しい予測モデルで、データ中の最初（1 番）の学生について卒業試験の実測値，その予測値，残差（実測値と予測値の差）はいくらか。
実測値［ 83 ］　予測値［ 87.789 ］　残差［ -4.789 ］

９）上と同様のモデルで、質問項目の値が入試点数 70、内申点数 3.5、勉強時間 5、出席率 70%の学生の卒業試験はいくらに予測されるか。　［ 67.6306 ］

演習１解答（多変量演習 3.txt）

１）多変量演習 3.txt で、全変数を使った以下の重回帰式はどのように与えられるか。
試験成績＝［ 9.0898 ］×評定平均＋［ 0.1816 ］×模試１
［ 0.0701 ］×模試２＋［ 0.3397 ］×模試３＋［ 0.1150 ］

２）この重回帰式の寄与率はいくらか。［ 0.906 ］

３）重回帰式はどのようになるか。説明変数に含まれないものは空欄のままにすること。
試験成績＝［ 13.6032 ］×評定平均＋［　　］×模試１
［　　］×模試２＋［ 0.3573 ］×模試３＋［ -0.6163 ］

４）寄与率はいくらになったか。［ 0.903 ］

５）上の重回帰式を新しい予測モデルにして良いと思うか。［思う・思わない］

６）データの中で最初の学生の予測試験成績はいくらか。［ 74.796 ］

７）新しい重回帰式を利用すると以下の点数の学生の試験成績は何点に予測されるか。

変数名	評定平均	模試１	模試２	模試３
成績	3.5	70	73	75

予測試験成績［ 73.7924 ］

演習２解答（多変量演習 4.txt）

１）重回帰式を求めよ。

総合評価＝　［ 0.2930 ］進む速さ　＋［ 0.1205 ］声の大きさ
＋［ 0.1142 ］黒板等　　＋［ 0.0593 ］私語注意
＋［ 0.1432 ］分かり易さ＋［ 0.3659 ］有益さ
＋［ −0.0653 ］受講態度　＋［ −0.0374 ］

２）この回帰式の寄与率を求めよ。［ 0.980 ］

３）最終的な回帰式はどのようになるか。不要な変数の係数欄は空欄のままでよい。

総合評価＝　［ 0.3054 ］進む速さ　＋［ 0.1326 ］声の大きさ
＋［ 0.0840 ］黒板等　　＋［　　］私語注意
＋［　　］分かり易さ＋［ 0.4026 ］有益さ
＋［　　］受講態度　＋［ −0.0790 ］

４）上の回帰式の寄与率を求めよ。［ 0.979 ］

５）上の回帰式の寄与率はすべての変数を使った場合に比べ大きく下がっているか。
［大きく下がっている・あまり下がっていない］

６）この式を新しい予測モデルとして採用するか。　［採用する・採用しない］

８）総合評価に影響を与える重要な説明変数を２つ挙げよ。
［ 有益さ ］［ 進む速さ ］

９）データ中の最初（1 番）の授業について、総合評価の実測値，その予測値，残差（実測値と予測値の差）はいくらか。
実測値［ 2.81 ］　予測値［ 2.860 ］　残差［ −0.050 ］

10）すべての質問項目の値が 3.5 の授業の総合評価はいくらに予測されるか。［ 3.57 ］

2.3　重回帰分析の理論

重回帰分析は、目的変数を複数の説明変数の線形回帰式で予測する手法である。データは以下の表１の形式で与えられる。

表１ 重回帰分析のデータ

目的変数	説明変数 1	…	説明変数 p
y_1	x_{11}	…	x_{p1}
y_2	x_{12}	…	x_{p2}
:	:		:
y_n	x_{1n}	…	x_{pn}

実測値は以下のような１次式と正規分布する誤差 ε_λ で与えられるものと考える。

$$y_\lambda = \sum_{i=1}^{p} b_i x_{i\lambda} + b_0 + \varepsilon_\lambda , \qquad \varepsilon_\lambda \sim N(0,\sigma^2) \text{分布［異なる } \lambda \text{ について独立］}$$

線形回帰式は偏回帰係数 b_i，b_0 を用いて、以下の形で与えられる。

$$Y_\lambda = \sum_{i=1}^{p} b_i x_{i\lambda} + b_0$$

これらの偏回帰係数は実測値と予測値のずれの 2 乗和EVが最小になるように決定される。

$$EV=\sum_{\lambda=1}^{n}(y_\lambda-Y_\lambda)^2 \qquad \text{最小化}$$

即ち、b_iとb_0についてのEVの微係数を 0 とおいて以下の式を得る。

$$b_i=(\mathbf{S}^{-1}\mathbf{S}_y)_i,\quad b_0=\bar{y}-\sum_{i=1}^{p}b_i\bar{x}_i$$

ここに、$\mathbf{S}^{-1}$は説明変数の共分散行列$\mathbf{S}$の逆行列、$\mathbf{S}_y$は目的変数と説明変数の分散共分散ベクトルである。

$$(\mathbf{S})_{ij}=\frac{1}{n-1}\sum_{\lambda=1}^{n}(x_{i\lambda}-\bar{x}_i)(x_{j\lambda}-\bar{x}_j),\quad (\mathbf{S}_y)_i=\frac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda-\bar{y})(x_{i\lambda}-\bar{x}_i)$$

偏回帰係数は変数の平均や分散によって影響を受け、係数の重要性が分かりにくいが、データを以下のように標準化して重回帰分析を行なうと変数の影響力の強さがはっきりと示される。ここに s_y^2， s_i^2は目的変数及び説明変数 iの不偏分散である。

$$\tilde{y}_\lambda=\frac{y_\lambda-\bar{y}}{s_y},\quad \tilde{x}_{i\lambda}=\frac{x_{i\lambda}-\bar{x}_i}{s_i}$$

これらの新しいデータ$\tilde{y}_\lambda$と$\tilde{x}_{i\lambda}$で作った重回帰式の偏回帰係数$\tilde{b}_i$を標準化偏回帰係数と言い、回帰式は以下のように表わされる。

$$\tilde{Y}_\lambda=\sum_{i=1}^{p}\tilde{b}_i\tilde{x}_{i\lambda}$$

標準化偏回帰係数と偏回帰係数との関係は$\tilde{b}_i=b_i\,s_i/s_y$で与えられる。

重相関係数 Rは実測値と予測値の相関係数であり、以下のように与えられる。

$$R=s_{yY}/(s_y s_Y)$$

ここに、s_{yY}は実測値 yと予測値 Yの共分散、s_y^2とs_Y^2は実測値と予測値の不偏分散である。

$$s_{yY}=\frac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda-\bar{y})(Y_\lambda-\bar{Y}),\quad s_y^2=\frac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda-\bar{y})^2,\quad s_Y^2=\frac{1}{n-1}\sum_{\lambda=1}^{n}(Y_\lambda-\bar{Y})^2$$

実測値の全変動 SVは回帰変動 RVと残差変動 EVの和として表わされる。

$$SV=\sum_{\lambda=1}^{n}(y_\lambda-\bar{y})^2=\sum_{\lambda=1}^{n}(y_\lambda-Y_\lambda)^2+\sum_{\lambda=1}^{n}(Y_\lambda-\bar{Y})^2=EV+RV$$

全変動に占める回帰変動の割合は、予測値が実測値を説明する割合を表わしていると考えられ、その値を寄与率という。寄与率は重相関係数の 2 乗に等しいことが示されるので、記号R^2で表わすことにする。

$$R^2=RV/SV$$

寄与率や重相関係数の値は説明変数の数が増えれば大きくなることが知られており、これを緩和するために以下のような自由度調整済み重相関係数$\bar{R}$が考えられている。

$$\overline{R} = \sqrt{1 - \frac{EV/(n-p-1)}{SV/(n-1)}}$$

重回帰式の有効性は回帰変動と残差変動を比べて、回帰変動が十分大きいことが重要で、この検定には、以下の性質が利用される。

$$F = \frac{RV/p}{EV/(n-p-1)} \sim F_{p,n-p-1} \text{分布}$$

重回帰式全体の有効性とは別に、それぞれの偏回帰係数の有効性も検討される。これらは偏回帰係数が 0 と異なることを示して確かめられる。この検定には以下の性質が利用される。

$b_i = 0$ の検定　　$t_i = \dfrac{b_i}{\sqrt{a^{ii} EV/(n-p-1)}} \sim t_{n-p-1}$ 分布

$b_0 = 0$ の検定　　$t_0 = \dfrac{b_0}{\sqrt{\left(\dfrac{1}{n} + \sum_{i=1}^{p}\sum_{j=1}^{p} \overline{x}_i \overline{x}_j a^{ij}\right) EV \Big/ (n-p-1)}} \sim t_{n-p-1}$ 分布

ここに a^{ij} は $\mathbf{A} = (n-1)\mathbf{S}$ としたときの行列 $\mathbf{A}$ の逆行列 $\mathbf{A}^{-1}$ の i, j 成分である。

説明変数 i を除く他の説明変数で作った $x_{i\lambda}$ の予測回帰式を以下のように書く。

$$X_{i\lambda} = b_1^{(i)} x_{1\lambda} + \cdots + b_{i-1}^{(i)} x_{i-1\lambda} + b_{i+1}^{(i)} x_{i+1\lambda} + \cdots + b_p^{(i)} x_{p\lambda} + b_0^{(i)}$$

また、説明変数 i を除く他の説明変数で作った目的変数の予測回帰式を以下のように書く。

$$Y_{i\lambda} = b_1'^{(i)} x_{1\lambda} + \cdots + b_{i-1}'^{(i)} x_{i-1\lambda} + b_{i+1}'^{(i)} x_{i+1\lambda} + \cdots + b_p'^{(i)} x_{p\lambda} + b_0'^{(i)}$$

実測値からこれらの予測値を引いた値をそれぞれ $x'_{i\lambda}$，$y'_{i\lambda}$ として、

$$x'_{i\lambda} = x_{i\lambda} - X_{i\lambda}, \quad y'_{i\lambda} = y_\lambda - Y_{i\lambda},$$

この $x'_{i\lambda}$ と $y'_{i\lambda}$ の相関係数を偏相関係数と呼び、$\tilde{r}_{iy}$ で表わす。偏相関係数は他の変数の影響を除いた相関係数と見ることができ、以下のように表わすこともできる。

$$\tilde{r}_{iy} = -r^{iy} \Big/ \sqrt{r^{ii} r^{yy}}$$

ここに r^{iy}，r^{ii}，r^{yy} は、目的変数と説明変数を合せた相関行列 $\mathbf{R}$ の逆行列 $\mathbf{R}^{-1}$ の成分である。

$$\mathbf{R} = \begin{pmatrix} 1 & r_{y1} & \cdots & r_{yp} \\ r_{1y} & 1 & \cdots & r_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{py} & r_{p1} & \cdots & 1 \end{pmatrix}, \quad \mathbf{R}^{-1} = \begin{pmatrix} r^{yy} & r^{y1} & \cdots & r^{yp} \\ r^{1y} & r^{11} & \cdots & r^{1p} \\ \vdots & \vdots & \ddots & \vdots \\ r^{py} & r^{p1} & \cdots & r^{pp} \end{pmatrix}$$

また、モデルの適合度を表すのに、AIC の値が利用されることがあるが、これは以下のように定義される（以下の $1/n$ で定義している本もある）。

$$AIC = n\left(\log(2\pi) + 1\right) + n\log\left(EV/n\right) + 2p$$

最後にダービン・ワトソン比は、重回帰式の残差がレコード毎に独立になっているかどうかを調べる指標で、残差をε_λとして、以下のように与えられる。

$$DW=\sum_{\lambda=2}^{N}(\varepsilon_\lambda-\varepsilon_{\lambda-1})^2\Big/\sum_{\lambda=1}^{N}\varepsilon_\lambda^2$$

ε_λと$\varepsilon_{\lambda-1}$が独立になっていれば、2に近く、相関の値が0.5で1、-0.5で3に近い値となる。

2.4 重回帰分析の行列による解説【補足】

一般的な表式を考える前に、

$$\mathbf{y}=\mathbf{X}\mathbf{b}+b_0\mathbf{1}+\mathbf{u}$$

$$\frac{\partial L}{\partial b_0}=-2\mathbf{1}'(\mathbf{y}-\mathbf{X}\mathbf{b}-b_0\mathbf{1})=-2N(\bar{y}-\bar{\mathbf{x}}'\mathbf{b}-b_0)=0,\quad \text{ここに}\bar{\mathbf{x}}=\begin{pmatrix}\bar{x}_1\\ \vdots\\ \bar{x}_p\end{pmatrix}$$

$$\frac{\partial L}{\partial \mathbf{b}}=-2\mathbf{X}'(\mathbf{y}-\mathbf{X}\mathbf{b}-b_0\mathbf{1})=-2\mathbf{X}'[\mathbf{y}-\mathbf{X}\mathbf{b}-(\bar{y}-\bar{\mathbf{x}}'\mathbf{b})\mathbf{1}]$$

$$=-2\mathbf{X}'[(\mathbf{y}-\bar{\mathbf{y}})-(\mathbf{X}-\bar{\mathbf{X}})\mathbf{b}]=-2(\mathbf{X}-\bar{\mathbf{X}})'[(\mathbf{y}-\bar{\mathbf{y}})-(\mathbf{X}-\bar{\mathbf{X}})\mathbf{b}]=\mathbf{0}$$

ここに、$\bar{\mathbf{y}}=\begin{pmatrix}\bar{y}\\ \vdots\\ \bar{y}\end{pmatrix}$, $\bar{\mathbf{X}}=\begin{pmatrix}\bar{x}_1 & \cdots & \bar{x}_p\\ \vdots & \ddots & \vdots\\ \bar{x}_1 & \cdots & \bar{x}_p\end{pmatrix}$, $\mathbf{Z}=\mathbf{X}-\bar{\mathbf{X}}$, $\mathbf{v}=\mathbf{y}-\bar{\mathbf{y}}$

これより、$\mathbf{b}$について考えると（b_0については一般論のところで）

$$\hat{\mathbf{b}}=\mathbf{C}^{-1}(\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{y}-\bar{\mathbf{y}}),\quad \mathbf{C}=(\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{X}-\bar{\mathbf{X}})$$

次に、$\hat{\mathbf{b}}$の分布を考える。

$$\hat{\mathbf{b}}=\mathbf{C}^{-1}(\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{y}-\bar{\mathbf{y}})=\mathbf{C}^{-1}(\mathbf{X}-\bar{\mathbf{X}})'[(\mathbf{X}-\bar{\mathbf{X}})\mathbf{b}+\mathbf{u}]$$

$$=\mathbf{b}+\mathbf{C}^{-1}(\mathbf{X}-\bar{\mathbf{X}})'\mathbf{u}$$

$\hat{\mathbf{b}}$の共分散は

$$\mathbf{S}=E[(\hat{\mathbf{b}}-\mathbf{b})(\hat{\mathbf{b}}-\mathbf{b})']=E[\mathbf{C}^{-1}(\mathbf{X}-\bar{\mathbf{X}})'\mathbf{u}\{\mathbf{C}^{-1}(\mathbf{X}-\bar{\mathbf{X}})'\mathbf{u}\}']$$

$$=\mathbf{C}^{-1}E[(\mathbf{X}-\bar{\mathbf{X}})'\mathbf{u}\mathbf{u}'(\mathbf{X}-\bar{\mathbf{X}})]\mathbf{C}^{-1}=\mathbf{C}^{-1}\mathbf{W}\mathbf{C}^{-1}$$

$$(\mathbf{W})_{ij}=E[(\mathbf{X}-\bar{\mathbf{X}})'\mathbf{u}\mathbf{u}'(\mathbf{X}-\bar{\mathbf{X}}))_{ij}]=\sum_{\lambda=1}^{N}\sum_{\lambda'=1}^{N}E[(x_{i\lambda}-\bar{x}_i)u_\lambda u_{\lambda'}(x_{j\lambda'}-\bar{x}_j)]$$

$$=\sum_{\lambda=1}^{N}E[(x_{i\lambda}-\bar{x}_i)u_\lambda^2(x_{j\lambda}-\bar{x}_j)]=[(\mathbf{X}-\bar{\mathbf{X}})'\mathbf{U}(\mathbf{X}-\bar{\mathbf{X}})]_{ij}$$

$$(\mathbf{U})_{\lambda\lambda'}=u_\lambda^2\delta_{\lambda\lambda'}$$

推定値としては以下とする。

$$(\mathbf{W})_{ij}=\frac{N}{N-p-1}\sum_{\lambda=1}^{N}(x_{i\lambda}-\bar{x}_i)\hat{u}_\lambda^2(x_{j\lambda}-\bar{x}_j)$$

結合仮説$\mathbf{b}=\mathbf{0}$の検定は、

$$\mathbf{S}=E[(\hat{\mathbf{b}}-\mathbf{b})(\hat{\mathbf{b}}-\mathbf{b})']=\mathbf{C}^{-1}\mathbf{W}\mathbf{C}^{-1}$$

$$\hat{\mathbf{b}}'\mathbf{S}^{-1}\hat{\mathbf{b}} = \hat{\mathbf{b}}'(\mathbf{C}^{-1}\mathbf{W}\mathbf{C}^{-1})^{-1}\hat{\mathbf{b}} = \hat{\mathbf{b}}'\mathbf{C}\mathbf{W}^{-1}\mathbf{C}\hat{\mathbf{b}}$$
$$= \hat{\mathbf{b}}'(\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{X}-\bar{\mathbf{X}})\mathbf{W}^{-1}(\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{X}-\bar{\mathbf{X}})\hat{\mathbf{b}}$$
$$= (\mathbf{Y}-\bar{\mathbf{Y}})'(\mathbf{X}-\bar{\mathbf{X}})\mathbf{W}^{-1}(\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{Y}-\bar{\mathbf{Y}})$$
$$\hat{\mathbf{b}}'\mathbf{S}^{-1}\hat{\mathbf{b}}\big/p \sim F_{p,\infty}$$

2.5　重回帰分析の拡張した形式による解説【補足】

行列と分布の公式

この節で使う統計の公式を上げておく。$\mathbf{u}$ が確率変数である。

(公式 1) $Cov(\mathbf{Au}, \mathbf{Bu}) = \mathbf{A}\Sigma_{\mathbf{u}}\mathbf{B}'$

(公式 2) $\mathbf{A}\Sigma_{\mathbf{u}}\mathbf{B}' = \mathbf{0}$ ならば、$\mathbf{Au}$ と $\mathbf{Bu}$ は独立した分布

(公式 3) $\mathbf{u}(m\times 1) \sim N(\mathbf{0}, \Sigma_{\mathbf{u}})$ のとき、

$$\mathbf{d} + \mathbf{Au} \sim N(\mathbf{d}, \mathbf{A}\Sigma_{\mathbf{u}}\mathbf{A}')$$
$$\mathbf{u}'\Sigma_{\mathbf{u}}^{-1}\mathbf{u} \sim \chi_m^2$$

(公式 4) $\mathbf{u} \sim N(0, \mathbf{I}_m)$ で $\mathbf{C}(m\times m)$ がべき等行列（$\mathbf{CC} = \mathbf{C}$）のとき、

$$\mathbf{u}'\mathbf{Cu} \sim \chi_r^2 \quad \text{但し、} rank(\mathbf{C}) = r$$

均一分散誤差の場合の理論

目的変数を k 個の説明変数と定数項で回帰する重回帰式を以下のように仮定する。

$$\mathbf{y} = \mathbf{Zd} + \mathbf{u}$$

ここに、

$$\mathbf{y}(N\times 1),\quad \mathbf{Z} = (\mathbf{1} \quad \mathbf{X}(N\times k)),\quad \mathbf{d}' = (b_0 \quad \mathbf{b}'(1\times k)),$$
$$\mathbf{u}(N\times 1) \sim N(\mathbf{0}, \sigma^2\mathbf{I}_N)$$

最小 2 乗法で以下の量の最小化を考える。

$$L = (\mathbf{y}-\mathbf{Zd})'(\mathbf{y}-\mathbf{Zd})$$

これを回帰係数 $\mathbf{d}$ で微分して、回帰係数の推定値 $\hat{\mathbf{d}}$ を求めると以下となる。

$$\frac{\partial L}{\partial \mathbf{d}} = -2\mathbf{Z}'(\mathbf{y}-\mathbf{Zd}) = \mathbf{0} \quad \text{より、} \hat{\mathbf{d}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

以後、推定値は真の値 a に対して $\hat{a}$ で表す。この $\hat{\mathbf{d}}$ を書き換えると、

$$\hat{\mathbf{d}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{Zd}+\mathbf{u}) = \mathbf{d} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u}$$

となり、これを用いると $\hat{\mathbf{d}}$ の平均と分散は以下となる。

$$E[\hat{\mathbf{d}}] = \mathbf{d} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'E[\mathbf{u}] = \mathbf{d}$$
$$Cov[\hat{\mathbf{d}}, \hat{\mathbf{d}}] = E[(\hat{\mathbf{d}}-\mathbf{d})(\hat{\mathbf{d}}-\mathbf{d})'] = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'E[\mathbf{uu}']\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}$$
$$= \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1} \equiv \Sigma_{\hat{\mathbf{d}}}$$

これより、【公式３】を用いると

$$\hat{\mathbf{d}} - \mathbf{d} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u} \sim N(\mathbf{0}, (\mathbf{Z}'\mathbf{Z})^{-1}\sigma^2)$$

よって、

$$(\hat{\mathbf{d}}-\mathbf{d})_i \Big/ \sqrt{(\mathbf{Z}'\mathbf{Z})^{-1}]_{ii}}\sigma \sim N(0,1) \qquad (1)$$

回帰係数の検定

推定値を使って $\mathbf{y}=\mathbf{Z}\hat{\mathbf{d}}+\hat{\mathbf{u}}$ と定義すると、

$$\hat{\mathbf{u}}=-\mathbf{Z}\hat{\mathbf{d}}+\mathbf{y}=-\mathbf{Z}(\hat{\mathbf{d}}-\mathbf{d})+\mathbf{u}$$
$$=-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u}+\mathbf{u}=(\mathbf{I}-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')\mathbf{u}$$

表式の簡単化のために、以下の定義をする。

$$\mathbf{P}_\mathbf{Z}\equiv\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}',\quad \mathbf{M}_\mathbf{Z}\equiv\mathbf{I}-\mathbf{P}_\mathbf{Z}$$

対象行列 $\mathbf{P}_\mathbf{Z}$ と $\mathbf{M}_\mathbf{Z}$ はべき等行列であり、$\mathbf{P}_\mathbf{Z}\mathbf{M}_\mathbf{Z}=\mathbf{M}_\mathbf{Z}\mathbf{P}_\mathbf{Z}=\mathbf{0}$ も成り立つ。

また、べき等行列の固有値は 0 か 1 であることから、以下も分かる。

$$rank(\mathbf{P}_\mathbf{Z})=k+1,\quad rank(\mathbf{M}_\mathbf{Z})=N-k-1$$

これらを使うと、

$$\hat{\mathbf{u}}=(\mathbf{I}-\mathbf{P}_\mathbf{Z})\mathbf{u}=\mathbf{M}_\mathbf{Z}\mathbf{u}$$

ここで、u_λ の分散 σ^2 が不明であるので、以下の推定値 $\hat{\sigma}^2$ で置き換える。

$$\hat{\sigma}^2=\hat{\mathbf{u}}'\hat{\mathbf{u}}/(N-k-1)=EV/(N-k-1)$$

最初にこの置き換えの妥当性をみるために、この式が σ^2 の不偏推定量であることを示す。それには以下の関係を使う。

$$\hat{\mathbf{u}}'\hat{\mathbf{u}}=\mathbf{u}'\mathbf{M}_\mathbf{Z}\mathbf{M}_\mathbf{Z}\mathbf{u}=\mathbf{u}'\mathbf{M}_\mathbf{Z}\mathbf{u}$$

これを用いると、

$$E[\hat{\mathbf{u}}'\hat{\mathbf{u}}]=E[\mathbf{u}'\mathbf{M}_\mathbf{Z}\mathbf{u}]=\sigma^2\sum_{\lambda=1}^{N}(\mathbf{M}_\mathbf{Z})_{\lambda\lambda}=\sigma^2 tr(\mathbf{M}_\mathbf{Z})=(N-k-1)\sigma^2$$
$$E[\hat{\sigma}^2]=E[\hat{\mathbf{u}}'\hat{\mathbf{u}}/(N-k-1)]=\sigma^2$$

次に、$\mathbf{u}/\sigma\sim N(0,\mathbf{I}_N)$ であるから、【公式４】を用いて以下となる。

$$EV/\sigma^2=\hat{\mathbf{u}}'\hat{\mathbf{u}}/\sigma^2=(\mathbf{u}'/\sigma)\mathbf{M}_\mathbf{Z}(\mathbf{u}/\sigma)\sim\chi^2_{N-k-1} \qquad (2)$$

最後に $\hat{\mathbf{d}}-\mathbf{d}$ と EV が独立であることを示す。$\mathbf{M}_\mathbf{Z}$ がべき等行列であり、

$$EV=\hat{\mathbf{u}}'\hat{\mathbf{u}}=\mathbf{u}'\mathbf{M}_\mathbf{Z}\mathbf{u}=\mathbf{u}'\mathbf{M}_\mathbf{Z}\mathbf{M}_\mathbf{Z}\mathbf{u}$$

の関係から、$\hat{\mathbf{d}}-\mathbf{d}$ と $\mathbf{M}_\mathbf{Z}\mathbf{u}$ の独立性を示せばよい。

$$\hat{\mathbf{d}}-\mathbf{d}=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u},\quad \mathbf{M}_\mathbf{Z}\mathbf{u}=(\mathbf{I}-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')\mathbf{u}$$

であるから、

$$(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}_\mathbf{u}(\mathbf{I}-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')'=\sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{I}(\mathbf{I}-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')'$$
$$=\sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'-\sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'=\mathbf{0}$$

よって、【公式２】より、$\hat{\mathbf{d}}-\mathbf{d}$ と EV は独立である。

(1)式と(2)式、および$\hat{\mathbf{d}}-\mathbf{d}$と$EV$の独立性より、

$$t_i \equiv \frac{(\hat{\mathbf{d}}-\mathbf{d})_i}{\sqrt{[(\mathbf{Z}'\mathbf{Z})^{-1}]_{ii}EV/(N-k-1)}} \sim t_{N-k-1} \quad , \quad EV=\hat{\mathbf{u}}'\hat{\mathbf{u}}$$

結合仮説の検定

ここではq個の制約が付いた検定$\mathbf{R}(\hat{\mathbf{d}}-\mathbf{d})=\mathbf{0}(q\times 1)$を考える。これが同時に成り立つことを検定する場合、【公式３】を利用する。

$$[\mathbf{R}(\hat{\mathbf{d}}-\mathbf{d})]'\boldsymbol{\Sigma}^{-1}_{\mathbf{R}(\hat{\mathbf{d}}-\mathbf{d})}\mathbf{R}(\hat{\mathbf{d}}-\mathbf{d}) \sim \chi^2_q$$

ここに、$rank(\mathbf{R})=q$

$$\boldsymbol{\Sigma}_{\mathbf{R}(\hat{\mathbf{d}}-\mathbf{d})}=\boldsymbol{\Sigma}_{\mathbf{R}\hat{\mathbf{d}}}=\mathbf{R}\boldsymbol{\Sigma}_{\hat{\mathbf{d}}}\mathbf{R}'=\sigma^2\mathbf{R}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{R}'$$

より、$\mathbf{R}\mathbf{d}=\mathbf{r}$とすると

$$[\mathbf{R}(\hat{\mathbf{d}}-\mathbf{d})]'\boldsymbol{\Sigma}^{-1}_{\mathbf{R}\hat{\mathbf{d}}}\mathbf{R}(\hat{\mathbf{d}}-\mathbf{d})$$

$$=\frac{1}{\sigma^2}(\mathbf{R}\hat{\mathbf{d}}-\mathbf{r})'[\mathbf{R}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\mathbf{d}}-\mathbf{r}) \sim \chi^2_q$$

また、(2)式より、

$$EV/\sigma^2=\hat{\mathbf{u}}'\hat{\mathbf{u}}/\sigma^2=\mathbf{u}'\mathbf{M}_{\mathbf{x}}\mathbf{u}/\sigma^2 \sim \chi^2_{N-k-1}$$

さらに、以前の議論から$\hat{\mathbf{d}}-\mathbf{d}$と$EV$は独立であるので、

$$F=\frac{(\mathbf{R}\hat{\mathbf{d}}-\mathbf{r})'[\mathbf{R}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\mathbf{d}}-\mathbf{r})/q}{EV/(N-k-1)} \sim F_{q,N-k-1}$$

例として、単独の検定$b_i=0$について考える。その場合は以下である。

$$\mathbf{R}(1\times(k+1))=[0 \quad \mathbf{0}(1\times(i-1)) \quad 1 \quad \mathbf{0}(1\times(k-i))], \quad r=0$$

また、有効性の検定$b_1=0,\cdots,b_k=0$の場合は以下となる。

$$\mathbf{R}(k\times(k+1))=[\mathbf{0}(k\times 1) \quad \mathbf{I}(k\times k)], \quad \mathbf{r}(k\times 1)=\mathbf{0}$$

また、結合仮説検定$b_1=0,\cdots,b_q=0$の場合は以下である。

$$\mathbf{R}(q\times(k+1))=[\mathbf{0}(q\times 1) \quad \mathbf{I}(q\times q) \quad \mathbf{0}(q\times(k-q))], \quad \mathbf{r}(q\times 1)=\mathbf{0}$$

定数項を分離した形式との比較

これらの検定を、定数項を分離する形式と比較してみる。回帰係数の検定では、

$$t_i=\frac{(\hat{\mathbf{d}}-\mathbf{d})_i}{\sqrt{[(\mathbf{Z}'\mathbf{Z})^{-1}]_{ii}EV/(N-k-1)}} \sim t_{N-k-1}, \qquad \mathbf{Z}'\mathbf{Z}=\begin{pmatrix} N & N\bar{\mathbf{x}}' \\ N\bar{\mathbf{x}} & \mathbf{X}'\mathbf{X} \end{pmatrix}$$

$(\mathbf{Z}'\mathbf{Z})^{-1}$はどのような形をしているのであろうか。そのために、以下の公式を利用する。

$\mathbf{A}=\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$ならば、$\mathbf{A}^{-1}=\mathbf{B}=\begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}$とすると、

$$\mathbf{B}_{11}=\mathbf{A}_{11}^{-1}+\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{C}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}, \quad \mathbf{B}_{12}=-\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{C}^{-1}, \quad \mathbf{B}_{22}=\mathbf{C}^{-1}$$

ここに、$\mathbf{C}=\mathbf{A}_{22}-\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$

この公式を用いると、

$(\mathbf{Z}'\mathbf{Z})^{-1}=\mathbf{B}=\begin{pmatrix}\mathbf{B}_{11} & \mathbf{B}_{12}\\ \mathbf{B}_{21} & \mathbf{B}_{22}\end{pmatrix}$　として、

$\mathbf{B}_{11}=1/N+\bar{\mathbf{x}}'\mathbf{C}^{-1}\bar{\mathbf{x}}$，　$\mathbf{B}_{12}=-\bar{\mathbf{x}}'\mathbf{C}^{-1}$，　$\mathbf{B}_{21}=-\mathbf{C}^{-1}\bar{\mathbf{x}}$，　$\mathbf{B}_{22}=\mathbf{C}^{-1}$

ここに、

$\mathbf{C}=\mathbf{X}'\mathbf{X}-N\bar{\mathbf{x}}\bar{\mathbf{x}}'=\mathbf{X}'\mathbf{X}-\bar{\mathbf{X}}'\bar{\mathbf{X}}=(\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{X}-\bar{\mathbf{X}})$

$\bar{\mathbf{x}}=\begin{pmatrix}\bar{x}_1\\ \vdots\\ \bar{x}_k\end{pmatrix}$，　$\bar{\mathbf{X}}=\begin{pmatrix}\bar{\mathbf{x}}'\\ \vdots\\ \bar{\mathbf{x}}'\end{pmatrix}=\begin{pmatrix}\bar{x}_1 & \cdots & \bar{x}_k\\ \vdots & \ddots & \vdots\\ \bar{x}_1 & \cdots & \bar{x}_k\end{pmatrix}$

以上、まとめると

$(\mathbf{Z}'\mathbf{Z})^{-1}=\begin{pmatrix}1/N+\bar{\mathbf{x}}'\mathbf{C}^{-1}\bar{\mathbf{x}} & -\bar{\mathbf{x}}'\mathbf{C}^{-1}\\ -\mathbf{C}^{-1}\bar{\mathbf{x}} & \mathbf{C}^{-1}\end{pmatrix}$，　　$\mathbf{C}=(\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{X}-\bar{\mathbf{X}})$

この関係を使って、もう一つ式を作っておく。

$\mathbf{y}=\begin{pmatrix}y_1\\ \vdots\\ y_N\end{pmatrix}$，　$\bar{y}=\dfrac{1}{N}\sum_{i=1}^{N}y_i$，　$\bar{\mathbf{y}}=\begin{pmatrix}\bar{y}\\ \vdots\\ \bar{y}\end{pmatrix}$

$$(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}=\begin{pmatrix}1/N+\bar{\mathbf{x}}'\mathbf{C}^{-1}\bar{\mathbf{x}} & -\bar{\mathbf{x}}'\mathbf{C}^{-1}\\ -\mathbf{C}^{-1}\bar{\mathbf{x}} & \mathbf{C}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{1}'\\ \mathbf{X}'\end{pmatrix}\mathbf{y}$$

$$=\begin{pmatrix}1/N+\bar{\mathbf{x}}'\mathbf{C}^{-1}\bar{\mathbf{x}} & -\bar{\mathbf{x}}'\mathbf{C}^{-1}\\ -\mathbf{C}^{-1}\bar{\mathbf{x}} & \mathbf{C}^{-1}\end{pmatrix}\begin{pmatrix}N\bar{y}\\ \mathbf{X}'\mathbf{y}\end{pmatrix}=\begin{pmatrix}\bar{y}+N\bar{\mathbf{x}}'\mathbf{C}^{-1}\bar{\mathbf{x}}\bar{y}-\bar{\mathbf{x}}'\mathbf{C}^{-1}\mathbf{X}'\mathbf{y}\\ -\mathbf{C}^{-1}\bar{\mathbf{x}}N\bar{y}+\mathbf{C}^{-1}\mathbf{X}'\mathbf{y}\end{pmatrix}$$

$$=\begin{pmatrix}\bar{y}+\bar{\mathbf{x}}'\mathbf{C}^{-1}\bar{\mathbf{X}}'\bar{\mathbf{y}}-\bar{\mathbf{x}}'\mathbf{C}^{-1}\mathbf{X}'\mathbf{y}\\ -\mathbf{C}^{-1}\bar{\mathbf{X}}'\bar{\mathbf{y}}+\mathbf{C}^{-1}\mathbf{X}'\mathbf{y}\end{pmatrix}=\begin{pmatrix}\bar{y}-\bar{\mathbf{x}}'\mathbf{C}^{-1}(\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{y}-\bar{\mathbf{y}})\\ \mathbf{C}^{-1}(\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{y}-\bar{\mathbf{y}})\end{pmatrix}$$

以上より

$$t_0=\frac{\hat{b}_0-b_0}{\sqrt{[(\mathbf{Z}'\mathbf{Z})^{-1}]_{00}EV/(N-k-1)}}$$

$$=\frac{\hat{b}_0-b_0}{\sqrt{(1/N+\bar{\mathbf{x}}'\mathbf{C}^{-1}\bar{\mathbf{x}})EV/(N-k-1)}}=\sim t_{N-k-1}$$

$$t_i=\frac{\hat{b}_i-b_i}{\sqrt{[(\mathbf{Z}'\mathbf{Z})^{-1}]_{ii}EV/(N-k-1)}}=\frac{\hat{b}_i-b_i}{\sqrt{[\mathbf{C}^{-1}]_{ii}EV/(N-k-1)}}\sim t_{N-k-1}$$

また、有効性の検定は、

$$F=\frac{(\mathbf{R}\hat{\mathbf{b}}-\mathbf{r})'[\mathbf{R}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\mathbf{b}}-\mathbf{r})/q}{EV/(N-k-1)}\sim F_{q,N-k-1}$$

より、$\mathbf{R}(k\times(k+1))=\begin{bmatrix}\mathbf{0}(k\times1) & \mathbf{I}(k\times k)\end{bmatrix}$，　$\mathbf{r}(k\times1)=\mathbf{0}$　として、

$$F = \frac{\hat{\mathbf{b}}'\mathbf{C}\hat{\mathbf{b}}/k}{EV/(N-k-1)} \sim F_{k,N-k-1}$$

不均一分散誤差の場合の理論

ここでは不均一分散誤差の場合の重回帰分析について説明するが、均一分散の場合にそって一から始めることにする。

目的変数を k 個の説明変数と定数項で回帰する重回帰式を以下のように仮定する。

$$\mathbf{y}=\mathbf{Z}\mathbf{d}+\mathbf{u}$$

ここに、$\mathbf{y}(N\times 1)$，$\mathbf{Z}=\begin{pmatrix}\mathbf{1} & \mathbf{X}(N\times k)\end{pmatrix}$，$\mathbf{d}'=\begin{pmatrix}b_0 & \mathbf{b}'(1\times k)\end{pmatrix}$，$\mathbf{u}(N\times 1)$

$$\mathbf{u}'(1\times N)=\begin{pmatrix}u_1 & \cdots & u_N\end{pmatrix},\quad u_\lambda \sim N(\mathbf{0},\sigma_\lambda^2),\quad Cov(u_\lambda,u_{\lambda'})=\sigma_\lambda^2\delta_{\lambda\lambda'}$$

最小２乗法で以下の量の最小化を考える。

$$L=(\mathbf{y}-\mathbf{Z}\mathbf{d})'(\mathbf{y}-\mathbf{Z}\mathbf{d})$$

回帰係数 $\mathbf{d}$ で微分して、回帰係数の推定値 $\hat{\mathbf{d}}$ を求めると以下となる。

$$\frac{\partial L}{\partial \mathbf{d}}=-2\mathbf{Z}'(\mathbf{y}-\mathbf{Z}\mathbf{d})=\mathbf{0}\quad \text{より、}\ \hat{\mathbf{d}}=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

この $\hat{\mathbf{d}}$ を書き換えると、

$$\hat{\mathbf{d}}=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{Z}\mathbf{d}+\mathbf{u})=\mathbf{d}+(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u}$$

となり、これを用いると $\hat{\mathbf{d}}$ の平均と分散は、

$$E[\hat{\mathbf{d}}]=\mathbf{d}+(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'E[\mathbf{u}]=\mathbf{d}$$

$$Cov[\hat{\mathbf{d}},\hat{\mathbf{d}}]=E[(\hat{\mathbf{d}}-\mathbf{d})(\hat{\mathbf{d}}-\mathbf{d})']=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'E[\mathbf{u}\mathbf{u}']\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}$$

$$=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{U}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}$$

ここに、$(\mathbf{U})_{\lambda\lambda'}=E[u_\lambda u_{\lambda'}]=\sigma_\lambda^2\delta_{\lambda\lambda'}$

以上より、回帰係数は以下の分布となる。

$$\hat{\mathbf{d}}-\mathbf{d}=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u}\sim N(\mathbf{0},\mathbf{\Sigma}_{\hat{\mathbf{d}}}),\qquad \mathbf{\Sigma}_{\hat{\mathbf{d}}}\equiv(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{U}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}$$

分散の推定値としては $\mathbf{U}$ の代わりに自由度を考えて以下と置く。

$$(\hat{\mathbf{U}})_{\lambda\lambda'}=\frac{N\hat{u}_\lambda^2}{N-k-1}\delta_{\lambda\lambda'}$$

計算には上の推定値を用いて、

$$\hat{\mathbf{d}}-\mathbf{d}=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u}\sim N(\mathbf{0},\hat{\mathbf{\Sigma}}_{\hat{\mathbf{d}}}),\qquad \hat{\mathbf{\Sigma}}_{\hat{\mathbf{d}}}\equiv(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{U}}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}$$

回帰係数の検定

上で述べた関係式より、回帰係数の検定には以下の関係を用いる。

$$(\hat{b}_i-b_i)\Big/\sqrt{(\hat{\mathbf{\Sigma}}_{\hat{\mathbf{d}}})_{ii}}\sim N(0,1)$$

結合仮説の検定

ここでは制約が付いた q 個の検定 $\mathbf{R}(\hat{\mathbf{b}}-\mathbf{b})=\mathbf{0}(q\times 1)$ を考える。これが同時に成り立つことを検定する場合、【公式 3】を利用する。

$$[\mathbf{R}(\hat{\mathbf{d}}-\mathbf{d})]'\mathbf{\Sigma}^{-1}_{\mathbf{R}(\hat{\mathbf{d}}-\mathbf{d})}\mathbf{R}(\hat{\mathbf{d}}-\mathbf{d})\sim\chi_q^2$$

ここに、$rank(\mathbf{R}) = q$ とする。また、

$$\boldsymbol{\Sigma}_{\mathbf{R}(\hat{\mathbf{d}}-\mathbf{d})} = \boldsymbol{\Sigma}_{\mathbf{R}\hat{\mathbf{d}}} = \mathbf{R}\boldsymbol{\Sigma}_{\hat{\mathbf{d}}}\mathbf{R}' \equiv \mathbf{R}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{U}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{R}'$$

より、$\mathbf{Rd} = \mathbf{r}$ とすると

$$[\mathbf{R}(\hat{\mathbf{d}}-\mathbf{d})]'\boldsymbol{\Sigma}_{\mathbf{R}\hat{\mathbf{d}}}^{-1}\mathbf{R}(\hat{\mathbf{d}}-\mathbf{d})$$

$$= (\mathbf{R}\hat{\mathbf{d}}-\mathbf{r})'[\mathbf{R}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{U}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\mathbf{d}}-\mathbf{r}) \sim \chi_q^2$$

検定のために推定値を使って以下の関係を利用する。

$$(\mathbf{R}\hat{\mathbf{d}}-\mathbf{r})'[\mathbf{R}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{U}}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\mathbf{d}}-\mathbf{r}) \sim \chi_q^2$$

また、F 検定に書式を合わせると以下となる。

$$F = (\mathbf{R}\hat{\mathbf{d}}-\mathbf{r})'[\mathbf{R}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{U}}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\mathbf{d}}-\mathbf{r})\big/q \sim F_{q,\infty}$$

標準化残差とてこ比について

復習になるが、最小２乗法を用いると、偏回帰係数の推定値 $\hat{\mathbf{d}}$ は以下となる。

$$\hat{\mathbf{d}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{Z}\mathbf{d}+\mathbf{u}) = \mathbf{d} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u}$$

残差 $\mathbf{u}$ の大きさの異常性や残差の説明変数依存性などを調べるために、標準化残差という指標が使われる。残差の不偏分散を $V_e = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(N-p-2)$ として、標準化残差は $\mathbf{u}\big/\sqrt{V_e}$ で与えられる。

次に、回帰分析の予測値は以下のように与えられるが、

$$\hat{\mathbf{y}} = \mathbf{Z}\hat{\mathbf{d}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \equiv \mathbf{H}\mathbf{y}$$

これを見ると実測値の変化が予測値に影響を与えることが分かる。この $\mathbf{H}$ の対角成分をてこ比と呼ぶ。さらに詳しく見てみよう。上式は以下のように書ける。

$$\hat{\mathbf{y}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \begin{pmatrix}\mathbf{1} & \mathbf{X}\end{pmatrix}\begin{pmatrix}1/N + \bar{\mathbf{x}}'\mathbf{C}^{-1}\bar{\mathbf{x}} & -\bar{\mathbf{x}}'\mathbf{C}^{-1} \\ -\mathbf{C}^{-1}\bar{\mathbf{x}} & \mathbf{C}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{1}' \\ \mathbf{X}'\end{pmatrix}\mathbf{y}$$

$$= \begin{pmatrix}\mathbf{1} & \mathbf{X}\end{pmatrix}\begin{pmatrix}(1/N + \bar{\mathbf{x}}'\mathbf{C}^{-1}\bar{\mathbf{x}})\mathbf{1}' - \bar{\mathbf{x}}'\mathbf{C}^{-1}\mathbf{X}' \\ -\mathbf{C}^{-1}\bar{\mathbf{x}}\mathbf{1}' + \mathbf{C}^{-1}\mathbf{X}'\end{pmatrix}\mathbf{y}$$

$$= \left[\mathbf{1}(1/N + \bar{\mathbf{x}}'\mathbf{C}^{-1}\bar{\mathbf{x}})\mathbf{1}' - \mathbf{1}\bar{\mathbf{x}}'\mathbf{C}^{-1}\mathbf{X}' - \mathbf{X}\mathbf{C}^{-1}\bar{\mathbf{x}}\mathbf{1}' + \mathbf{X}\mathbf{C}^{-1}\mathbf{X}'\right]\mathbf{y}$$

$$= \left[1/N\,\mathbf{1}\mathbf{1}' + (\mathbf{X}-\bar{\mathbf{X}})\mathbf{C}^{-1}(\mathbf{X}-\bar{\mathbf{X}})'\right]\mathbf{y}$$

ここに、$\mathbf{1}\bar{\mathbf{x}}' = \bar{\mathbf{X}}$，$\mathbf{C} = (\mathbf{X}-\bar{\mathbf{X}})'(\mathbf{X}-\bar{\mathbf{X}})$

これを成分で書き表すと以下となる。

$$\hat{y}_\lambda = \frac{1}{N}\sum_{\mu=1}^{N} y_\mu + \sum_{\mu=1}^{N}\sum_{j=1}^{p}\sum_{k=1}^{p}(x_{j\lambda}-\bar{x}_j)S^{jk}(x_{k\mu}-\bar{x}_k)y_\mu \equiv \sum_{\mu=1}^{N} h_{\lambda\mu}y_\mu$$

よって、てこ比 $h_{\lambda\lambda}$ は以下のように表される。

$$h_{\lambda\lambda} = \frac{1}{N} + \sum_{j=1}^{p}\sum_{k=1}^{p}(x_{j\lambda}-\bar{x}_j)S^{jk}(x_{k\lambda}-\bar{x}_k)$$

てこ比 $h_{\lambda\lambda}$ を見ると、平均から離れているデータほど、$\hat{y}_\lambda$ への影響は大きくなる。この係数がてこ比と呼ばれているのも分かる。てこ比の目安は、てこ比の平均の 2.5 倍未満と言われている。てこ比の平均は以下である。

$$\frac{1}{N}\sum_{\lambda=1}^{N} h_{\lambda\lambda} = \frac{1}{N} tr\mathbf{H} = \frac{1}{N}(1 + tr\mathbf{I}_p) = (p+1)/N$$

予測値の分布

てこ比に関連して予測値の分布についても見ておく。

$$\hat{\mathbf{y}} = \mathbf{Z}\hat{\mathbf{d}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{Z}\mathbf{d}+\mathbf{u})$$
$$= \mathbf{Z}\mathbf{d} + \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u}$$

上の関係から、予測値の共分散は以下で与えられることが分かる。

$$\mathbf{\Sigma}_{\hat{\mathbf{y}}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Sigma}_{\mathbf{u}}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$$

これから、以下のようになる。但し、使用に対しては $\mathbf{d}$ の代わりに $\hat{\mathbf{d}}$ の推測値を用いる。

均一分散の場合

$$\mathbf{\Sigma}_{\hat{\mathbf{y}}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{U}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \sigma^2\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \sigma^2\mathbf{H}$$

ここに、$(\mathbf{U})_{\lambda\lambda'} = E[u_\lambda u_{\lambda'}] = \sigma^2\delta_{\lambda\lambda'}$　　$\hat{\sigma}^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(N-k-1)$

$$\hat{y}_\lambda = (\mathbf{Z}\mathbf{d})_\lambda \pm t_{N-p-2}(\alpha/2)\sqrt{(\mathbf{H})_{\lambda\lambda}V_e}\,, \qquad V_e = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(N-p-2)$$

不均一分散の場合

$$\mathbf{\Sigma}_{\hat{\mathbf{y}}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{U}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \mathbf{Z}\mathbf{\Sigma}_{\hat{\mathbf{d}}}\mathbf{Z}'$$

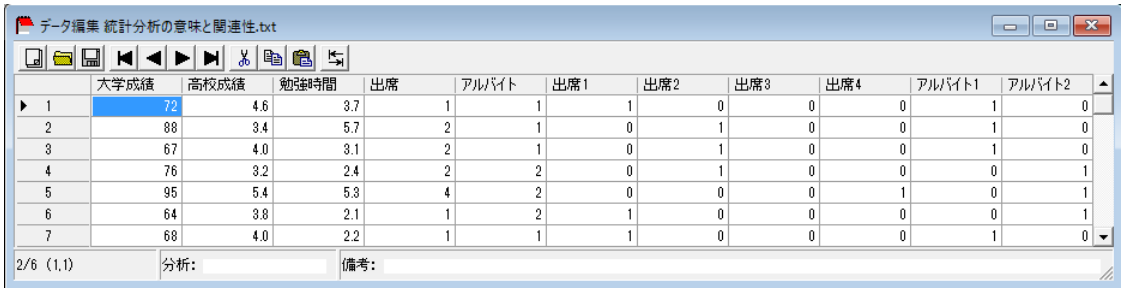
ここに、$(\mathbf{U})_{\lambda\lambda'} = E[u_\lambda u_{\lambda'}] = \sigma_\lambda^2\delta_{\lambda\lambda'}$　　$(\hat{\mathbf{U}})_{\lambda\lambda'} = \dfrac{N\hat{u}_\lambda^2}{N-k-1}\delta_{\lambda\lambda'}$

$$\hat{y}_\lambda = (\mathbf{Z}\mathbf{d})_\lambda \pm Z(\alpha/2)\sqrt{(\mathbf{\Sigma}_{\hat{\mathbf{y}}})_{\lambda\lambda}}\,, \qquad \mathbf{\Sigma}_{\hat{\mathbf{d}}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{U}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}$$

2.6　重回帰分析への質的データの投入法【補足】

一般に、重回帰分析は量的な目的変数を量的な説明変数で予測する手法で、数量化 I 類は量的な目的変数を質的な説明変数で予測する手法であると言われている。しかし、重回帰分析の説明変数には量的データだけでなく質的データも使うことができ、質的データだけの場合、その結果は数量化 I 類の結果と一致する。以上のことを実際のデータ（統計分析の意味と関連性.txt 2 頁目）を使って調べてみる。分析は、メニュー［分析－多変量解析他－予測手法－］の中の重回帰分析と数量化 I 類を使っている。

図 1 のデータは、大学成績を高校成績、勉強時間、出席状況、アルバイトの有無で予測しようとしたものである。ここに出席とアルバイトは質的なデータである。

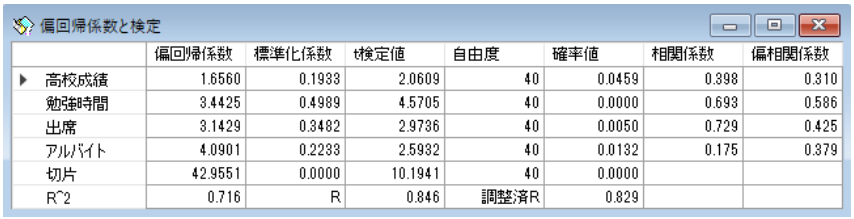

データ編集 統計分析の意味と関連性.txt

	大学成績	高校成績	勉強時間	出席	アルバイト	出席1	出席2	出席3	出席4	アルバイト1	アルバイト2
1	72	4.6	3.7	1	1	1	0	0	0	1	0
2	88	3.4	5.7	2	1	0	1	0	0	1	0
3	67	4.0	3.1	2	1	0	1	0	0	1	0
4	76	3.2	2.4	2	2	0	1	0	0	0	1
5	95	5.4	5.3	4	2	0	0	0	1	0	1
6	64	3.8	2.1	1	2	1	0	0	0	0	1
7	68	4.0	2.2	1	1	1	0	0	0	1	0

2/6 (1,1)　分析:　備考:

図１　データ

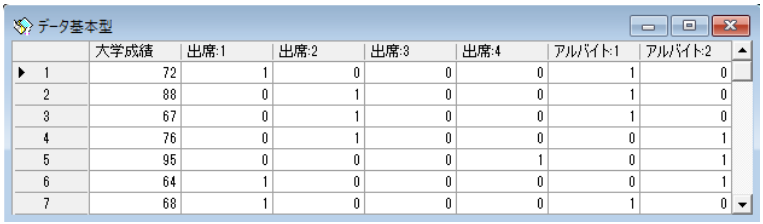
まず大学成績を目的変数に、高校成績、勉強時間、出席、アルバイトを説明変数にした重回帰分析の結果を図 2 に示す。但し、出席とアルバイトは質的データをそのまま量的データのように使っている。右下の青い部分は結果表示では別に表されるが、重要なので加えている。


偏回帰係数と検定

	偏回帰係数	標準化係数	t検定値	自由度	確率値	相関係数	偏相関係数
高校成績	1.6560	0.1933	2.0609	40	0.0459	0.398	0.310
勉強時間	3.4425	0.4989	4.5705	40	0.0000	0.693	0.586
出席	3.1429	0.3482	2.9736	40	0.0050	0.729	0.425
アルバイト	4.0901	0.2233	2.5932	40	0.0132	0.175	0.379
切片	42.9551	0.0000	10.1941	40	0.0000		
R^2	0.716	R	0.846	調整済R	0.829		

図２　順序尺度をそのまま利用した重回帰分析結果

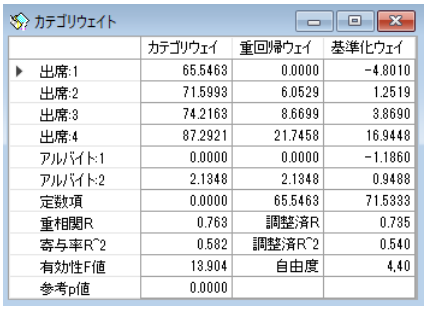
次に大学成績を目的変数に、出席とアルバイトを説明変数にして、数量化 I 類の方法で予測式を作ってみる。数量化 I 類ではデータを図 3 のような 0/1 データに変換して（ダミー変数化ともいう）重回帰分析を実行する。その係数をカテゴリウェイトという。


データ基本型

	大学成績	出席:1	出席:2	出席:3	出席:4	アルバイト:1	アルバイト:2
1	72	1	0	0	0	1	0
2	88	0	1	0	0	1	0
3	67	0	1	0	0	1	0
4	76	0	1	0	0	0	1
5	95	0	0	0	1	0	1
6	64	1	0	0	0	0	1
7	68	1	0	0	0	1	0

図３　数量化 I 類の計算のための形式

ここでは変数をアイテム、その中の分類をカテゴリという。分析結果を図 4 に示す。


カテゴリウェイト

	カテゴリウェイ	重回帰ウェイ	基準化ウェイ
出席:1	65.5463	0.0000	-4.8010
出席:2	71.5993	6.0529	1.2519
出席:3	74.2163	8.6699	3.8690
出席:4	87.2921	21.7458	16.9448
アルバイト:1	0.0000	0.0000	-1.1860
アルバイト:2	2.1348	2.1348	0.9488
定数項	0.0000	65.5463	71.5333
重相関R	0.763	調整済R	0.735
寄与率R^2	0.582	調整済R^2	0.540
有効性F値	13.904	自由度	4,40
参考p値	0.0000		

図４　数量化 I 類分析結果

数量化 I 類では、カテゴリウェイトの値がカテゴリ間の制約条件（合計が 1 になる）から一意的に定まらず、ここでは値が 3 種類の形式で与えられている。最も結果の解釈が容易

なカテゴリウェイトは基準化カテゴリウェイトである。この 0/1 形式のデータを用いて重回帰分析を行った結果と一致するのが、重回帰カテゴリウェイトである。但し、重回帰分析には、上に述べた制約条件から多重共線性という問題が生じ、すべての 0/1 データを使うことができない。そのため各アイテムの先頭カテゴリ（もちろん他のカテゴリでもよい）を除いて分析を実行する。分析結果を図 5 に示す。

偏回帰係数と検定

	偏回帰係数	標準化係数	t検定値	自由度	確率値	相関係数	偏相関係数
出席2	6.0529	0.2940	2.7341	40	0.0093	0.075	0.397
出席3	8.6699	0.2710	2.5247	40	0.0156	0.110	0.371
出席4	21.7458	0.7507	7.0178	40	0.0000	0.678	0.743
アルバイト2	2.1348	0.1165	1.1015	40	0.2773	0.175	0.172
切片	65.5463	0.0000	38.4690	40	0.0000		
R^2	0.582	R	0.763	調整済R	0.735		

図 5　質的データを用いた重回帰分析結果

この結果の偏回帰係数を数量化 I 類の重回帰カテゴリウェイトと比較すると同じものであることが分かる。

再び説明変数に量的データと質的データを混ぜることを考える。数量化 I 類の手法を用いるのであれば、0/1 データで各アイテムの先頭カテゴリを削除して重回帰分析を実行する。

偏回帰係数と検定

	偏回帰係数	標準化係数	t検定値	自由度	確率値	相関係数	偏相関係数
高校成績	1.4138	0.1650	1.6314	38	0.1111	0.398	0.256
勉強時間	3.3710	0.4885	4.3446	38	0.0001	0.693	0.576
出席2	3.7513	0.1822	1.9451	38	0.0592	0.075	0.301
出席3	4.2551	0.1330	1.3956	38	0.1709	0.110	0.221
出席4	10.8429	0.3743	2.8967	38	0.0062	0.678	0.425
アルバイト2	3.7623	0.2054	2.2645	38	0.0293	0.175	0.345
切片	51.3881	0.0000	11.7027	38	0.0000		
R^2	0.723	R	0.850	調整済R	0.824		

図 6　0/1 データに変換した順序尺度を用いた重回帰分析結果

ここで、図 2 の順序尺度をそのまま使った重回帰分析の結果とこの数量化 I 類の方法を使った結果を比較してみよう。前者の寄与率が 0.716 であるのに対して、この方法の寄与率は 0.723 と向上している。しかし、データ利用の便利さを考えると、前者の方法を使うことはそれほど悪いことではない。ただ、質的データが順序尺度でなく名義尺度の場合は、0/1 データに変換する方法を使う必要があるだろう。また、アルバイトのように 2 つのカテゴリの場合は、どちらの方法でも結果は同じである。この考え方は判別分析やその他の分析の場合にも適用できる。

ここで述べた質的データを 0/1 データにして投入する方法では、先頭カテゴリを 1 つ取り除いて利用していた。しかし、理論的には他のカテゴリを取り除くことも全く問題がない。実際にやってみると、結果の偏回帰係数はそのアイテム変数のところで異なった値を示すが、他の変数については変わらないし、重相関係数なども同じ値となる。これをどのように解釈すればよいのだろうか。

ここで、あるカテゴリ変数を取り除くということはそのカテゴリ変数の偏回帰係数を 0 と置くことと考えてみる。そうすると別々に見えた偏回帰係数間の関係が見えてくる。どの変数の偏回帰係数を 0 に置いたかに関わらず、偏回帰係数間の差は別々のモデルで同じなのである。つまり、各カテゴリ変数の偏回帰係数の値は、0 と置いたカテゴリ変数を基準と

した偏回帰係数の値になっている。このことから、これらの偏回帰係数は相対的な指標であり、その確率値などは、0 と置いたカテゴリ変数からの差の検定の確率値になっていることが分かる。

これらのカテゴリ変数をまとめたアイテムについては、絶対的な重要性が 2.2 節の終わりで説明した結合仮説検定で示され、その中のカテゴリ変数間の関係はこの重回帰分析によって示されるのである。

３．判別分析

3.1　判別分析とは

判別分析は、質的データの分類を、複数の変数で予測する手法である。その際、予測の精度やどの変数の影響が強いかなどが検討される。

判別分析の目的は 2 群（多群の場合は後に説明する）を判別する最適な１次式を求めることである。2 群の場合の判別関数は、試験の合否が勉強時間と模擬試験の平均点で表されると仮定する場合、以下のように与えられる。

判別関数　$z=b_1$ 勉強時間＋b_2 平均点＋b_0

ここに、b_1, b_2, b_0 は理論から決まるパラメータである。

判別関数で群を分けるのは、実測値を判別関数に代入した値の判別得点である。群の判別は、判別得点が 0 以上か 0 未満かによる。我々のソフトでは判別得点が 0 以上の場合、群の名前がシフト JIS コードの小さい方に判別される。例えば１群と 2 群では１群、a 群と b 群では a 群となる。

群分けを図で表現すると図 1 のようになる。ここで、白点と黒点をする線が、判別関数=0 となる線である。

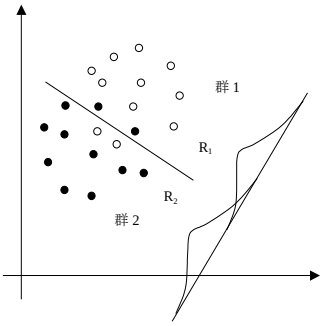


図１　判別分析での群分け

判別に影響を与える変数は、標準化係数の絶対値の大きい変数、または各係数の有効性の検定の確率値の小さい変数である。誤判別の程度は、誤判別確率で与えられるが、これには実測値と理論値がある。実測値は実際に与えられたデータで行った判別から求められるものであるが、理論値は変数の多変量正規性と等共分散性を仮定して理論的に求められる。また、各係数の有効性の検定も同じ仮定の元に求められているので、使用に際しては、ある程度注意が必要である。

マハラノビス距離とは 2 つの群がどの程度離れているかを表す指標で、標準化された距離の 2 乗で与えられている。マハラノビス距離と誤判別確率の理論値との関係は以下の表１で与えられる。

表１　マハラノビス距離と誤判別確率

マハラノビス距離	1	4	9	16	25
誤判別確率	0.309	0.159	0.067	0.023	0.006

判別分析では、予め 2 つの群に差を付けて判別する方法もある。判別の生起確率は元々

の判別群の大きさが大きく異なる場合に利用する。例えば、非常に難しい試験の場合、合格群は不合格群に比べて小さくなる。即ち、合格の確率は不合格の確率に比べて小さくなる。このような場合、データ数に応じて、判別に補正を加える。また、誤判別した場合の損失（ダメージ）が大きく異なるような場合も、判別に補正を加えることが考えられる。

今までは 2 群の場合の話をしたが、3 群以上（2 群も含む）の場合には、各群に対応した判別関数を作る。即ち、3 群の場合は、3 つの判別関数を作る。群の判別は、これらの判別関数の中にデータを代入して、最大の値となった群に属するものと判別する。

特に、2 群の場合は 2 つの判別関数の大きい方の群に属するものとするが、これまでの 2 群の場合は、2 つの判別関数の差を取って、それを 1 つの判別関数として、正になるか負になるかで判別していたのである。

以下の例を用いて実際に 2 群の判別分析を行い、結果をまとめてみよう。

例

入学試験の合否と勉強時間･模擬試験の平均点のデータを求めたところ以下のような結果を得た（判別分析 1.txt）。判別分析を実行し、結果をまとめよ。

合否	勉強時間	平均点
1	5.6	70.2
1	5.9	74.2
1	4.1	72.7
:	:	:
2	3.8	47.9
2	3.9	70.8
2	3.8	67.4
:	:	:

正規性の検定から、2 群とも正規性があるとみなされ、等共分散の検定でも共分散に差があるとは言えなかった。以上から判別分析が有効に適用可能であると判断した。

2 群の生起確率を同じとし、誤判別損失を等しいとすると、判別分析によって、以下の判別関数が得られる。

y=2.246*勉強時間+0.201*平均点-23.019

データはこの判別関数の値をもとに、判別の分点を 0 として、2 群に分けられる。

係数の有効性の検定では、勉強時間が $p<0.001$（$p=0.0001$）、平均点が $p<0.001$（$p=0.0006$）のように、両方とも有意に 0 でないことが示された。このことから 2 つの変数とも有効であると思われる。

マハラノビス距離 5.682 から、理論的な誤判別確率として p=0.117 が予想される。また、実際に判定を行うと、1 群を 2 群と間違える割合が 7.7%、その逆が 5.9%となる。これらの数値から、判別はかなりうまく行われたものと思われる。

3.2　プログラムの利用法

メニュー［分析－多変量解析他－判別手法－判別分析］をクリックすると、図 1 のような判別分析実行画面が表示される。

図 1　判別分析実行画面

データの形式は、先頭列で群分けする場合と最初から群分けされている場合が扱える。但し、後者の場合、予め群の数を入力しておかなければならない。各群の生起確率や誤判別損失の値は、オプションボタンの「指定する」を選び、テキストボックス内に値をカンマ区切りで入力することによって、自由に設定することができる。但し、確率の値は合計が 1 になることが必要であるので、無限小数の場合は 1/3 のように、分数で入力する。これらのデフォルト値は生起確率が「各群同じ」、誤判別損失が「各群 1 とする」である。

データとして判別分析 1.txt を使った場合の結果を以下に示す。2 群の判別の場合、「等共分散の検定」ボタンで等共分散性を調べることができる。図 2 に「等共分散の検定」の出力結果を示す。図 3 と図 4 に 2 群の判別分析と判別得点の出力結果を示す。判定は判別得点を判別の分点 0 と比較して決定される。

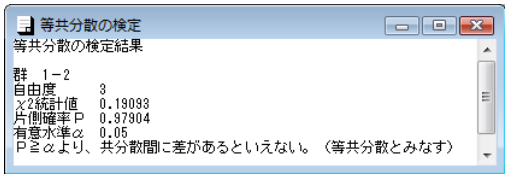


図 2　等共分散の検定出力結果（判別分析 1.txt）

図 3　判別分析出力結果（2 群の形式）

図 4　判別得点（2 群の形式）

標準化係数の定数項は、重回帰分析などでは 0 になるが、判別分析では、判別の分点を 2 つの群の群別平均のデータ数による加重平均ではなく、単純平均にしていることから、2 つの群のデータ数が異なる場合、一般に 0 にならない。

比較のために同じデータを用いて 3 群以上の判別のプログラムを実行した出力結果を図 5 と図 6 に示す。本来は 3 群以上で利用すべきであるが、2 群の判別で用いても問題はない。

判別分析

	1群判別関	2群判別関	1群標準化	2群標準化	偏λ	検定確率
勉強時間	8.737	6.491	10.195	7.574	0.576	0.0001
平均点	1.083	0.883	12.298	10.019	0.642	0.0006
定数項	−61.851	−38.833	47.197	47.576		
群間の分離	Wilks λ	0.401	検定確率	0.0000		
マハラノビスの距離	1群	2群				
1群	0.000	5.682				
2群	5.682	0.000				
誤判別確率	1群を他群と	2群を他群と				
実測から	0.077	0.059				
判別(数・確率)	1群予測	2群予測	1群予測	2群予測		
1群実測	12	1	0.923	0.077		
2群実測	1	16	0.059	0.941		

図 5　判別分析出力結果（3 群以上の形式）

判別得点：既存データの判別

	所属群	1群	2群	判別群
6	1	53.3136	52.9857	1
7	1	43.6412	44.4156	2
8	1	72.6009	67.6956	1
9	1	58.3039	56.9886	1
10	1	54.6531	52.7597	1
11	1	50.2115	49.1412	1
12	1	65.9266	61.8816	1
13	1	62.2250	59.9949	1
14	2	23.2397	28.1079	2
15	2	48.9213	48.9682	2
16	2	44.3643	45.3183	2
17	2	37.7561	39.9345	2

図 6　判別得点（3 群以上の形式）

ここで Wilks λ と横の検定確率は分析の正当性の指標であり、右端の偏 λ と検定確率は変数の重要性の指標である。

次に我々は正準形式に基づく判別の結果を示す。これは正準判別分析とも呼ばれている。正準相関分析における判別関数は、変数の数≧分割数、の場合は、分割数－1 個作られる。同じ判別分析 1.txt のデータを用いた結果を図 7 に示す。

正準判別分析

	判別1	標準化1
勉強時間	0.942	1.100
平均点	0.084	0.956
定数項	−9.656	−0.159
固有値	1.495	
寄与率	1.000	
累積寄与率	1.000	
判別の分点	0	
	1群を他群と	2群を他群と
誤判別確率	0.077	0.059

図 7　正準判別分析結果（判別分析 1.txt，2 群）

生起確率が同じで誤判別損失が 1 の場合、2 群のマハラノビス形式と正準形式の同等性から、判別関数の係数は比例している。また、判別の分点は 2 つの形式とも 0 に設定している。

判別得点

	所属群	判別得点1	判別群
1	1	1.532	1
2	1	2.151	1
3	1	0.329	1
4	1	2.298	1
5	1	2.886	1
6	1	0.138	1
7	1	−0.325	2
8	1	2.058	1
9	1	0.552	1

図 8　正準判別分析の判別得点

以後は 3 群以上である判別分析 3.txt、1 頁目のデータを用いる。正準判別分析の出力結果を図 9 に示す。

正準判別分析

	判別1	判別2	標準化1	標準化2	偏λ	確率値
がくの長さ	-0.829	0.024	-0.687	0.020	0.938	0.0103
がくの幅	-1.534	2.165	-0.669	0.943	0.766	0.0000
花弁の長さ	2.201	-0.932	3.886	-1.645	0.669	0.0000
花弁の幅	2.810	2.839	2.142	2.164	0.743	0.0000
定数項	-2.105	-6.661	0.000	0.000		
固有値	32.192	0.285				
寄与率	0.991	0.009				
累積寄与率	0.991	1.000				
群間の分離	Wilksλ	0.023	確率値	0.0000		

図 9　正準判別分析出力結果（判別分析 3.txt）

ここに Wilksλと横の検定確率は分析の正当性の指標であり、右端の偏λと検定確率は変数の重要性の指標である。ここでは標準化係数の定数項が 0 になっているが、これは 3 つの群のデータ数がすべて同じであることによる結果で、一般に定数項は 0 と異なる。

3 群の判別得点は 2 つの固有値に対応して図 10 のように 2 種類出力される。これは 2 次元上の点であるので、「軸設定」を行い、「散布図」ボタンをクリックすることにより、図 11 のような散布図が表示される。

判別得点

	所属群	判別得点1	判別得点2
1	1	-8.062	0.300
2	1	-7.129	-0.787
3	1	-7.490	-0.265
4	1	-6.813	-0.671
5	1	-8.132	0.514
6	1	-7.702	1.462
7	1	-7.213	0.356
8	1	-7.605	-0.012
9	1	-6.561	-1.015
10	1	-7.343	-0.947

図 10　正準判別分析の判別得点

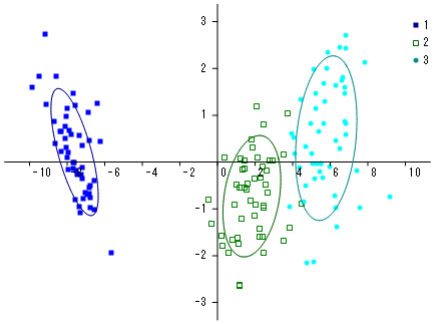


図 11　判別得点散布図

ここには、各群の分布を 2 変量正規分布とみなした場合の、1.5σ の確率楕円が示されている。確率楕円の大きさ、軸の正負の向き等はメニューで変更できる。

この 2 変量正規分布の密度関数式は、グラフメニュー「設定－正規楕円半径－密度関数数式」で図 12 のように表示される。

群別正規分布密度関数

```
).2897*exp(-0.9541*(1.4208*(x-(-7.6075))^2+1.2221*(y-(0.2151))^2+(1.8181)*(x-(-7.6075))*(y-(0.2151))))
0.1863*exp(-0.5387*(0.9506*(x-(1.8251))^2+1.3374*(y-(-0.7279))^2+(-0.6047)*(x-(1.8251))*(y-(-0.7279))))
0.1280*exp(-0.5264*(0.8448*(x-(5.7825))^2+0.7278*(y-(0.5128))^2+(-0.3515)*(x-(5.7825))*(y-(0.5128))))
```

図 12　2 変量正規分布密度関数式

この式をコピーし、分析メニュー「数学－2 変量関数グラフ」のテキストボックスに貼り付けて（[Shift+Ins] または [Ctrl+v]）、（図 11 より範囲を設定、分割数を 50 に増加、色を指定に）表示させると、図 13 のように 3 つの密度関数グラフを重ね合わせて視覚化することもできる。これによってどの程度分離ができているのか直感的に見ることもできる。

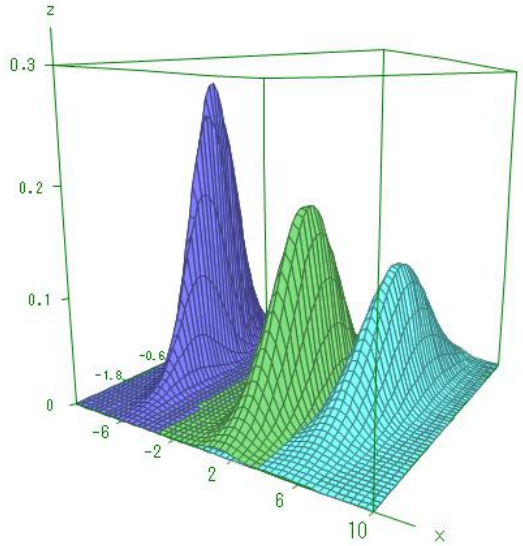


図 13　確率密度関数の視覚化

予測について

機械学習などでは、訓練データに対して独立なテストデータを用いて目的変数の予測を行う。図 1 の分析実行画面の一番下に「新データ予測」ボタンがあるが、これは一度判別分析を実行した後、新たなデータを選択して予測を行うためのボタンである。テストデータに目的変数を含む場合は「目的変数付き」チェックボックスにチェックを入れて、「新データ予測」ボタンをクリックする。テストデータで予測値を計算し、適合性を正解率の値で示してくれる。これは「先頭列で群分け」のとき有効であるが、正準判別分析では使えない。

問題１（判別分析 2.txt）

Samples¥判別分析 2.txt は、適性の有無の判定（有：1，無：2）と適性検査の結果とＳＰＩの結果を与えたデータである。判定を適性検査とＳＰＩで予測する判別分析を行い、以下の問いに答えよ。但し、事象の生起確率は各群同じ、誤判別損失は２群とも１とすること。

１）判別関数を求めよ。

判別得点＝［　　］適性検査＋［　　］ＳＰＩ＋［　　］

２）どちらの変数が判定に影響があると思われるか。［適性検査・ＳＰＩ］

３）実測値から求めた誤判別の確率は？

適性有りを無しと［　　］　適性無しを有りと［　　］

４）先頭（1 番）の人の判別得点はいくらか。［　　］

５）適性検査 50 点，ＳＰＩ 55 点の人の判別得点はいくらか、またその人の適性の有無を判定せよ。　判別得点［　　］　適性［有り・無し］

問題２（判別分析 3.txt）

Samples¥判別分析 3.txt はあやめの種類をがくの長さと幅、花弁の長さと幅で３群に分類したデータである。あやめの群を他の変数の１次式で判別する３群以上の判別分析を行い、以下の問題に答えよ。但し、設定は前問と同じとする。

１）３つの判別得点の式を求めよ。

判別得点１＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］

判別得点２＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］

判別得点３＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］

２）実測値から求めた誤判別確率はいくらか。

群１を他と［　　］群２を他と［　　］群３を他と［　　］

３）先頭のデータの３つの判別得点を求めよ。

判別得点１［　　］判別得点２［　　］判別得点３［　　］

これは何群に予想されたか。　［1 群・2 群・３群］

演習１

多変量演習 5.txt のデータを用いて、生起確率をデータ数から、誤判別損失を各群１として判別分析を行い、以下の問いに答えよ。合否の欄で、1 は合格、2 は不合格である。

１）判別関数を求めよ。
判別値＝　［　　］内申＋［　　］模試１
　　＋［　　］模試２＋［　　］
２）判別の分点［　　］
３）実測値から求めた誤判別の確率は？
合格を不合格と［　　］　不合格を合格と［　　］
４）元の設定で、各係数の有効性の検定で、5%の有意水準で有意でない変数はどれか。
変数［　　］　検定確率［　　］
５）その変数を取り除いて再度判別分析を行い、判別関数を求めよ。但し、取り除いた変数のところは空欄とせよ。
判別値＝　［　　］内申＋［　　］模試１
　　＋［　　］模試２＋［　　］
６）この場合、実測値から見た誤判別の確率はどうなるか。
合格を不合格と［　　］　不合格を合格と［　　］
７）元のモデルとこの新しいモデルとで誤判別確率に大きな差があると思われるか。
［大きな差がある・大した差ではない］と思われる。
８）新しいモデルで、先頭（1 番）の人の判別値はいくらか。［　　］
９）新しいモデルで、内申 3.4 点，模試１ 65 点，模試２ 70 点の人の判別値はいくらか、またその人の合否を判定せよ。
判別値［　　］　判定［合格・不合格］

演習２

多変量演習 6.txt のデータはある職業の適性について調べた結果である。適性は、１．適性あり、２．努力しだい、３．適性なしに分類され、それを予測するデータとして回答者の年齢、学力テスト、体力テスト、面接（10 段階）の結果が含まれている。

１ページ目のデータを用いて、生起確率をデータ数から、誤判別損失を各群１として判別分析を行い、以下の問いに答えよ。

１）３つの判別得点の式を求めよ。但し定数項は判別の分点を引いたものとする。
判別得点１＝［　　］年齢＋［　　］学力テスト
　　＋［　　］体力テスト＋［　　］面接＋［　　］
判別得点２＝［　　］年齢＋［　　］学力テスト
　　＋［　　］体力テスト＋［　　］面接＋［　　］
判別得点３＝［　　］年齢＋［　　］学力テスト
　　＋［　　］体力テスト＋［　　］面接＋［　　］
２）実測値から求めた誤判別確率はいくらか。
適性ありを他と［　　］　努力しだいを他と［　　］　適性なしを他と［　　］

３）先頭の人の３つの判別得点を求めよ。
判別得点１［　　］判別得点２［　　］判別得点３［　　］
４）先頭の人はどのように予測されているか。
［適性あり・努力しだい・適性なし］

問題１解答（判別分析 2.txt）
１）判別関数を求めよ。
判別得点＝［ -0.190 ］適性検査＋［ 0.645 ］ＳＰＩ＋［ -20.467 ］
２）どちらの変数が判定に影響があると思われるか。［適性検査・ＳＰＩ］
３）実測値から求めた誤判別の確率は？
適性有りを無しと［ 0.053 ］　適性無しを有りと［ 0.095 ］
４）先頭（1 番）の人の判別得点はいくらか。［ -5.118 ］
５）適性検査 50 点，ＳＰＩ 55 点の人の判別得点はいくらか、またその人の適性の有無を判定せよ。　判別得点［ 5.508 ］　適性［有り・無し］

問題２解答（判別分析 3.txt）
１）３つの判別得点の式を求めよ。
判別得点１＝［ 23.544 ］がくの長さ＋［ 23.588 ］がくの幅
＋［ -16.431 ］花弁の長さ＋［ -17.398 ］花弁の幅＋［ -85.210 ］
判別得点２＝［ 15.698 ］がくの長さ＋［ 7.073 ］がくの幅
＋［ 5.211 ］花弁の長さ＋［ 6.434 ］花弁の幅＋［ -71.754 ］
判別得点３＝［ 12.446 ］がくの長さ＋［ 3.685 ］がくの幅
＋［ 12.767 ］花弁の長さ＋［ 21.079 ］花弁の幅＋［--103.270 ］
２）実測値から求めた誤判別確率はいくらか。
群１を他と［ 0 ］群２を他と［ 0.040 ］群３を他と［ 0.020 ］
３）先頭のデータの３つの判別得点を求めよ。
判別得点１［ 90.940 ］判別得点２［ 41.644 ］判別得点３［-4.808 ］
これは何群に予想されたか。　［1 群・2 群・3 群］

演習１解答（多変量演習 5.txt）
１）判別関数を求めよ。
判別値＝　［ 2.100 ］内申＋［ 0.163 ］模試１＋［ 0.103 ］模試２＋［ -23.980 ］
２）判別の分点［ 0 ］

３）実測値から求めた誤判別の確率は？
合格を不合格と［ 0.185 ］　不合格を合格と［ 0.140 ］
４）元の設定で、各係数の有効性の検定で、5%の有意水準で有意でない変数はどれか。
変数［ 内申 ］　検定確率［ 0.0954 ］
５）その変数を取り除いて再度判別分析を行い、判別関数を求めよ。但し、取り除いた変数のところは空欄とせよ。
判別値＝　［　　］内申＋［ 0.199 ］模試１
＋［ 0.131 ］模試２＋［ -21.348 ］
６）この場合、実測値から見た誤判別の確率はどうなるか。
合格を不合格と［ 0.148 ］　不合格を合格と［ 0.140 ］
７）元のモデルとこの新しいモデルとで誤判別確率に大きな差があると思われるか。
［大きな差がある・大した差ではない］と思われる。

８）新しいモデルで、先頭（1 番）の人の判別値はいくらか。［ 3.782 ］
９）新しいモデルで、内申 3.4 点，模試１ 65 点，模試２ 70 点の人の判別値はいくらか、またその人の合否を判定せよ。
判別値［ 0.757 ］　判定［合格・不合格］

演習２解答（多変量演習 6.txt）

１）３つの判別得点の式を求めよ。但し定数項は判別の分点を引いたものとする。
判別得点１＝［ 4.541 ］年齢＋［ 0.683 ］学力テスト
＋［ 1.019 ］体力テスト＋［ 0.241 ］面接＋［ −124.813 ］
判別得点２＝［ 4.976 ］年齢＋［ 0.630 ］学力テスト
＋［ 0.952 ］体力テスト＋［ −0.219 ］面接＋［ −127.538 ］
判別得点３＝［ 5.288 ］年齢＋［ 0.437 ］学力テスト
＋［ 0.913 ］体力テスト＋［ −0.575 ］面接＋［ −124.026 ］
２）実測値から求めた誤判別確率はいくらか。
適性ありを他と［ 0.222 ］　努力しだいを他と［ 0.389 ］　適性なしを他と［ 0.143 ］
３）先頭の人の３つの判別得点を求めよ。
判別得点１［ 113.479 ］判別得点２［ 117.419 ］判別得点３［ 117.568 ］
４）先頭の人はどのように予測されているか。
［適性あり・努力しだい・適性なし］

3.3　判別分析の理論

判別分析は外的基準によって群別に分類されたデータから、群を判別するための線形関数を見出すことを目的としている。データは例えば 2 群の場合、表 1 のような形式で与えられる。

表 1 判別分析のデータ（2 群の場合）

群 1			群 2		
変数 1	…	変数 p	変数 1	…	変数 p
x_{11}^1	…	x_{p1}^1	x_{11}^2	…	x_{p1}^2
x_{12}^1	…	x_{p2}^1	x_{12}^2	…	x_{p2}^2
：		：	：		：
$x_{1n_1}^1$	…	$x_{pn_1}^1$	$x_{1n_2}^2$	…	$x_{pn_2}^2$

変数の一般的な表式 $x_{i\lambda}^{\alpha}$ において、α は群、i は変数、λ はレコード番号を表わす。

1）マハラノビス距離を用いた方法

ここでは、最初に 2 群の場合の理論について考える。2 つの群 G_1 と G_2 について、群 $G_1 \cup G_2$ から、G_α（$\alpha = 1,2$）の要素を取り出す確率を P_α とし、G_α の要素を G_β（$\alpha \neq \beta$）と誤判別する損失を $C_{\beta\alpha}$ とする。また、群 α の確率密度関数を $f_\alpha(\mathbf{x})$ とすると、G_α の要素を G_β と誤判別する確率 $Q_{\beta\alpha}$ は以下となる。

$$Q_{\beta\alpha} = \int_{R_\beta} f_\alpha(\mathbf{x}) d\mathbf{x}$$

ここに領域 R_β は、R_β 内の要素を G_β の要素と判別する領域である。これから、誤判別に

よる損失Lは以下のように与えられる。

$$L = C_{21}P_1Q_{21} + C_{12}P_2Q_{12}$$
$$= C_{21}P_1\int_{R_2} f_1(\mathbf{x})d\mathbf{x} + C_{12}P_2\int_{R_1} f_2(\mathbf{x})d\mathbf{x}$$
$$= C_{21}P_1\int_{R_1\cup R_2} f_1(\mathbf{x})d\mathbf{x} + \int_{R_1}[C_{12}P_2 f_2(\mathbf{x}) - C_{21}P_1 f_1(\mathbf{x})]d\mathbf{x}$$

これより、損失を最小にするためには$\boldsymbol{R}_1$として第２項の被積分関数が負になる領域を選べばよい。即ち各群の領域として、以下のような領域を考えれば良いことが分かる。

$$R_1 = \{\mathbf{x} \mid C_{12}P_2 f_2(\mathbf{x}) - C_{21}P_1 f_1(\mathbf{x}) \le 0\},$$
$$R_2 = \{\mathbf{x} \mid C_{12}P_2 f_2(\mathbf{x}) - C_{21}P_1 f_1(\mathbf{x}) > 0\}$$

これを$h = C_{12}P_2 / C_{21}P_1$として書き換えて、以下のような条件を得る。

$$R_1 = \{\mathbf{x} \mid \log f_1(\mathbf{x}) / f_2(\mathbf{x}) - \log h \ge 0\},$$
$$R_2 = \{\mathbf{x} \mid \log f_1(\mathbf{x}) / f_2(\mathbf{x}) - \log h < 0\}$$

ここに、判別の分点は０である。

今、群αの変数iの平均$\bar{x}_i^{\alpha}$と各群共通な共分散s_{ij}をそれぞれ以下のように求め、

$$\bar{x}_i^{\alpha} = \frac{1}{n_\alpha}\sum_{\lambda=1}^{n_\alpha} x_{i\lambda}, \quad s_{ij} = \frac{1}{n_1+n_2-2}\sum_{\alpha=1}^{2}\sum_{\lambda=1}^{n_\alpha}(x_{i\lambda}^{\alpha} - \bar{x}_i^{\alpha})(x_{j\lambda}^{\alpha} - \bar{x}_j^{\alpha}),$$

これらを成分とする平均ベクトル$\bar{\mathbf{x}}^{\alpha}$と共分散行列$\mathbf{S}$を用いて、以下の多変量正規分布の確率密度関数を考える。

$$f_\alpha(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k \mid \mathbf{S} \mid}} \exp\left[-\frac{1}{2}{}^t(\mathbf{x} - \bar{\mathbf{x}}^{\alpha})\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}^{\alpha})\right]$$

これを判別関数に代入して以下の線形判別関数を得る。

$$z = \log f_1(\mathbf{x}) / f_2(\mathbf{x}) - \log h$$
$$= {}^t\mathbf{x}\mathbf{S}^{-1}(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2) - \frac{1}{2}{}^t(\bar{\mathbf{x}}^1 + \bar{\mathbf{x}}^2)\mathbf{S}^{-1}(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2) - \log h$$

$\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)$ とすると、判別関数は以下のように書くことができる。

$$z = {}^t\mathbf{x}\mathbf{a} - \frac{1}{2}{}^t(\bar{\mathbf{x}}^1 + \bar{\mathbf{x}}^2)\mathbf{a} - \log h \qquad (1)$$

判別関数は、変数x_i の標準化値u_i と不偏分散s_i を用いて以下のように書くこともできる。

$$z = {}^t\mathbf{u}\mathbf{c} + {}^t\bar{\mathbf{x}}\mathbf{a} - \frac{1}{2}{}^t(\bar{\mathbf{x}}^1 + \bar{\mathbf{x}}^2)\mathbf{a} - \log h, \qquad c_i = a_i s_i \qquad (2)$$

この係数$\mathbf{c}$を標準化係数と呼ぶ。標準化係数は変数の重要性をみるときに利用される。

判別関数（1）は各群の平均$\bar{\mathbf{x}}^{\alpha}$から、$\mathbf{x}$までのマハラノビスの平方距離$D^{2(\alpha)}$の差として以下のように定義することもできる。

$$z = \frac{1}{2}(D^{2(2)} - D^{2(1)}) - \log h \quad , \quad D^{2(\alpha)} = {}^t(\mathbf{x} - \bar{\mathbf{x}}^{\alpha})\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}^{\alpha})$$

このzは$\log h$が０の場合、$\mathbf{x}$ が２つの群別平均の中央である$(\bar{\mathbf{x}}^1 + \bar{\mathbf{x}}^2)/2$のとき、０になっている。

変数 z の確率分布は、個体 $\mathbf{x}$ が群 1 に属するか、群 2 に属するかに応じて、以下のような正規分布に従うことが知られている。

$z \sim N(D^2/2, D^2)$　　$\mathbf{x} \in G_1$ の場合

$z \sim N(-D^2/2, D^2)$　　$\mathbf{x} \in G_2$ の場合

ここに、D^2 は群平均 $\overline{\mathbf{x}}^1$ と $\overline{\mathbf{x}}^2$ のマハラノビスの平方距離で、以下のように定義される。

$$D^2 = {}^t(\overline{\mathbf{x}}^1 - \overline{\mathbf{x}}^2)\mathbf{S}^{-1}(\overline{\mathbf{x}}^1 - \overline{\mathbf{x}}^2)$$

この性質から誤判別の理論確率は以下で与えられることが分かる

$$Q_{21} = \int_{-\infty}^{\log h} \frac{1}{\sqrt{2\pi D^2}} \exp\left[-\frac{(z - D^2/2)^2}{2D^2} \right] dz = Z\left(\frac{\log h - D^2/2}{D} \right)$$

$$Q_{12} = \int_{\log h}^{\infty} \frac{1}{\sqrt{2\pi D^2}} \exp\left[-\frac{(z + D^2/2)^2}{2D^2} \right] dz = 1 - Z\left(\frac{\log h + D^2/2}{D} \right)$$

これは判別分析の有効性を示している。

判別分析では、判別関数の係数についてもその有効性を検定できる。変数 i の係数が 0 であるかどうかの検定は、以下の性質を利用する。

$$F_i = \frac{(n_1 + n_2 - p - 1) n_1 n_2 (D^2 - D_i^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D_i^2} \sim F_{1, n_1 + n_2 - p - 1} \text{分布}$$

ここに、D_i^2 は両群の変数 i を除いたマハラノビスの平方距離である。

以上のような理論では、線形判別関数で表わされる判別分析がうまく利用できる条件は、分布が多変量正規分布に従うことに加えて 2 群の共分散が等しいことである。この検定には以下の性質が利用される。

$$\chi^2 = \left[1 - \left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{n_1 + n_2 - 2} \right) \frac{2p^2 + 3p - 1}{6(p+1)} \right] \log \frac{|\mathbf{S}|^{n_1 + n_2 - 2}}{|\mathbf{S}^1|^{n_1 - 1} |\mathbf{S}^2|^{n_2 - 1}} \sim \chi^2_{p(p+1)/2} \text{分布}$$

ここに、$\mathbf{S}^\alpha$ は群 α の共分散行列である。しかし、後に述べるような正準形式では、2 群の場合、分布の形を仮定することなく同等な結論を導く。

３群以上（群の数を m）の判別には以下の判別関数を考え、z^α が最大になる群 α に属するものと判定する。

$$z^\alpha = {}^t\mathbf{x}\mathbf{S}^{-1}\overline{\mathbf{x}}^\alpha - \frac{1}{2} {}^t\overline{\mathbf{x}}^\alpha \mathbf{S}^{-1} \overline{\mathbf{x}}^\alpha + \log C_\alpha P_\alpha m$$

但し、C_α は群 α を他の群と間違えた場合の損失である。定数項に含まれる m は、各群の生起確率が同じで誤判別損失が 1 の場合、これらを考えない理論と繋がるように、定数項を 0 にするための定数である。

$\mathbf{a}^\alpha = \mathbf{S}^{-1}\overline{\mathbf{x}}^\alpha$ として、この判別関数は以下のように書くこともできる。

$$z^\alpha = {}^t\mathbf{x}\mathbf{a}^\alpha - \frac{1}{2} {}^t\overline{\mathbf{x}}^\alpha \mathbf{a}^\alpha + \log C_\alpha P_\alpha m \qquad (3)$$

2 群の場合と同様に、判別関数は変数 x_i の標準化値 u_i と不偏分散 s_i を用いて以下のように書くこともできる。

$$z^{\alpha} = {}^{t}\mathbf{u}\mathbf{c}^{\alpha} + {}^{t}\bar{\mathbf{x}}\mathbf{a}^{\alpha} - \frac{1}{2}{}^{t}\bar{\mathbf{x}}^{\alpha}\mathbf{a}^{\alpha} + \log C_{\alpha}P_{\alpha}m, \qquad c_i^{\alpha} = a_i^{\alpha}s_i \tag{4}$$

この係数$\mathbf{c}^{\alpha}$を標準化係数と呼ぶ。

上で与えた 2 群の場合の判別関数は、この判別関数を用いて$z = z^1 - z^2$として求めることができる。

2) 正準形式を用いた方法

正準形式の判別分析（正準判別分析と呼ばれる）は、判別関数の拡がりを最大化するように係数を求めるもので、特に 3 群以上の場合は、判別得点を複数次元の空間上に配置し、判別をより分かり易く表現する手法である。これまでのプログラムでは、数量化Ⅱ類でその中の主要な１次元を取り出して判別する方法を導入している。以下に正準判別分析の理論を示す。

正準判別分析は、判別群で分けられたデータについて、「群間分散／群内分散」を最大化するように線形判別関数の係数を決定する手法である。判別関数を以下のように表す。ここにz_0は後に決める定数項である。

$$z = \sum_{i=1}^{p} a_i x_i + z_0$$

判別群をα, 群別のデータの番号をλ, 変数の番号をi, としてデータを$x_{i\lambda}^{\alpha}$（$\alpha = 1,\cdots,m$, $\lambda = 1,\cdots,n_{\alpha}$, $i = 1,\cdots,p$）と表す。このデータを用いて、群αのλ番目の判別関数の値z_{λ}^{α}は以下のようになる。

$$z_{\lambda}^{\alpha} = \sum_{i=1}^{p} a_i x_{i\lambda}^{\alpha} + z_0$$

このz_{λ}^{α}による群間分散s_B^2，群内分散s^2を以下のように定義する。

$$s_B^2 = \frac{1}{n-m}\sum_{\alpha=1}^{m} n_{\alpha}\left(\bar{z}^{\alpha} - \bar{z}\right)^2, \quad s^2 = \frac{1}{n-m}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_{\alpha}}\left(z_{\lambda}^{\alpha} - \bar{z}^{\alpha}\right)^2$$

ここに、$\bar{z}^{\alpha} = \frac{1}{n_{\alpha}}\sum_{\lambda=1}^{n_{\alpha}} z_{\lambda}^{\alpha}$ ，$\bar{z} = \frac{1}{n}\sum_{\alpha=1}^{m} n_{\alpha}\bar{z}^{\alpha}$ ，$n = \sum_{\alpha=1}^{m} n_{\alpha}$ である。

これより、$\bar{x}_i^{\alpha} = \frac{1}{n_{\alpha}}\sum_{\lambda=1}^{n_{\alpha}} x_{i\lambda}^{\alpha}$ ，$\bar{x}_i = \frac{1}{n}\sum_{\alpha=1}^{m} n_{\alpha}\bar{x}_i^{\alpha}$ として、s_B^2とs^2は以下のようになる。

$$s_B^2 = \frac{1}{n-m}\sum_{\alpha=1}^{m} n_{\alpha}\left[\sum_{i=1}^{p} a_i\left(\bar{x}_i^{\alpha} - \bar{x}_i\right)\right]^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} a_i b_{ij} a_j$$

$$s^2 = \frac{1}{n-m}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_{\alpha}}\left[\sum_{i=1}^{p} a_i\left(x_{i\lambda}^{\alpha} - \bar{x}^{\alpha}\right)\right]^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} a_i s_{ij} a_j$$

ここに、

$$b_{ij}=\frac{1}{n-m}\sum_{\alpha=1}^{m}n_\alpha\left(\overline{x}_i^\alpha-\overline{x}_i\right)\left(\overline{x}_j^\alpha-\overline{x}_j\right)$$

$$s_{ij}=\frac{1}{n-m}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}\left(x_{i\lambda}^\alpha-\overline{x}_i^\alpha\right)\left(x_{j\lambda}^\alpha-\overline{x}_j^\alpha\right)$$

である。行列の成分として、$\left(\mathbf{B}\right)_{ij}=b_{ij}$ ，$\left(\mathbf{S}\right)_{ij}=s_{ij}$ ，$\left(\mathbf{a}\right)_i=a_i$ とすると、s_B^2 と s^2 はこれらの行列を用いて次のように書ける。

$$s_B^2={}^t\mathbf{aBa}\quad ,\quad s^2={}^t\mathbf{aSa}$$

ここに、$n\geq m$ の場合、一般に $rank\left(\mathbf{B}\right)=m-1$，$rank\left(\mathbf{S}\right)=n-m$ である。

群間分散を群内分散で割った分散比 ρ は以下のようになる。

$$\rho=s_B^2/s^2={}^t\mathbf{aBa}/{}^t\mathbf{aSa}$$

この分散比を最大化するには、以下の解を求める。

$$\partial\rho/\partial\mathbf{a}=\frac{1}{\left(s^2\right)^2}\left[\partial s_B^2/\partial\mathbf{a}\,s^2-s_B^2\,\partial s^2/\partial\mathbf{a}\right]=\mathbf{0}$$

$\partial s_B^2/\partial\mathbf{a}=2\mathbf{Ba}$ ，$\partial s_W^2/\partial\mathbf{a}=2\mathbf{Sa}$ であるので、上の式は以下となる。

$$\mathbf{Ba}=\rho\mathbf{Sa}\qquad(5)$$

これを対称行列の固有方程式にするために、適当な下三角行列 $\mathbf{F}$ を用いて対称行列$\mathbf{S}$ を $\mathbf{S}=\mathbf{F}\,{}^t\mathbf{F}$ のように書いて、上式を以下のようにする。

$$\mathbf{F}^{-1}\mathbf{B}\,{}^t\mathbf{F}^{-1}\,{}^t\mathbf{Fa}=\rho\,{}^t\mathbf{Fa}$$

ここで $\mathbf{A}=\mathbf{F}^{-1}\mathbf{B}\,{}^t\mathbf{F}^{-1}$ ，$\mathbf{u}={}^t\mathbf{Fa}$ （$\mathbf{a}={}^t\mathbf{F}^{-1}\mathbf{u}$）とすると、上式は以下のような対称行列の固有方程式となる。

$$\mathbf{Au}=\rho\mathbf{u}\qquad(6)$$

${}^t\mathbf{uu}=1$ の規格化条件を付けてr番目の固有値$\rho^{(r)}$について方程式を解いた答えを、$\mathbf{u}^{(r)}$とすると、正準判別関数の係数は以下で与えられる。

$$\mathbf{a}^{(r)}={}^t\mathbf{F}^{-1}\mathbf{u}^{(r)}$$

以上より、第r番目の固有値に対応する判別関数$z^{(r)}$は以下のようになる。

$$z^{(r)}={}^t\mathbf{xa}^{(r)}-{}^t\tilde{\mathbf{x}}^\alpha\mathbf{a}^{(r)}\qquad(7)$$

ここに$\tilde{\mathbf{x}}^\alpha=\frac{1}{m}\sum_{\alpha=1}^{m}\overline{\mathbf{x}}^\alpha$ である。定数項については、後に述べる２群の場合のマハラノビス形式と正準形式の同一性から、各固有ベクトルに対応する判別関数の群別平均の単純平均が０になるように決めた。

マハラノビス形式と同様、変数x_iの標準化値u_iと不偏分散s_iを用いて判別関数は以下のように書くこともできる。

$$z^{(r)}={}^t\mathbf{uc}^{(r)}+{}^t\overline{\mathbf{x}}\mathbf{a}^{(r)}-{}^t\tilde{\mathbf{x}}\mathbf{a}^{(r)},\qquad c_i^{(r)}=a_i^{(r)}s_i\qquad(8)$$

この係数$\mathbf{c}^{(r)}$を標準化係数と呼ぶ。

(6) 式から、

$$\rho^{(r)} = {}^t\mathbf{u}^{(r)}\mathbf{A}\mathbf{u}^{(r)} = {}^t\mathbf{u}^{(r)}\mathbf{F}^{-1}\mathbf{B}\,{}^t\mathbf{F}^{-1}\mathbf{u}^{(r)} = {}^t\mathbf{a}^{(r)}\mathbf{B}\mathbf{a}^{(r)} = s_B^{(r)2}$$

となり、r番目の固有値は群間分散の第r成分に等しくなる。この性質を用いて、r番目の固有値に対する変動の寄与率$P^{(r)}$を以下で与える。

$$P^{(r)} = \rho^{(r)} \Big/ \sum_{k=1}^{m-1} \rho^{(k)}$$

3) 2群におけるマハラノビスの形式と正準形式の同等性

さて、ここで述べてきた従来の理論とマハラノビスの距離を用いた判別分析とはどのような関係にあるのだろうか。(5)式について再考する。ここに方程式を再度挙げておく。

$$\mathbf{B}\mathbf{a} = \rho\mathbf{S}\mathbf{a}$$

行列$\mathbf{B}$は成分を用いて書くと以下のように表される。

$$\begin{aligned} b_{ij} &= \frac{1}{n-m}\sum_{\alpha=1}^{m} n_\alpha \left(\bar{x}_i^\alpha - \bar{x}_i\right)\left(\bar{x}_j^\alpha - \bar{x}_j\right) \\ &= \frac{1}{n(n-m)}\sum_{\alpha=1}^{m}\sum_{\beta=1}^{m} n_\alpha n_\beta \left(\bar{x}_i^\alpha \bar{x}_j^\alpha - \bar{x}_i^\alpha \bar{x}_j^\beta\right) \\ &= \frac{1}{2n(n-m)}\sum_{\alpha=1}^{m}\sum_{\beta=1}^{m} n_\alpha n_\beta \left(\bar{x}_i^\alpha - \bar{x}_i^\beta\right)\left(\bar{x}_j^\alpha - \bar{x}_j^\beta\right) \end{aligned}$$

これより、$(\mathbf{S}_B\mathbf{a})_{ij}$は以下のように書ける。

$$\begin{aligned} (\mathbf{S}_B\mathbf{a})_i &= \frac{1}{2n(n-m)}\sum_{\alpha=1}^{m}\sum_{\beta=1}^{m}\sum_{j=1}^{p} n_\alpha n_\beta \left(\bar{x}_i^\alpha - \bar{x}_i^\beta\right)\left(\bar{x}_j^\alpha - \bar{x}_j^\beta\right) a_j \\ &= \sum_{\alpha=1}^{m}\sum_{\beta=1}^{m} c_{\alpha\beta}\left(\bar{x}_i^\alpha - \bar{x}_i^\beta\right) \\ c_{\alpha\beta} &= \frac{n_\alpha n_\beta}{2n(n-m)}\sum_{j=1}^{p}\left(\bar{x}_j^\alpha - \bar{x}_j^\beta\right) a_j \end{aligned}$$

特に2群の判別の場合、方程式(5)は以下となる。

$$\rho\mathbf{S}\mathbf{a} = \mathbf{S}_B\mathbf{a} = c(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)$$

$$c = 2c_{12} = -2c_{21} = \frac{n_1 n_2}{n(n-2)}\sum_{j=1}^{p}\left(\bar{x}_j^1 - \bar{x}_j^2\right) a_j$$

ここにcはある定数になる。これより、$\mathbf{a}$の比率だけに着目すると以下を得る。

$$\mathbf{a} = \frac{c}{\rho}\mathbf{S}^{-1}(\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)$$

これは、(1)式で与えられたマハラノビス形式の判別関数の係数の定数倍である。よって、判別の分点を0にするような判別関数は以下となる。

$$z=\frac{c}{\rho}\,{}^{t}\mathbf{x}\mathbf{S}^{-1}(\bar{\mathbf{x}}^{1}-\bar{\mathbf{x}}^{2})-\frac{c}{2\rho}\,{}^{t}(\bar{\mathbf{x}}^{1}+\bar{\mathbf{x}}^{2})\mathbf{S}^{-1}(\bar{\mathbf{x}}^{1}-\bar{\mathbf{x}}^{2})$$

これは判別関数全体が定数倍となっただけで、判別結果は$-\log h$の項を除いて同等である。

4) Wilks のλについて

多群の判別と正準判別分析の出力結果に、群間の分離の状態を表す指標として Wilks のλ統計量及びそれから求めた F 値による分離の可能性の検定確率を加えた。また変数の重要性として偏λ統計量とそれから求めた F 値による検定確率を加えた。

λ統計量は群の分離の状態を 0 から 1 の間の値で与える指標で、0 に近いほど分離が良いとされる。λ統計量は全体の平方積和行列Tと群内の平方積和行列Wの行列式の比として以下のように定義される。

$$\lambda=|\mathrm{W}|/|T|$$

ここで、

$$(\mathbf{T})_{jk}=\sum_{l=1}^{g}\sum_{i=1}^{n_l}(x_{ij}^{(l)}-\bar{x}_j)(x_{ik}^{(l)}-\bar{x}_k)\quad,\quad(\mathbf{W})_{jk}=\sum_{l=1}^{g}\sum_{i=1}^{n_l}(x_{ij}^{(l)}-\bar{x}_j^{(l)})(x_{ik}^{(l)}-\bar{x}_k^{(l)})$$

であり、検定のためのF統計量は、nを全データ数、gを群の数、qを変数の数として、近似的に以下のように与えられる。

$$F=\frac{f_2}{f_1}\frac{1-\lambda^{1/c}}{\lambda^{1/c}}\sim F_{qa,bc+1-qa/b},\qquad a=g-1,b=n-1-(q+g)/2$$

$$c=\begin{cases}\left\{(q^2a^2-4)/(q^2+a^2-5)\right\}^{1/2} & for\ \ q^2+a^2\neq 5\\ 1 & for\ \ q^2+a^2=5\end{cases}$$

次に変数の重要性を見るための偏λ統計量の定義を示す。ある変数aを除いて求めたWilksλ統計量をλ_aとする。これと全体のλ統計量との比$\lambda_a^*=\lambda/\lambda_a$を変数$a$の偏λ統計量という。その検定のための$F$統計量$F_a$を以下に示す。

$$F_a=\frac{n-g-q}{g-1}\frac{1-\lambda_a^*}{\lambda_a^*}\sim F_{g-1,n-g-q}$$

これらにより、分析の有効性や変数の重要性が議論できる。

参考文献

[1] 田中豊・垂水共之編, Windows 版 統計解析ハンドブック 多変量解析, 共立出版社, 1995.

４．主成分分析

4.1　主成分分析とは

主成分分析の目的は、複数の変数を１次関数として組み合わせて、いくつかの特徴的な量を作り出すことである。例えば、身長、体重、胸囲、座高のデータを組み合わせると人間の体格に関する特徴的な量を作り出すことができるのではないかと考えられる。この特徴的な量は主成分 1、主成分 2 などと呼ばれ、以下のように作られる。

主成分 1＝a_{11} 身長＋a_{12} 体重＋a_{13} 胸囲＋a_{14} 座高

主成分 2＝a_{21} 身長＋a_{22} 体重＋a_{23} 胸囲＋a_{24} 座高

：　　：

これらを構成する変数の係数は、固有ベクトルと呼ばれるものの値で与えられる。各主成分の解釈は、係数の値と符号を見て利用者が判断する。この他に、主成分と変数との相関係数を因子負荷量と呼び、固有ベクトルの代わりに主成分の意味の解釈に用いられることがある。因子負荷量は固有ベクトルの値に比例する。

各主成分のばらつきである分散は、固有値と呼ばれる値で与えられる。また、各主成分の重要性を表す寄与率は、固有値の合計に対する各固有値の割合で与えられる。これは変動の何割を各主成分が受け持つかを表す量である。主成分分析は、図 1 のように、データの変動が大きくなる方向へ主軸を移動・回転させて向けることで、主成分を決めている。この背景には変動の大きさの中には何らかの特徴が隠されているという考えがある。

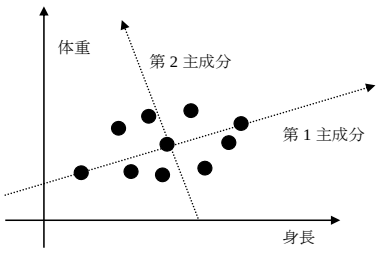


図 1　主成分の決定法

多次元の場合は、一番大きな変動の方向へ最初の軸を向け、次の大きな変動の方向へ最初の軸と直交させながら 2 番目の軸を向ける。これを続けて軸の向きを決めて行く。累積寄与率と呼ばれる寄与率の合計が 9 割を超えると、十分とする見方もある。また、大きな変動が区別できるところまでが、意味がある主成分の個数であるとする見方もある。これを調べるのが等固有値の検定である。

このように、いくつかの変数の性質をそれより少ない主成分でまとめることが主成分分析の目的である。主成分分析は、データそのものを使って共分散行列を作り、その行列から計算する場合と、変数について標準化の処理を行い、それを使って相関行列を作り、その行列から計算する場合がある。変数間の単位が異なったり、標準偏差の大きさが大きく違っていたりする場合などは、相関行列からの実行が一般的である。

主成分得点の値は、共分散行列を使う場合は固有ベクトルにデータの値を掛けて、相関行

列を使う場合は固有ベクトルに標準化したデータの値を掛けて計算される。主成分得点の分散は、固有ベクトルに一致するようになっているので、分散が 1 になるように標準化されているわけではない。（因子分析の場合は分散が 1 になっている。）

以下の例を考えて、まとめてみよう。

例

以下の健康診断のデータ（Samples¥主成分分析 1.txt）から、変数の 1 次関数として体格を表す特徴的な指標を作り、その意味を考察せよ。

身長	体重	胸囲	座高
148	41	72	78
160	49	77	86
159	45	80	86
153	43	76	83
:	:	:	:
148	38	70	78

まとめ

変数に身長、体重、胸囲、座高の 4 つをとって主成分分析を行なった。各変数の値に大きな差がないことから、ここでは共分散行列を基にした方法を用いている。変数は正規分布するものとみなされ、等固有値の検定も利用可能である。

第 1 主成分は 1 次式の係数の値（固有ベクトルの値）がすべて正であることから身体の大きさを表す変数であると考える。また、第 2 主成分は身長・座高と体重・胸囲で符号が違うことから、肥満の程度を表す変数であると考える。

これらの主成分の寄与率をみると、第 1 主成分が 0.891、第 2 主成分が 0.077 であり、他はすべて 0.02 以下になっている。また等固有値の検定より、第 1 主成分と第 2 主成分が利用可能であることが分かる。第 3 主成分以降については意味付けが困難であり、利用しない。最後に結果を式で表しておく。

身体の大きさを表す主成分

第 1 主成分＝0.6240 身長＋0.5592 体重＋0.4083 胸囲＋0.3622 座高

肥満の程度を表す主成分

第 2 主成分＝－0.6456 身長＋0.3456 体重＋0.6605 胸囲－0.1660 座高

4.2　プログラムの利用法

実際の主成分分析のメニュー画面を図 1 に与える。主成分分析は、表 1 に与えたデータの形から実行する場合に加え、それを集計した共分散行列や相関行列から実行する場合も想定される。それ故データの形式としてこれら 3 つの場合が含まれている。等固有値の検定にはデータ数も必要になることから、集計結果からの計算ではデータ数を入力する必要もある。計算を実行するモデルには、通常のデータから計算する「共分散行列から」と標準化されたデータから計算する「相関行列から」の 2 種類がある。勿論、データ形式で相関行列を選んだ場合は共分散行列からの計算はできない。

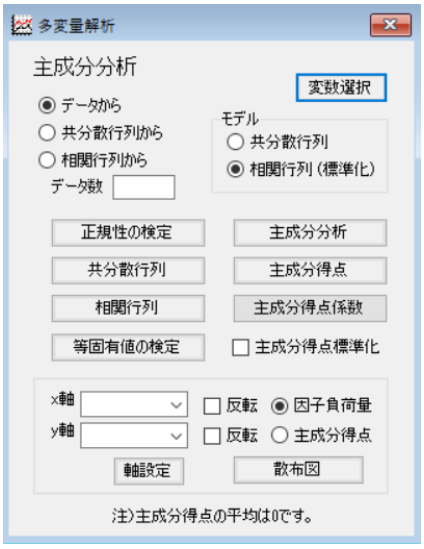


図１　主成分分析のメニュー

計算結果の表示としては「共分散行列」や「相関行列」も必要と思われるので加えてある。主成分分析は「主成分分析」ボタンで実行され、出力例は、図２に示される。

主成分分析

	主成分1	主成分2	主成分3	主成分4
固有値	3.541	0.313	0.079	0.066
寄与率	0.885	0.078	0.020	0.017
累積寄与率	0.885	0.964	0.983	1.000
固有ベクトル				
身長	0.497	−0.543	−0.450	−0.506
体重	0.515	0.210	−0.462	0.691
胸囲	0.481	0.725	0.175	−0.461
座高	0.507	−0.368	0.744	0.232
因子負荷量				
身長	0.935	−0.304	−0.127	−0.130
体重	0.968	0.118	−0.130	0.178
胸囲	0.905	0.406	0.049	−0.119
座高	0.954	−0.206	0.210	0.060

図２　主成分分析出力結果

等固有値の検定結果は図３に示される。

等固有値の検定 (by Anderson)

利用主成分	第1主成分	第2主成分	第3主成分	第4主成分
χ2値	67.0395	10.1275	0.1093	
自由度	9	5	2	
等固有値確率	0.0000	0.0717	0.9468	
利用可能性	可	不可	不可	不可

図３　等固有値の検定結果

ここに表示された第i主成分のχ^2値は、固有値を大きさの順に並べた場合、第i主成分以降の固有値がすべて等しいとみなせるかどうかの検定値であり、等固有値確率はその確率値を表わす。それゆえ等固有確率が有意水準より大きい主成分以降が利用に適さないことを示している。極端な例として、第 1 主成分の等固有値確率が有意水準より小さい場合、主成分分析自体があまり意味を持たない。

「主成分得点」の出力は各主成分毎に図 4 に与えられ、2 つの主成分に関する主成分得点の散布図は図 5 に与えられる。これによって主成分で見た場合の個体の類似度を把握することが容易となる。

主成分得点

	主成分1	主成分2	主成分3	主成分4
1	-0.069	0.234	-0.349	0.262
2	2.800	-0.383	-0.096	0.275
3	2.694	-0.017	0.354	-0.353
4	1.397	0.060	0.207	0.043
5	0.919	0.575	-0.087	-0.178
6	-2.790	-0.343	0.033	0.031
7	2.401	0.165	-0.461	0.160
8	-2.766	0.313	-0.032	0.218
9	1.529	1.676	-0.326	-0.007
10	2.479	-0.956	0.120	0.384

図 4　主成分得点出力結果

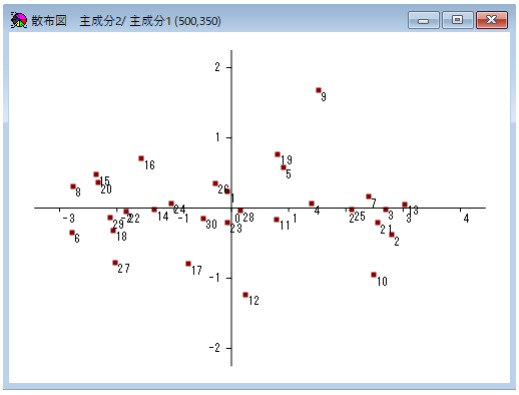


図 5　主成分得点散布図

デフォルトでは、主成分得点の各主成分の分散は図 2 の固有値の値になっているが、「主成分得点標準化」チェックボックスにチェックを入れることによって、因子得点のように標準化することもできる。

「主成分得点係数」ボタンをクリックすると、図 6 のような表示が得られるが、これは相関行列からの場合、標準化データから主成分得点を求める際の係数行列である。

主成分得点係数

	主成分1	主成分2	主成分3	主成分4
身長	0.497	-0.543	-0.450	-0.506
体重	0.515	0.210	-0.462	0.691
胸囲	0.481	0.725	0.175	-0.461
座高	0.507	-0.368	0.744	0.232

図 6　主成分得点係数行列（相関行列モデル）

主成分得点を標準化しない場合は、この値は固有ベクトルの値と同じである。また、共分散行列からのモデルでは、標準化されていない元のデータから主成分得点を求める係数となるが、この場合図 7 のように平行移動に相当する定数項が現れる。

主成分得点係数

	主成分1	主成分2	主成分3	主成分4
身長	0.624	-0.646	0.224	-0.379
体重	0.559	0.346	-0.746	-0.108
胸囲	0.408	0.660	0.624	-0.084
座高	0.362	-0.166	0.062	0.915
定数項	-172.859	48.281	-54.495	-5.863

図 7　主成分得点係数行列（共分散行列モデル）

問題（主成分分析 2.txt）

主成分分析 2.txt は生徒の教科別の成績データである。相関行列をもとにするモデルを用いて以下の問いに答えよ。

１）各主成分の固有値（分散の値)、寄与率、累積寄与率を求めよ。

	第 1 主成分	第 2 主成分	第 3 主成分	第 4 主成分	第 5 主成分
固有値					
寄与率					
累積寄与率					

２）主成分を２つ使うとすると、第 1 主成分と第 2 主成分の関数はどのように表されるか。

第 1 主成分＝［　　］英語＋［　　］数学
　　＋［　　］国語＋［　　］理科＋［　　］社会

第 2 主成分＝［　　］英語＋［　　］数学
　　＋［　　］国語＋［　　］理科＋［　　］社会

３）これら２つの主成分で説明できるのは全体の変動の何%か。［　　］%

４）これら２つの主成分はどのように意味づけられるか。

第 1 主成分　意味［　　］を表す指標

第 2 主成分　意味［　　］を表す指標

５）先頭（1 番）の生徒の２つの主成分得点を求めよ。

第 1 主成分得点［　　］　第 2 主成分得点［　　］

６）２つの主成分の意味を考えて、この生徒にはどんな特徴があるか。

［　　］

問題解答（主成分分析 2.txt）

１）各主成分の固有値（分散の値)、寄与率、累積寄与率を求めよ。

	第 1 主成分	第 2 主成分	第 3 主成分	第 4 主成分	第 5 主成分
固有値	4.200	0.427	0.229	0.112	0.032
寄与率	0.840	0.085	0.046	0.023	0.006
累積寄与率	0.840	0.925	0.971	0.994	1.000

２）主成分を２つ使うとすると、第 1 主成分と第 2 主成分の関数はどのように表されるか。

第 1 主成分＝［ 0.481 ］英語＋［ 0.460 ］数学
　　＋［ 0.459 ］国語＋［ 0.412 ］理科＋［ 0.420 ］社会

第 2 主成分＝［ 0.040 ］英語＋［ 0.098 ］数学
　　＋［ -0.221 ］国語＋［ 0.735 ］理科＋［ -0.633 ］社会

３）これら２つの主成分で説明できるのは全体の変動の何%か。［ 92.5 ］%

４）これら２つの主成分はどのように意味づけられるか。

第 1 主成分　意味［ 総合的な学力 ］を表す指標

第 2 主成分　意味［ 文系か理系か ］を表す指標

５）先頭（1 番）の生徒の２つの主成分得点を求めよ。

第 1 主成分得点［ -1.78 ］　第 2 主成分得点［ 1.60 ］

６） ２つの主成分の意味を考えて、この生徒にはどんな特徴があるか。
［ 学力は低く、理系 ］

4.3 主成分分析の理論

主成分分析は、変数の 1 次結合により、新しい意味付けのできる特徴的な変数を作り出すことを目的としている。この新しい変数を主成分と呼ぶ。主成分分析のデータ形式は表１で与えられる。

表１　主成分分析のデータ

変数１	変数２	…	変数 p
x_{11}	x_{21}	…	x_{p1}
x_{12}	x_{22}	…	x_{p2}
:	:	…	:
x_{1n}	x_{2n}	…	x_{pn}

我々は新しい変数として以下の１次式を考える。

$$y_\lambda = \sum_{i=1}^{p} u_i x_{i\lambda}$$

特徴的な変数とは、データの変化に最も敏感であることと考え、係数u_iは変数yの不偏分散s^2が最大になるように求める。但し、スケールの自由度を無くすため係数に${}^t\mathbf{uu}=1$の制約を付ける。ここに$\mathbf{u}$は成分がu_iの縦ベクトルである。

不偏分散s^2は係数ベクトル$\mathbf{u}$と共分散行列$\mathbf{S}$を用いて以下のように与えられる。

$$s^2 = \frac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda - \bar{y})^2 = {}^t\mathbf{uSu}, \qquad (\mathbf{S})_{ij} = \frac{1}{n-1}\sum_{\lambda=1}^{n}(x_{i\lambda}-\bar{x}_i)(x_{j\lambda}-\bar{x}_j)$$

この制約付き最大化問題は、Lagrange の未定定数法を用いて以下の量Lの極値問題となり、解は行列$\mathbf{S}$の固有方程式で与えられる。

$$L = {}^t\mathbf{uSu} - \lambda({}^t\mathbf{uu}-1) \quad \rightarrow \quad \mathbf{Su} = \lambda\mathbf{u}$$

この最大固有値に対する固有ベクトル$\mathbf{u}$を用いて作られた変数yを第 1 主成分といい、順次固有値の大きい方から第 2 主成分、第 3 主成分と呼ぶ。一般に p 変数の場合、第 p 主成分まで選ぶことができる。

係数u_iは変数の平均や分散から影響を受けるので、変数を標準化して分析を実行する場合も多い。この場合固有方程式は相関行列$\mathbf{R}$を用いて上と同様に与えられる。

$$\mathbf{Ru} = \lambda\mathbf{u}$$

正規化された固有ベクトルを求めることは、線形変換における座標回転の角度を決めることを意味する。即ち、主成分分析は、座標回転によって最も分散の大きな主軸を選び、さらにその主軸に直交し、分散が最大になるような軸を次々と定めてゆく方法である。

これらの固有方程式の第α固有値λ_αに対する固有ベクトル$\mathbf{u}^\alpha$の成分を以下のように表わす。

$${}^t\mathbf{u}^{\alpha} = (u_1^{\alpha} \quad u_2^{\alpha} \quad \cdots \quad u_p^{\alpha})$$

固有値λ_{α}は第α主成分の分散を表わすことが知られている。このことから、全分散s^2に対する第α主成分の分散の割合c_{α}は以下で与えられ、寄与率と呼ばれる。

$$c_{\alpha} = \lambda_{\alpha} \Big/ \sum_{i=1}^{p} \lambda_i$$

因子負荷量$r_{\alpha i}$は第α主成分と変数iの相関係数として与えられるが、これは共分散行列と相関行列を元にした場合に分けて、それぞれ以下のような形に表わされる。

$$r_{\alpha i} = \frac{\sqrt{\lambda_{\alpha}}\, u_i^{\alpha}}{s_i} \quad \text{（共分散行列から），} \qquad r_{\alpha i} = \sqrt{\lambda_{\alpha}}\, u_i^{\alpha} \quad \text{（相関行列から）}$$

ここでs_i^2は変数iの不偏分散である。

主成分得点y_{λ}^{α}は個体毎の第α主成分の値として以下のように定義される。

$$y_{\lambda}^{\alpha} = \sum_{i=1}^{p} u_i^{\alpha} x_{i\lambda}$$

主成分分析において主成分を区別するためには、その固有値の大きさに差がなければならない。そこで固有値を$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$とした場合、大きいほうから$r$個だけ値が異なり、残りは$\lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_p$となるかどうかのAndersonによるsphericityの検定を行なう。この検定には以下の性質が利用される。

$$\chi^2 = -n \sum_{\alpha=r+1}^{p} \log \lambda_{\alpha} + n(p-r) \log \left(\sum_{\alpha=r+1}^{p} \lambda_{\alpha} \Big/ (p-r) \right) \sim \chi^2_{(p-r-1)(p-r+2)/2}$$

５．因子分析

5.1　因子分析とは

データの中には、変数の背後に全体に共通するある因子があり、その因子の発現状態によって変数の性質が決定されると考えられるようなものがある。因子分析は各変数の背後にあるその共通因子を求め、それらの１次関数として各変数が表されるように係数を求め、変数への因子の影響の向きや強さを考える手法である。

例えば、身長、体重、胸囲、座高の変数の因子分析での考え方を因子数 2 の因子分析の表式で表してみよう。ここに因子分析では、データはすべて標準化されているものとする。

身長＝b_{11} 因子 1＋b_{12} 因子 2＋誤差

体重＝b_{21} 因子 1＋b_{22} 因子 2＋誤差

胸囲＝b_{31} 因子 1＋b_{32} 因子 2＋誤差

座高＝b_{41} 因子 1＋b_{42} 因子 2＋誤差

これは主成分分析と逆の関係である。

各因子の係数値は因子負荷量と呼ばれ、因子と変数との相関係数である。因子軸については直交しており、因子間の相関係数は 0 である。

因子軸と変数軸との間の関係を図で表すと図 1 のようになる。因子軸は変数の変動の大きな方向に回転させている。

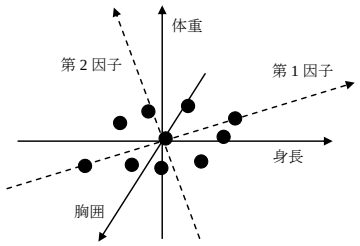


図１　因子軸と変数軸の関係

各因子の重要性は、各因子の寄与率で与えられる。因子数の目安は、相関行列の固有値で 1 より大きい固有値の数とみる見方もあるが、累積寄与率が 90%までとする見方もある。これらはあくまで目安であり、現実の分析ではこのようにうまく行かないことの方が多い。

求めた因子が各変数の値を説明する割合は、共通性という指標で与えられる。また、データごとの因子の値は因子得点と呼ばれる。

以下の例を用いて因子分析のまとめを作ってみる。

例

以下の健康診断のデータ（因子分析 1.txt）から、変数の背後にある体格を表す共通因子を求め、その意味を考察せよ。

身長	体重	胸囲	座高
148	41	72	78
160	49	77	86

153	43	76	83
151	42	77	80
:	:	:	:
151	36	74	80
141	30	67	76
148	38	70	78

まとめ

変数に身長、体重、胸囲、座高の４つをとって因子分析を行った。累積寄与率が 0.9 になるように因子数を 2 とし、因子負荷量推定法には主成分分析を用いて計算を実行した。

バリマックス回転の実施後､各変数は２つの因子と因子負荷量を用いて以下のように予測される。

身長＝0.898×因子１+0.402×因子２

体重＝0.639×因子１+0.737×因子２

胸囲＝0.399×因子１+0.908×因子２

座高＝0.846×因子１+0.487×因子２

第 1 因子は身長と座高の因子負荷量が大きいため、体の縦方向の大きさを代表する因子、第 2 因子は体重と胸囲の因子負荷量が大きいため、体の横方向の大きさを代表する因子と考えられる。

各変数を予測するこれらの因子の寄与率は第１因子が 0.522、第２因子が 0.441 で、２つの因子の累積寄与率は 0.964 である。

因子分析には因子の決め方により回転の自由度が残されており、この自由度を使って、より解釈し易い因子を構成し直すことができるという利点がある。これがバリマックス回転やプロマックス回転などと呼ばれる変換である。バリマックス回転は軸を直角に保ったままの回転であり、プロマックス回転は軸の直交性は考えず、より分類がはっきりとする方向へ軸を向ける。プロマックス回転は斜交回転と呼ばれ､他にもいくつかの方法が考えられている。

バリマックス回転やプロマックス回転を行うと､因子負荷量の値の大きなところだけに注目して因子の解釈ができるため、因子の意味が考え易くなる。もちろんこれらの回転を行っても因子分析の精度としてはすべて同じである。

5.2　プログラムの利用法

メニュー［分析－多変量解析他－分類手法－因子分析］を選択すると、因子分析の実行画面が図 1 のように表示される。データとしては主成分分析と同じように個体毎のデータ、共分散行列､相関行列が選択できる。因子負荷量を求める方法では､主因子法､主成分分析、最小２乗法、最尤法が利用できる。

図１　因子分析実行画面

図２に因子数を２としてバリマックス回転にチェックを入れ、「因子分析」のボタンをクリックした場合の出力結果を示す（因子分析 4.txt）。

因子負荷量

	因子1	因子2	共通性
国語	0.075	0.787	0.625
英語	0.221	0.737	0.591
社会	0.157	0.763	0.606
数学Ⅰ	0.851	0.203	0.766
数学Ⅱ	0.845	0.296	0.802
理科	0.889	0.040	0.792
寄与率	0.385	0.312	
累積寄与率	0.385	0.697	
符号調整済α	0.836	0.673	

図２　因子分析出力結果

結果には因子数で指定した数だけ因子負荷量と寄与率､累積寄与率が表示されている｡但し､主因子法では、固有値が 0 に近いところで負の値を取る場合も見つかっており、指定した個数より少なく表示されることもある｡符号調整済みαは､因子負荷量の符号が同じになるように､変数の符号を調整して因子負荷量の大きさで組み分けした場合の Cronbach のα係数である。これは、一般には 0.8 程度以上が良いとされている。

「因子得点」ボタンをクリックすると図３のように個体毎の因子得点が表示される。ここでは因子得点の推定に、回帰推定法を用いている。「散布図」ボタンをクリックすると図 4 のように因子得点１を横軸に因子得点２を縦軸にした散布図を作成する。

因子得点

	因子1	因子2
1	1.411	0.567
2	-1.339	-0.491
3	1.193	0.803
4	0.073	-0.385
5	0.787	-0.299
6	1.243	-1.697
7	-1.685	-0.981
8	0.439	-1.921
9	-1.844	-0.224

図３　因子得点出力画面

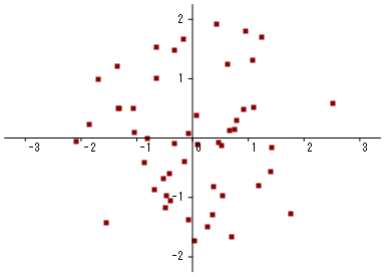


図４　因子得点散布図

「因子得点係数」ボタンをクリックすると、因子得点を求めるための係数が、図 5 のよ

うに表示される。実データから求める場合と標準化されたデータ（不偏分散による）から求める場合の２種類の係数が示されている。


因子得点の係数

	国語	英語	社会	数学Ⅰ	数学Ⅱ	理科	定数項
▶ 因子1(標準化データ)	-0.13276	-0.04763	-0.08572	0.38865	0.36543	0.44342	
因子2(標準化データ)	-0.47650	-0.41340	-0.44341	0.05726	-0.00231	0.16796	
因子1(実データ)	-0.00944	-0.00278	-0.00586	0.02523	0.02669	0.02156	-3.07657
因子2(実データ)	-0.03388	-0.02416	-0.03033	0.00372	-0.00017	0.00817	4.87724

図５　因子得点を求める場合の係数

「予測整合性」というボタンは、因子得点を計算して、逆に元のデータを予測し、実データと比較して、因子分析の効果を実感してもらうためのものである。その分析結果を図 6 に示す。


予測整合性（標準化変数値と予測値）

	国語	予測値	英語	予測値	社会	予測値	数学Ⅰ	予測値	数学
45	0.492	0.046	0.394	-0.006	-0.709	0.017	-0.491	-0.259	
46	0.919	0.095	-1.009	0.101	0.386	0.099	0.288	0.092	
47	0.492	0.446	-0.892	0.355	1.207	0.397	-0.621	-0.230	
48	-0.219	-0.834	-1.652	-0.877	-0.845	-0.862	-0.815	-0.749	
49	-0.717	-0.469	-0.191	-0.600	-0.709	-0.545	-1.010	-1.008	
50	-1.286	-1.348	0.219	-1.117	-2.350	-1.225	0.223	0.447	
平均値	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
標準偏差	1.000	0.782	1.000	0.761	1.000	0.771	1.000	0.866	
相関係数		0.782		0.761		0.771		0.866	

図６　実測値と予測値の比較結果

因子負荷量推定法で主成分分析を選ぶと、累積寄与率の値が主成分分析の同じ主成分数の場合と等しくなる。また、他の推定法に比べても累積寄与率の値は向上する。結果の出力変数の並びは変数選択順の他に、因子負荷量の大きさで２通りに並べ替える方法が利用できる。特にグループ化負荷量順では因子ごとに因子負荷量の大きい変数同士を並べて表示でき、因子の解釈がより容易になる。

因子負荷量推定法には、最小２乗法と最尤法もあるが、これらは共分散構造分析の手法を因子分析に利用したものである。最尤法は変数が多くなると計算に時間を要する。

図２の因子分析の結果の中に符号調整済 α という欄がある。ここに書かれた数字はクロンバッハ（Cronbach）の α 係数と呼ばれ、各因子を因子負荷量の大きい変数だけで表す（命名する）妥当性を示す指標である。通常 α≧0.8 の値が良いとされる。最近はマクドナルド（Mcdonald）の ω 係数の方が信頼性をよく表すという見方もあり、これまでのボタンを変更して「信頼性分析」のボタンを作り、ω 係数の値も図７のように表示することにした。


共分散/信頼性係数

	理科	数学Ⅰ	数学Ⅱ	
▶ 理科	422.864	197.320	190.251	
数学Ⅰ	197.320	237.231	146.607	
数学Ⅱ	190.251	146.607	187.462	
	無変換	符号調整	標準化	符号標準化
α係数	0.836	0.836	0.856	0.856
ω係数(最尤法ωh)	0.844	0.844	0.857	0.857

図７　信頼性分析実行結果

参考文献

[1] 田中豊・垂水共之編，Windows 版統計解析ハンドブック多変量解析，共立出版，1995.

問題（因子分析 3.txt）

Samples¥因子分析 3.txt は北海道各地の 2 月の気温データである。設定はデフォルト（バリマックス回転）として以下の問いに答えよ。注）江差町（えさし：南部），寿都町（すっつ：南部），小樽市（おたる：中部），留萌市（るもい：北部），天塩町（てしお：北部）

１）各都市間の相関行列の固有値を大きい順に３つ求めよ。

1	2	3

以後因子数を２つと決めて各質問に答えよ。

２）各因子の寄与率と累積寄与率を求めよ。

	第 1 因子	第 2 因子
寄与率		
累積寄与率		

３）因子数は２つでよいか。［よい・注意が必要］

４）各因子の因子負荷量を求めよ。

	江差	寿都	小樽	留萌	天塩
第 1 因子					
第 2 因子					

５）上の因子負荷量の値から各因子の意味を解釈せよ。

第 1 因子：[　　]の気温を代表する因子

第 2 因子：[　　]の気温を代表する因子

６）各地の気温の変動は因子によりどの程度説明されるか（共通性）。

江差	寿都	小樽	留萌	天塩

７）最初の 3 日間の因子の値（因子得点）を推定せよ。

	第 1 因子	第 2 因子
1		
2		
3		

８）この 3 日間、北海道はどのような気候だったか。

[　　]

９）このモデルは良いモデルと思うか。

［良いと思う・あまり良いと思わない］

演習（多変量演習 8.txt）

多変量演習 8.txt のデータはある学校で測定した小学６年生の運動適性テストの結果である。因子分析（バリマックス回転）を用いて特徴を分析し、以下の問いに答えよ。

１）各科目間の相関行列の固有値を大きい順に求めよ。

1	2	3	4	5

２）因子数を２つとして、因子分析を行い、寄与率を求めよ。

因子 1	因子 2

３）因子数を３つとして、因子分析を行い、寄与率を求めよ。

因子 1	因子 2	因子 3

本来なら因子数 3 が良いが、解釈の問題から、以後因子数を 2 と決めて各質問に答えよ。

４）各因子の因子負荷量を求めよ。

	立幅跳び	腹筋	腕立伏せ	往復走	5 分間走
第 1 因子					
第 2 因子					

５）この場合の各因子の意味を解釈せよ。

第 1 因子：[　　]を表す因子

第 2 因子：[　　]を表す因子

６）先頭から 3 人及び、特徴的な 9 番目の人の因子の値（因子得点）を推定せよ。

	第 1 因子	第 2 因子
1		
2		
3		
9		

７）9 番目の人にはどんな特徴があるか。因子負荷量の符号に注意して考えよ。

[　　]

８）各種目の変動は因子によりどの程度説明されるか。

立幅跳び	腹筋	腕立伏せ	往復走	5 分間走

９）予測値が２つの因子から予測されたことを考えると、この分析はうまくいったと思うか。

［まずまずうまくいった・うまくいっていない］

問題解答（因子分析 3.txt）

１）各都市間の相関行列の固有値を大きい順に３つ求めよ。

1	2	3
3.931	0.793	0.149

２）各因子の寄与率と累積寄与率を求めよ。

	第１因子	第２因子
寄与率	0.473	0.472
累積寄与率	0.473	0.945

３）因子数は２つでよいか。［よい・注意が必要］

４）各因子の因子負荷量を求めよ。

	江差	寿都	小樽	留萌	天塩
第１因子	0.948	0.906	0.675	0.423	0.185
第２因子	0.235	0.363	0.730	0.853	0.955

５）上の因子負荷量の値から各因子の意味を解釈せよ。

第１因子：［ 北海道南部 ］の気温を代表する因子

第２因子：［ 北海道北部 ］の気温を代表する因子

６）各地の気温の変動は因子によりどの程度説明されるか（共通性）。

江差	寿都	小樽	留萌	天塩
0.954	0.953	0.965	0.906	0.946

７）最初の３日間の因子の値（因子得点）を推定せよ。

	第１因子	第２因子
1	−0.870	0.061
2	−0.086	0.321
3	−0.295	0.778

８）この３日間、北海道はどのような気候だったか。

［(それぞれの平均に比べて）南部は少し寒く、北部は少し暖かい。］

９）このモデルは良いモデルと思うか。

［良いと思う・あまり良いと思わない］

演習解答（多変量演習 8.txt）

１）各科目間の相関行列の固有値を大きい順に求めよ。

1	2	3	4	5
3.248	0.677	0.584	0.335	0.156

２）因子数を２つとして、因子分析を行い、寄与率を求めよ。

因子１	因子２
0.559	0.226

３）因子数を３つとして、因子分析を行い、寄与率を求めよ。

因子１	因子２	因子３
0.443	0.247	0.213

４）各因子の因子負荷量を求めよ。

	立幅跳び	腹筋	腕立伏せ	往復走	５分間走
第１因子	0.907	0.813	0.712	0.858	0.258
第２因子	0.257	0.209	0.231	0.193	0.965

５）この場合の各因子の意味を解釈せよ。

第１因子：［ 瞬発力 ］を表す因子

第２因子：［ 持久力 ］を表す因子

６）先頭から３人及び、特徴的な９番目の人の因子の値（因子得点）を推定せよ。

	第１因子	第２因子
1	−0.682	−0.247
2	−1.076	0.397
3	−0.302	−3.233
9	1.119	1.854

７）９番目の人にはどんな特徴があるか。因子負荷量の符号に注意して考えよ。
［ 瞬発力も持久力も優れている。 ］

８）各種目の変動は因子によりどの程度説明されるか。

立幅跳び	腹筋	腕立伏せ	往復走	５分間走
0.888	0.705	0.561	0.773	0.998

９）予測値が２つの因子から予測されたことを考えると、この分析はうまくいったと思うか。
［まずまずうまくいった・うまくいっていない］

5.3 因子分析の理論

因子分析が取り扱うのは主成分分析等と同様に p 変数、n 個体（レコード）のデータ $x_{i\lambda}$（$i=1,2,\cdots,p,\ \lambda=1,2,\cdots,n$）である。これらのデータから各変数 x_i に内在すると思われる因子を抽出することが因子分析のねらいである。

因子分析では変数 x_i を標準化した変数 $t_i=(x_i-\bar{x}_i)/u_i$ を用いることが多いので、今後はこの変数 t_i を用いて議論を進める。ここで $\bar{x}_i$ は変数 x_i の標本平均、u_i は不偏分散から求めた標準偏差である。

因子分析では各データに内在すると考えられる共通因子 $f_{\alpha\lambda}$（$\alpha=1,2,\cdots,q\le p$）の線形結合によって、変数 t_i が以下のように表わされるものと考える。

$$t_i=\sum_{\alpha=1}^{q}a_{i\alpha}f_\alpha+\varepsilon_i \qquad (1)$$

係数 $a_{i\alpha}$ は α 因子の因子負荷量と呼ばれている。ここで ε_i は誤差であり、共通因子 f_α との相関や互いの相関はないものと考える。

$$E(f_\alpha\varepsilon_i)=0,\quad E(\varepsilon_i\varepsilon_j)=0\ \ (i\neq j)$$

また共通因子 f_α についても互いの相関はなく、平均０、分散１に標準化されているものとする。

$$E(f_\alpha f_\beta)=\delta_{\alpha\beta},\quad E(f_\alpha)=0,\quad V(f_\alpha)=1$$

これを直交回転モデルという。一方共通因子間に相関があるモデルを斜交回転モデルといい、後にその性質を論じる。

これらを利用すると変数 x_i と x_j との相関係数 r_{ij} は以下のように表わせる。

$$r_{ij}=E(t_it_j)=\sum_{\alpha=1}^{q}a_{i\alpha}a_{j\alpha}\ \ (i\neq j),\quad r_{ii}=V(t_i)=\sum_{\alpha=1}^{q}a_{i\alpha}^2+V(\varepsilon_i)=1$$

ここで、$h_i=\sum_{\alpha=1}^{q}a_{i\alpha}^2=1-V(\varepsilon_i)$ と置くと、上式は以下のように表わされる。

$$\mathbf{A}^t\mathbf{A}=\mathbf{R},$$

$$\mathbf{A}=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p1} & \cdots & a_{pq} \end{pmatrix},\quad \mathbf{R}=\begin{pmatrix} h_1 & r_{12} & \cdots & r_{1p} \\ r_{12} & h_2 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & h_p \end{pmatrix} \qquad (2)$$

この中で特にh_iは共通性と呼ばれる。

共通性の和を取ると、

$$\sum_{i=1}^{p} h_i=\sum_{i=1}^{p}\left(\sum_{\alpha=1}^{q} a_{i\alpha}^2\right)=\sum_{\alpha=1}^{q}\left(\sum_{i=1}^{p} a_{i\alpha}^2\right)$$

$$=\sum_{i=1}^{p} V\left(t_i\right)-\sum_{i=1}^{p} V\left(\varepsilon_i\right)=p-\sum_{i=1}^{p} V\left(\varepsilon_i\right)$$

となるが、この関係式を利用し、誤差$\varepsilon_{i\lambda}$が 0 に近づけば左辺はpに近づくことを考えて、因子αの寄与率を以下のように定義する。

$$P_\alpha=\sum_{i=1}^{p} a_{i\alpha}^2 \Big/ p$$

我々は (2) 式を解いて因子負荷量$a_{i\alpha}$を求めようとするが、その求め方にも主因子法、主成分分析法、最小 2 乗法、最尤法等種々の方法がある[9]。ここでは主因子法と主成分分析法について説明する。最小 2 乗法と最尤法については次節で説明する。

主因子法では、最初に適当な推定値h_iを用いて、因子負荷量$a_{i\alpha}$を計算し、その値を使って再度$h_i=\sum_{\alpha=1}^{q} a_{i\alpha}^2$で共通性$h_i$を計算し、それをまた推定値として再び因子負荷量を計算する。これを共通性h_iが収束するまで（このプログラムでは前回との差が 0.001 以下になるまで）繰り返すという方法で近似値を求める。その際最初の共通性h_iの推定値には変数x_iと他の変数の重相関係数や他との相関係数の中で最大のものなどが利用される。ここで推定値h_iを与えた因子負荷量$a_{i\alpha}$の計算法としては、対角成分を共通性h_iで置き換えた相関行列$\mathbf{R}$の固有値と固有ベクトルを利用する。即ち、第α因子の因子負荷量$a_{i\alpha}$は、行列$\mathbf{R}$の固有値λ_αと規格化された固有ベクトル$u_{i\alpha}$を使って、以下のように与えられる。

$$a_{i\alpha}=\sqrt{\lambda_\alpha}\,u_{i\alpha}$$

主成分分析法では、相関行列$\mathbf{R}$をそのまま使い、固有値と固有ベクトルによって因子負荷量$a_{i\alpha}$を計算する。第α因子の因子負荷量$a_{i\alpha}$は、相関行列$\mathbf{R}$の固有値λ_αと規格化された固有ベクトル$u_{i\alpha}$を使って以下のように与える。

$$a_{i\alpha}=\sqrt{\lambda_\alpha}\,u_{i\alpha}$$

共通性は$h_i=\sum_{\alpha=1}^{p} a_{i\alpha}^2$のように因子負荷量から計算する。

次に各因子、各個体毎の因子得点 $f_{\alpha\lambda}$ の値について考える。前にも述べたとおり、誤差項が特定できない限り、一般に観測値 $x_{i\lambda}$ から因子得点 $f_{\alpha\lambda}$ を決定することはできない。我々のプログラムでは因子得点推定のために、Bartlett の重み付き最小 2 乗推定法と回帰推定法を用意している。

Bartlett の重み付き最小 2 乗推定法は、分散で重み付けされた誤差の 2 乗項

$$\sum_{\lambda=1}^{n}\sum_{i=1}^{p}\varepsilon_{i\lambda}^2/u_i^2=\sum_{\lambda=1}^{n}\sum_{i=1}^{p}(t_{i\lambda}-\sum_{\alpha=1}^{q}a_{i\alpha}f_{\alpha\lambda})^2\Big/u_i^2$$

が最小になるように仮定して、因子得点 $f_{\alpha\lambda}$ を推定する。この解は成分が

$$(\mathbf{F})_{\lambda\alpha}=f_{\alpha\lambda},\quad (\mathbf{T})_{\lambda i}=t_{i\lambda},\quad (\mathbf{A})_{i\alpha}=a_{i\alpha},\ (\mathbf{D})_{ij}=u_i^2\delta_{ij},$$

のように与えられる行列 $\mathbf{F},\mathbf{T},\mathbf{A},\mathbf{D}$ を用いて以下のように求められる。

$$\mathbf{F}=\mathbf{TD}^{-1}\mathbf{A}({}^t\mathbf{AD}^{-1}\mathbf{A})^{-1} \qquad (3)$$

また回帰推定法では、(1)式から、共通因子の推定値と変数は以下のような関係にあると考える。

$$t_{i\lambda}=\sum_{\alpha=1}^{q}a_{i\alpha}\hat{f}_{\alpha\lambda}$$

これから、

$$\sum_{i=1}^{p}a_{i\alpha}t_{i\lambda}=\sum_{i=1}^{p}\sum_{\beta=1}^{q}a_{i\alpha}a_{i\beta}\hat{f}_{\beta\lambda}=\sum_{\beta=1}^{q}\lambda_\alpha\delta_{\alpha\beta}\hat{f}_{\beta\lambda}=\lambda_\alpha\hat{f}_{\alpha\lambda}$$

となり、以下を得る。

$$\hat{f}_{\alpha\lambda}=\frac{1}{\lambda_\alpha}\sum_{i=1}^{p}a_{i\alpha}t_{i\lambda}=\sum_{i=1}^{p}\sum_{j=1}^{p}r^{ij}a_{j\alpha}t_{i\lambda}\equiv\sum_{i=1}^{p}b_{\alpha i}t_{i\lambda} \qquad (4)$$

ここで、$\mathbf{Ra}_\alpha=\lambda\mathbf{a}_\alpha\Leftrightarrow\mathbf{R}^{-1}\mathbf{a}_\alpha=(1/\lambda_\alpha)\mathbf{a}_\alpha$ の関係を用いた。これにより、因子得点を求める係数 $b_{\alpha i}$ は以下のように与えられる。

$$b_{\alpha i}=\sum_{j=1}^{p}r^{ij}a_{j\alpha} \qquad (5)$$

この関係は、(4)式が $\hat{f}_\alpha$ を推定する重回帰分析の式（目的変数には実測値がないが）であると考えることによっても導かれる。重回帰式の標準化係数は $b_i=\sum_{j=1}^{p}r^{ij}r_{jy}$ であり、r_{jy} は変数 j と目的変数の相関係数である。この場合目的変数は因子 α なので、相関係数は因子負荷量 $a_{j\alpha}$ である。我々のプログラムのデフォルトでは、回帰推定法を採用している。

ここで求めた因子負荷量 $a_{i\alpha}$ には、$a_{i\alpha}^*=\sum_{\beta=1}^{q}o_{\alpha\beta}a_{i\beta}$，$\sum_{\gamma=1}^{q}o_{\alpha\gamma}o_{\beta\gamma}=\delta_{\alpha\beta}$ のような回転の自由度が存在する。この変換により、(1)式は以下のように変わり、因子も回転を受ける。

$$t_i=\sum_{\alpha=1}^{q}a_{i\alpha}^*f_\alpha^*+\varepsilon_i\ ,\quad f_\alpha^*=\sum_{\beta=1}^{q}o_{\alpha\beta}f_\beta$$

しかし、(2)式、寄与率、因子の平均と分散や直交性は不変である。この性質を利用して、

因子負荷量の各因子の分散を最大化するように回転させると因子の解釈が容易になる。この直交回転をバリマックス回転という。

このようにして推測された共通因子からデータはどの程度推測できるのであろうか。実際に以下の式によってデータを推測し、観測値との相関係数を調べてみるとモデルの良さが実感できる。

$$\hat{t}_{i\lambda} = \sum_{\alpha=1}^{q} a_{i\alpha} \hat{f}_{\alpha\lambda} \qquad (6)$$

最後に直交性を保たずに、より因子の説明を付けやすいように軸の向きを決める斜交回転について述べておく[1]。斜交回転の軸に用いられる用語として、プライマリ因子軸とは、斜交回転をした場合の斜交軸のことであり、参考因子軸とは、プライマリ因子軸と直交する斜交軸のことである。直交回転の場合の因子負荷量行列は、斜交回転の場合、因子パターン行列と因子構造行列に区別される。斜交軸の座標の取り方として、図１のベクトル **a** の座標の取り方で、（a_1, a_2）が因子パターン行列に基づく取り方、（a'_1, a'_2）が因子構造行列に基づく取り方である。

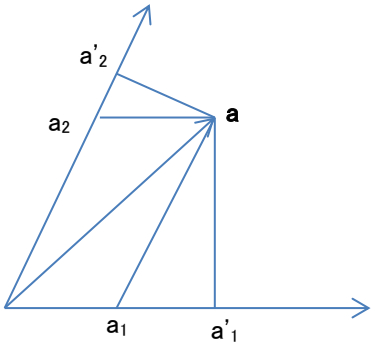


図１　斜交軸

直交回転の場合の因子負荷量行列は因子と変量の相関を与えるが、斜交回転の場合、因子と変量の相関を与えるものは因子構造行列である。以後プロマックス回転における因子パターン行列と因子構造行列の求め方を段階に分けて示す。

ステップ１

直交回転後の因子負荷行列 **A** から始める。

A の各要素を、各行の２乗和が１となるように共通性を用いて基準化する。

各列の絶対値最大の成分が±1（我々の場合はバリマックス回転ですでに正）となるように定数倍する。

A の各要素を k 乗したものを目標行列 $\mathbf{A}^*$ とする。ここで k が奇数の場合はそのまま、偶数の場合は要素の符号をかけておく。通常 k は３か４を指定するが、我々の場合は４にしている。

これによって、絶対値が１に近いものを除き、他の要素は０に近づく。

ステップ２

回転後の$\mathbf{A}$が$\mathbf{A}^*$と最小２乗法の意味で最も近くなるような変換（プロクラステス変換）行列$\mathbf{T}_r$は以下の式で与えられる。

$$\mathbf{T}_r = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{A}^*$$

ステップ３

プライマリ因子の因子構造を計算するための回転行列$\mathbf{T}_p$は、$\mathbf{T}_p = (\mathbf{T}_r')^{-1}$の列ノルムを１に基準化した行列である。

ステップ４

参考因子軸に沿った成分である因子構造行列$\mathbf{S}_p$、プライマリ因子軸に沿った成分である因子パターン行列$\mathbf{P}_p$、因子間の相関行列Φ_pを以下より求める。ここで、結果には因子構造行列$\mathbf{S}_p$と因子パターン行列$\mathbf{P}_p$を用いる。

$$\mathbf{S}_p = \mathbf{A}\mathbf{T}_p, \quad \mathbf{P}_p = \mathbf{A}(\mathbf{T}_p')^{-1}, \quad \Phi_p = \mathbf{T}_p'\mathbf{T}_p$$

最後に、斜交回転の場合、因子得点の計算と因子からのデータの推測で、使われる行列が異なることを注意しておく。Bartlett の重み付き最小２乗推定法の場合、(3)式の$\mathbf{A}$には因子パターン行列、(6)式の$a_{i\alpha}$にも因子パターン行列を使う。一方、回帰推定法では、(5)式の$a_{i\alpha}$には因子構造行列、(6)式の$a_{i\alpha}$には因子パターン行列を使う。予測される結果は、因子の回転によらず同じ結果となる。

斜交回転では因子と変数との相関を表す因子構造行列は分離が悪くなるため、因子の解釈には向かない。そのため、因子パターン行列を使って因子が強く影響する変数から推測する方法が採られる。

最後に回帰推定法を用いて上で述べたことを示しておこう。$\mathbf{X}$を標準化された観測データ、$\mathbf{f}$を直交回転での因子得点とする。回帰推定法ではこれらは以下の関係にある。

$$\mathbf{X} = \mathbf{f}\mathbf{A}', \quad \mathbf{f} = \mathbf{X}\mathbf{R}^{-1}\mathbf{A}$$

今、斜交回転によって、$\mathbf{f} \to \mathbf{f}^*, \mathbf{A} \to \mathbf{A}^*$と変換したとする。そのとき、

$$\mathbf{X} = \mathbf{f}^*\mathbf{A}^{*\prime}, \quad \mathbf{f}^* = \mathbf{X}\mathbf{R}^{-1}\mathbf{A}\mathbf{T}_p, \quad \mathbf{A}^{*\prime} = \mathbf{T}_p^{-1}\mathbf{A}'$$

のように変換すれば結果は同じである。最後の式は$\mathbf{A}^* = \mathbf{A}(\mathbf{T}_p')^{-1}$である。よって、変換は以下となる。

$$\mathbf{f}^* = \mathbf{X}\mathbf{R}^{-1}\mathbf{S}_p, \quad \mathbf{A}^* = \mathbf{P}_p$$

同じように考えれば Bartlett の重み付き最小２乗推定法の場合も理解できる。

5.4　最尤法と最小２乗法による因子分析

標準化された観測変数を以下のように p 個の因子（内生変数）で表すものとする。

$$x_{i\lambda} = \sum_{j=1}^{p} a_{ij} f_{p\lambda} + b_i e_{i\lambda}$$

これを用いると、相関係数 s_{ij} は以下のように書ける。

$$\begin{aligned} s_{ij} &= \frac{1}{N}\sum_{\lambda=1}^{N} x_{i\lambda}x_{j\lambda} = \frac{1}{N}\sum_{\lambda=1}^{N}(\sum_{k=1}^{p} a_{ik} f_{k\lambda} + b_i e_{i\lambda})(\sum_{l=1}^{p} a_{jl} f_{l\lambda} + b_j e_{j\lambda}) \\ &= \sum_{k=1}^{p}\sum_{l=1}^{p} a_{ik}a_{jl}\frac{1}{N}\sum_{\lambda=1}^{N} f_{k\lambda} f_{l\lambda} + b_i b_j \frac{1}{N}\sum_{\lambda=1}^{N} e_{i\lambda}e_{j\lambda} \\ &= \sum_{k=1}^{p} a_{ik}a_{jk} + b_i b_j \delta_{ij} \equiv \mathbf{\Sigma}_{ij} \end{aligned}$$

ここに、因子と誤差について、以下の関係があるものとする。

$$\frac{1}{N}\sum_{\lambda=1}^{N} f_{k\lambda} f_{l\lambda} = \delta_{kl}\,,\quad \frac{1}{N}\sum_{\lambda=1}^{N} e_{i\lambda}e_{j\lambda} = \delta_{ij}\,,\quad \sum_{\lambda=1}^{N} f_{k\lambda}e_{i\lambda} = 0$$

最尤法では、以下の尤度を考える。

$$\begin{aligned} L &= \prod_{\lambda=1}^{N} \frac{1}{\sqrt{(2\pi)^p\,|\mathbf{\Sigma}|}} \exp\left[-\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(\mathbf{\Sigma}^{-1})_{ij}\, x_{i\lambda}x_{j\lambda}\right] \\ &= (2\pi)^{-pN/2}\,|\mathbf{\Sigma}|^{-N/2}\,\exp\left[-\frac{N}{2}tr(\mathbf{\Sigma}^{-1}\mathbf{S})\right] \end{aligned}$$

これを最大化するために、符号を反対にした対数尤度の最小化を考える。

$$\begin{aligned} -\log L &= \frac{pN}{2}\log(2\pi) + \frac{N}{2}\log|\mathbf{\Sigma}| + \frac{N}{2}tr(\mathbf{\Sigma}^{-1}\mathbf{S}) \\ &= \frac{N}{2}\left[tr(\mathbf{\Sigma}^{-1}\mathbf{S}) - \log|\mathbf{\Sigma}^{-1}|\right] + const. \end{aligned}$$

この形から、$\mathbf{\Sigma}$ と $\mathbf{S}$ が完全に一致する場合に 0 になるように、評価関数 $f_{ML}(\mathbf{a},\mathbf{b})$ を以下のように定義してこれを最小化する。

$$f_{ML}(\mathbf{a},\mathbf{b}) = tr(\mathbf{\Sigma}^{-1}\mathbf{S}) - \log|\mathbf{\Sigma}^{-1}\mathbf{S}| - p$$

最小２乗法では、評価関数 $f_{MS}(\mathbf{a},\mathbf{b})$ を以下のように定義してこれを最小化する。

$$f_{MS}(\mathbf{a},\mathbf{b}) = \sum_{i=1}^{p}\sum_{j=1}^{p}(\Sigma_{ij} - s_{ij})^2$$

評価関数が複雑なため、パラメータの初期値の与え方が重要であるが、我々のプログラムでは、確実に値の求まる主成分分析法の結果を初期値として用いている。

5.5　α係数とω係数の考え方

Cronbach のα係数と Mcdonald のω係数は複数の変数をひとまとめに考える妥当性を与える指標である。定義にはデータをそのまま使うことが多いが、標準化して使うこともある。ここでは両方の場合でこれらの指標について定義を与えておく。また、変数については逆転項目になっている場合があるが、これは変数の符号を変えることによって対応する。以下記号として、p を変数の数、Σ を分散共分散行列、σ_{ij} をその成分、$\mathbf{R}$ を相関行列、r_{ij} をその成分とする。

Cronbach のα係数について

観測データ x_i から出発した場合の定義は以下である。

$$\alpha = \frac{p}{p-1}\left[1-\frac{\sum_{i=1}^{p}\sigma_{ii}}{\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma_{ij}}\right] = \frac{p}{p-1}\left[1-\frac{tr\Sigma}{V_X}\right]$$

ここに、$V_X = \mathbf{1}'\Sigma\mathbf{1} = \sum_{i=1}^{r}\sum_{j=1}^{r}\sigma_{ij}$，　$tr\Sigma = \sum_{i=1}^{r}\sigma_{ii}$

また、以下の関係が成り立つことから、V_X は変数の合計の分散と考えることができる。

$$Var\left(\sum_{i=1}^{p}x_i\right) = Var\left(\sum_{i=1}^{p}(x_i-\mu_i)\right) = E\left(\sum_{i=1}^{p}(x_i-\mu_i)\sum_{j=1}^{p}(x_j-\mu_j)\right)$$
$$= \sum_{i=1}^{p}\sum_{j=1}^{p}\sigma_{ij} = V_X$$

標準化データ $y_i = (x_i-\mu_i)\big/\sqrt{\sigma_{ii}}$ から出発した場合の定義は以下である。

$$\alpha = \frac{p}{p-1}\left[1-\frac{\sum_{i=1}^{p}r_{ii}}{\sum_{i=1}^{p}\sum_{j=1}^{p}r_{ij}}\right] = \frac{p}{p-1}\left[1-\frac{trR}{R_X}\right] == \frac{p}{p-1}\left[1-\frac{p}{R_X}\right] < \frac{p}{p-1}\left[1-\frac{p}{p^2}\right] = 1$$

ここに、$R_X = \mathbf{1}'R\mathbf{1} = \sum_{i=1}^{p}\sum_{j=1}^{p}r_{ij}$，　$trR = \sum_{i=1}^{p}r_{ii} = p$

同様に、以下の関係が成り立つ。

$$Var\left(\sum_{i=1}^{p}y_i\right) = Var\left(\sum_{i=1}^{p}y_i\right) = E\left(\sum_{i=1}^{p}y_i\sum_{j=1}^{p}y_j\right)$$
$$= \sum_{i=1}^{p}\sum_{j=1}^{p}r_{ij} = R_X$$

観測データから出発したα係数はデータの大小に左右されるが、標準化データから出発した場合はその心配はない。

McDonald のω係数（1978, ω_h）について

観測データ x_i から出発した場合は以下の仮定をおく。

$$x_i = \mu_i + \sqrt{\sigma_{ii}}\lambda_i f + e_i$$

ここに、μ_i は平均値、λ_i は１因子で表される因子分析の因子負荷量、f はその因子で標準化された変数、e_i は変数に固有の値である。但し、因子分析は標準化変数を元に計算されたものとする。ω係数は以下のように定義される。

$$\omega = \frac{Var\left(\sum_{i=1}^{p}\sqrt{\sigma_{ii}}\lambda_i f\right)}{Var\left(\sum_{i=1}^{p}x_i\right)} = \frac{\left(\sum_{i=1}^{p}\sqrt{\sigma_{ii}}\lambda_i\right)^2}{V_X}$$

これは変数の合計の変動がどれだけ１つの因子で表されるかということを考えた指標ととらえることができる。

標準化データ $y_i = (x_i - \mu_i)/\sqrt{\sigma_{ii}}$ から出発した場合は以下を仮定する。

$$y_i = \lambda_i f + e_i'$$

ここに e_i' は変数に固有の値である。ω係数は以下のように定義される。

$$\omega = \frac{Var\left(\sum_{i=1}^{p}\lambda_i f\right)}{Var\left(\sum_{i=1}^{p}y_i\right)} = \frac{\left(\sum_{i=1}^{p}\lambda_i\right)^2}{R_X}$$

上のように考えると McDonald のωは非常に素直な定義と考えられる。また、観測データから出発した指標は単純に変数を足すことを前提にした指標､標準化データから出発した指標は標準化したのち足すことを前提にした指標と考えられる。

ω係数についての変数の仮定は、まさに 1 因子の因子分析の仮定である。係数 λ_i の値は様々な方法で推定されるが、SPSS では最尤法を利用しているようである。我々は、変数が 3 つ以上の場合は最尤法、2 つの場合は、最尤法が識別不能になるので、主成分分析法を利用している。ω係数には Mcdonald が定義した ω_h (1978)と ω_t (1999) があるようだが､我々は因子分析内で使うことを考えて、 ω_h の定義を利用した。

６．クラスター分析

6.1　クラスター分析とは

複数の変数の値によって、個体を分類したり、また個体の特徴によって、変数を分類したりする手法をクラスター分析という。特に、分類を段階的に行う手法を階層的クラスター分析という。

クラスター分析には、回答の類似度で個体（レコード）を分類する場合と回答の類似度で変数を分類する場合がある。回答の類似度は個体間の距離（非類似度）で表され、階層的クラスター分析では、距離の近いものからクラスターと呼ばれる組を構成して行く。この個体間の距離には、量的データと質的データとで以下のようにいくつかの定義がある。

量的データ：ユークリッド距離、標準化ユークリッド距離、マハラノビス距離　等

質的 0/1 データ：類似比、一致係数、φ係数　等

また、変数間の距離にもいくつかの定義がある。

量的データ：1−相関係数、1−|相関係数|、1−順位相関係数、1−|順位相関係数|

質的データ：平均平方根一致係数を使ったもの、1−一致係数、1−クラメールのＶ　等

これらは、データの性質により使い分けられる。

個体間、変数間の距離を定義した後、それらを使ってさらにクラスター間の距離を定義する。クラスター間の距離の定義はクラスター構成法とも呼ばれ、最短距離法，最長距離法、群平均法、重心法、メジアン法、ウォード法等、様々な方法が考えられている。特にクラスターの分離が分かり易い最長距離法や、クラスター内の分散を最小化するウォード法等がよく利用される。

クラスターを構成する過程を表示するには、縦軸にクラスター間の距離を取ったデンドログラムと呼ばれる図が利用される。我々のプログラムではこれに付随してクラスター間の距離を数値的に示す表も出力するようにしている。

表 1 は個人の酒類の好みを 1 から 9 の点数で表したデータである。

表１　酒類の好みを表したデータ（クラスター分析 1.txt）

	日本酒	焼酎	ビール	ウィスキー	ワイン
増川	1	2	9	6	5
西山	3	1	7	5	4
三好	5	3	4	2	2
芝田	3	6	2	8	3
尾崎	4	6	9	3	4
藤田	7	2	5	4	5
細川	7	5	4	3	2

このデータを元にクラスター分析を行ったまとめを示しておく。

酒類の好みを与えたクラスター分析 1.txt のデータから、クラスター分析を用いて各人の好みの分類を行った。距離測定法はユークリッド距離、クラスター構成法は最短距離法と最長距離法を用いて図 1 と図 2 のようなデンドログラムを得た。

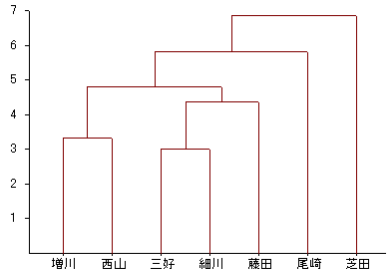


図１　最短距離法

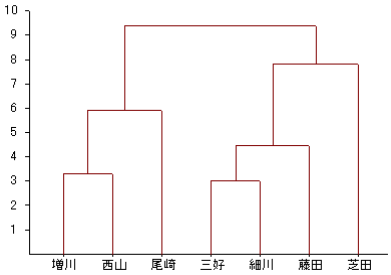


図２　最長距離法

これらによると、クラスターは大きく増川・西山さんと三好・細川・藤田さんに分かれる。尾崎さんと芝田さんは比較的独立な存在であると思われる。

酒の種類については、距離測定法は１－相関係数を用いる方法で、クラスター構成法は最短距離法と最長距離法を用いて分析を行い、図３と図４のようなデンドログラムを得た。

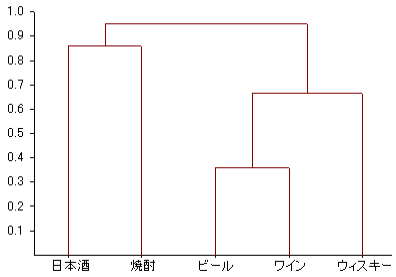


図３　最短距離法

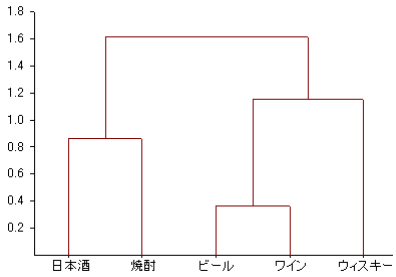


図４　最長距離法

これらによると２つのクラスター構成法とも日本酒・焼酎、ビール・ワイン・ウィスキーに分かれている。

6.2　プログラムの利用法

メニュー［分析－多変量解析－分類手法－クラスター分析］を選択して表示される、クラスター分析の実行画面を図１に示す。

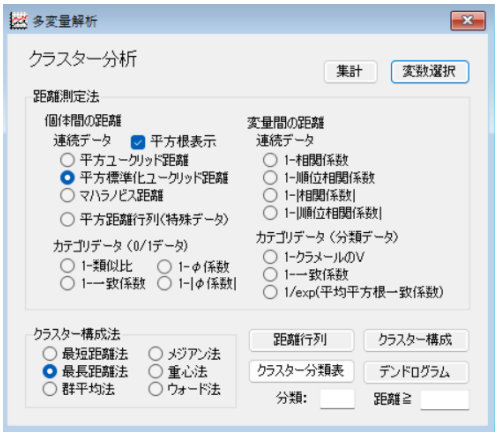


図１　クラスター分析実行画面

変数を選択して「距離行列」ボタンをクリックした場合の出力結果を図 2 に示す。これは各要素の類似度（距離）を表示したものである。


要素間類似度

	増川	西山	三好	芝田	尾崎	藤田	細川
▶ 増川	0.0000	1.5660	4.0301	3.8370	2.8785	3.2378	4.5335
西山	1.5660	0.0000	2.7501	3.4648	2.7428	2.2134	3.4122
三好	4.0301	2.7501	0.0000	3.5335	2.9089	2.7711	1.4079
芝田	3.8370	3.4648	3.5335	0.0000	3.6640	3.8004	3.2402
尾崎	2.8785	2.7428	2.9089	3.6640	0.0000	2.9377	2.8272
藤田	3.2378	2.2134	2.7711	3.8004	2.9377	0.0000	2.8338
細川	4.5335	3.4122	1.4079	3.2402	2.8272	2.8338	0.0000

図 2　類似度行列

クラスター分析で最も利用する「デンドログラム」の出力結果を図 3 に与える。

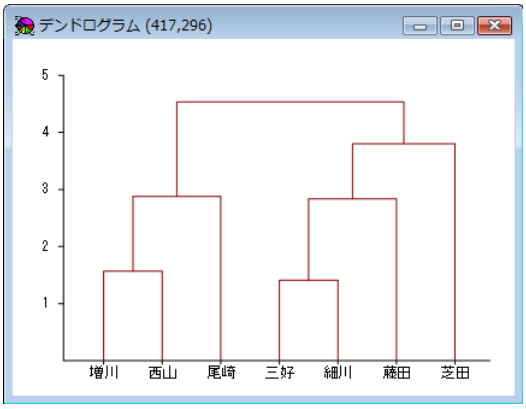


図 3　デンドログラム

デンドログラムでは構成の際の類似度が読みづらいので構成順を表にして示す。「クラスター構成」ボタンをクリックすると図 4 に示される結果が表示される。


クラスター分析結果

	クラスター名	クラスター名	類似度
▶ 1	E:三好	E:細川	1.4079
2	E:増川	E:西山	1.5660
3	C:三好	E:藤田	2.8338
4	C:増川	E:尾崎	2.8785
5	C:三好	E:芝田	3.8004
6	C:増川	C:三好	4.5335

図 4　クラスターの構成

クラスター名の先頭に E の付いたものは要素（Element）名、C の付いたものはクラスター（Cluster）である。クラスター名はデンドログラムで表示される左端の要素名で代表される。例えば、最初の行は、要素「三好」と要素「増川」が結合され、クラスター「三好」になる、と読む。また、3 番目の行は、クラスター「三好」と要素「藤田」が結合され、クラスター「三好」になる、と読む。

「クラスター分類表」ボタンをクリックすると、例えば、図 3 のデンドログラムを表形式で表した図 5 のクラスター分類表が表示される。これはクラスター構成の各段階での分類を表示している。これによって例えば全体を 3 分割するときに各個体がどのクラスターに属するか簡単に知ることができる。また、これを利用して 2 つのクラスター間での有意差検定などを行いたい場合、この表の列をコピーして元データに加え、簡単に群分けするこ

とができるようになる。また、「分類：」テキストボックスに数値（分類数＜要素数）を入力するとその部分だけ取り出して表示する。

図５　クラスター分類表（右は分類数指定の場合）

データに欠損値がある場合、この分類表ではその部分を空欄にして表示する。我々はこの考えをすべての多変量解析に適用し、予測値には欠損値も加えて表示するように変更した。

問題１（クラスター分析 4.txt）

クラスター分析 4.txt はある野球チームの今年度の成績である。これについてクラスター分析を行い以下の問いに答えよ。

１）ユークリッド距離及び標準化ユークリッド距離を用いた場合、山下と田中の距離はいくらか。ユークリッド距離［　　］　標準化ユークリッド距離［　　］

２）各変数の標準偏差はいくらか。

打率	安打	本塁打	打点	盗塁

３）上の結果から、距離測定法はどちらを利用すべきか。

［ユークリッド距離・標準化ユークリッド距離］　以後はこの距離を用いる。

４）クラスター構成法を最長距離法とする場合、最初にクラスターを構成するのはどの要素とどの要素でそれらの距離はいくらか。

［　　］と［　　］で距離［　　］

５）最長距離法の場合、4 分類か 5 分類が適当と思われるが、4 分類の場合、各クラスターにはどのような要素が含まれるか。

［　　］［　　］［　　］［　　］

６）最長距離法と最短距離法とでどちらの分類が理解しやすいと思われるか。

［最長距離法・最短距離法］

８）１－相関係数の距離測定法で最長距離法を用いて変数を 3 分類すると各クラスターに含まれる要素はどのようになるか。

［　　］［　　］［　　］

問題２（クラスター分析 3.txt）

クラスター分析 3.txt のデータを用いてクラスター分析を行い、以下の問いに答えよ。

１）個体の分類

距離測定法は標準化ユークリッド距離、クラスター構成法は最長距離法を用いると、3 分類の場合、各クラスターに含まれる要素はどうなるか。

［　　］［　　］［　　］

２）変数の分類

距離測定法は１－相関係数、クラスター構成法は最長距離法を用いると、2 分類の場合、各クラスターに含まれる要素はどうなるか。

［　　］［　　］

演習１

多変量演習 9.txt は学生による授業評価のデータであり、レコード（個体）は１つの授業で調べた質問項目（変数）ごとの平均を表している。このデータからクラスター分析を用いて、個体や変数の類似性の特徴を見出したい。以下の質問に答えよ。

１）ユークリッド距離を用いた場合、1 番と 12 番の距離はいくらか。［　　］

２）クラスター構成法を最長距離法、距離測定法をユークリッド距離とする場合、最初にクラスターを構成するのは何番と何番でそれらの距離はいくらか。

個体［　　］番と個体［　　］番で、距離［　　］

３）上の設定で、最初にクラスターとクラスター、またはクラスターと要素の結合になるのはどのようなクラスター（要素）か。それらに含まれる要素を示せ。またその際の距離はいくらか。

クラスター［　　］とクラスター（要素）［　　］　距離［　　］

４）上の設定でクラスター分析を実行し、4 つのクラスターに分けたとき、それらのクラスターに含まれる要素（授業の番号）は何か。

［　　］［　　］［　　］［　　］

５）5 番が含まれるクラスターと 10 番が含まれるクラスターの最も大きな特徴は何か。5 番［　　］　10 番［　　］

６）距離測定法を標準化ユークリッド距離（各変数を標準化したときのユークリッド距離）に変えた場合、クラスター構成は大きく変わるか。

［変わる・あまり変わらない］　　注）標準化値＝（値－平均値）／標準偏差

７）これにはどんな理由が考えられるか。

各変数の［　　］があまり違わないから。

８）距離測定法をユークリッド距離とし、クラスター構成法を最短距離法に変えるとクラスター構成は大きく変わるか。［変わる・あまり変わらない］

９）ユークリッド距離の場合、その他のクラスター構成法は最長距離法と最短距離法のどちらに近いか。［最長距離法・最短距離法］

各質問についての分類を行いたいが、距離測定法を１－相関係数として以下の問いに答えよ。

10)最長距離法で上の距離測定法を用いる場合、最初にクラスターを構成するのは何と何で、そのときの距離はいくらか。

変数［　　］と変数［　　］で、距離［　　］

11）上の設定でクラスター分析を行い、変数を３つのクラスターに分類する場合、それらのクラスターに含まれる要素（変数）は何か。

［　　］［　　］［　　］

問題１解答（クラスター分析 4.txt）

１）ユークリッド距離及び標準化ユークリッド距離を用いた場合、山下と田中の距離はいくらか。ユークリッド距離［ 21.657 ］　標準化ユークリッド距離［ 4.696 ］

２）各変数の標準偏差はいくらか。

打率	安打	本塁打	打点	盗塁
0.027	6.177	3.778	5.839	3.516

３）上の結果から、距離測定法はどちらを利用すべきか。

［ユークリッド距離・標準化ユークリッド距離］　以後はこの距離を用いる。

４）クラスター構成法を最長距離法とする場合、最初にクラスターを構成するのはどの要素とどの要素でそれらの距離はいくらか。

［ 小川 ］と［ 青木 ］で距離［ 0.737 ］

５）最長距離法の場合、4 分類か 5 分類が適当と思われるが、4 分類の場合、各クラスターにはどのような要素が含まれるか。

［ 山下 ］［ 田中,鈴木,小川,青木,荻原,高田 ］［ 黒岩,田村,岩崎 ］［ 斉藤,井上 ］

６）最長距離法と最短距離法とでどちらの分類が理解しやすいと思われるか。

［最長距離法・最短距離法］

８）１－相関係数の距離測定法で最長距離法を用いて変数を 3 分類すると各クラスターに含まれる要素はどのようになるか。

［ 打率,安打 ］［ 盗塁 ］［ 本塁打,打点 ］

問題２解答（クラスター分析 3.txt）

１）個体の分類

距離測定法は標準化ユークリッド距離、クラスター構成法は最長距離法を用いると、3 分類の場合、各クラスターに含まれる要素はどうなるか。

［ 宮地,後藤,黒木,大成 ］［ 大門,貝田,坂田,福井 ］［ 武田,田口 ］

２）変数の分類

距離測定法は１－相関係数、クラスター構成法は最長距離法を用いると、2 分類の場合、各クラスターに含まれる要素はどうなるか。

［ 身長,座高 ］［ 体重,胸囲 ］

演習１解答（多変量演習 9. txt）

１）ユークリッド距離を用いた場合、1 番と 12 番の距離はいくらか。［ 0.404 ］

２）クラスター構成法を最長距離法、距離測定法をユークリッド距離とする場合、最初にクラスターを構成するのは何番と何番でそれらの距離はいくらか。

個体［ 15 ］番と個体［ 16 ］番で、距離［ 0.199 ］

3）上の設定で、最初にクラスターとクラスター、またはクラスターと要素の結合になるのはどのようなクラスター（要素）か。それらに含まれる要素を示せ。またその際の距離はいくらか。

クラスター［ 8,14 ］とクラスター（要素）［ 15,16 ］　距離［ 0.568 ］

4）上の設定でクラスター分析を実行し、４つのクラスターに分けたとき、それらのクラスターに含まれる要素（授業の番号）は何か。

［ 1,12,4,3,18,2,20 ］［ 8,14,15,16,17 ］［ 9,10,19 ］［ 5,11,13,6,7 ］

5）5 番が含まれるクラスターと 10 番が含まれるクラスターの最も大きな特徴は何か。

5 番［ 高い評価 ］　10 番［ 低い評価 ］

6）距離測定法を標準化ユークリッド距離（各変数を標準化したときのユークリッド距離）に変えた場合、クラスター構成は大きく変わるか。

［変わる・あまり変わらない］　　注）標準化値＝（値－平均値）／標準偏差

7）これにはどんな理由が考えられるか。

各変数の［ 標準偏差 ］があまり違わないから。

8）距離測定法をユークリッド距離とし、クラスター構成法を最短距離法に変えるとクラスター構成は大きく変わるか。［変わる・あまり変わらない］

9）ユークリッド距離の場合、その他のクラスター構成法は最長距離法と最短距離法のどちらに近いか。［最長距離法・最短距離法］

各質問についての分類を行いたいが、距離測定法を１－相関係数として以下の問いに答えよ。

10）最長距離法で上の距離測定法を用いる場合、最初にクラスターを構成するのは何と何で、そのときの距離はいくらか。

変数［ 分かり易さ ］と変数［ 有益さ ］で、距離［ 0.100 ］

11）上の設定でクラスター分析を行い、変数を３つのクラスターに分類する場合、それらのクラスターに含まれる要素（変数）は何か。

［ 進む速さ,黒板等,分かり易さ,有益さ ］［ 声の大きさ,受講態度 ］［ 死後注意 ］

6.3　クラスター分析の理論

クラスター分析は個体や変数間の様々に定義された距離に基づき、これらを分類する手法である。その中でもここで取り扱うのはクラスターを 1 つずつまとめてゆく階層的方法と呼ばれるものである。クラスター分析のデータは変数と個体のシート形式で、表 1 のように与えられる。

表 1 クラスター分析のデータ

	変数 1	変数 2	…	変数 p
個体 1	x_{11}	x_{21}	…	x_{p1}
個体 2	x_{12}	x_{22}	…	x_{p2}
:	:	:	:	:
個体 n	x_{1n}	x_{2n}	…	x_{pn}

クラスター分析には距離の測定方法やクラスターの構成法にさまざまな種類があるが、ここでは利用者の理解し易い代表的な数種のものについて取り上げている。距離の測定は 2 つの個体または変数の間で定義される。これらが複数個集まったクラスター間の距離の定義にはクラスター構成法を利用する。

ここではまず、距離の測定方法を個体間のものと変数間のものに分けて説明する。個体 μ と個体 ν との距離には以下のようなものがある。最初に量的なデータに対してその定義を示す。

ユークリッド距離　$$d_{\mu\nu}^2 = \sum_{i=1}^{p} (x_{i\mu} - x_{i\nu})^2$$

標準化ユークリッド距離　$$d_{\mu\nu}^2 = \sum_{i=1}^{p} \frac{1}{s_i^2} (x_{i\mu} - x_{i\nu})^2$$

マハラノビス距離　$$d_{\mu\nu}^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} (x_{i\mu} - x_{i\nu}) s^{ij} (x_{j\mu} - x_{j\nu})$$

ここに s_i^2 は変数 i の不偏分散、添え字の上に付いた s^{ij} は共分散行列 $\mathbf{S}$ の逆行列 $\mathbf{S}^{-1}$ の i, j 成分である。

$$s_i^2 = \frac{1}{n-1} \sum_{\lambda=1}^{n} (x_{i\lambda} - \bar{x}_i)^2 , \quad (\mathbf{S})_{ij} = s_{ij} = \frac{1}{n-1} \sum_{\lambda=1}^{n} (x_{i\lambda} - \bar{x}_i)(x_{j\lambda} - \bar{x}_j)$$

ここでは標準化ユークリッド距離とマハラノビスの距離がよく利用される。

次に 0/1 の値で与えられるカテゴリデータに対しては、以下の統計量を距離として用いる。

類似比　$d_{\mu\nu} = a/(a+b+c)$

一致係数　$d_{\mu\nu} = (a+d)/(a+b+c+d)$

ファイ係数　$d_{\mu\nu} = (ad-bc)\Big/\sqrt{(a+b)(c+d)(a+c)(b+d)}$

ここに、a, b, c, d は以下のように与えられる。

$$a=\sum_{i=1}^{p}x_{i\mu}x_{i\nu}\ ,\quad b=\sum_{i=1}^{p}x_{i\mu}(1-x_{i\nu})\ ,\quad c=\sum_{i=1}^{p}(1-x_{i\mu})x_{i\nu}\ ,\quad d=\sum_{i=1}^{p}(1-x_{i\mu})(1-x_{i\nu})$$

ここではファイ係数が重要である。

次に、変数 i, j 間の距離について述べる。数値データに対しては、以下の統計量を距離として用いる。

相関　　$d_{ij}=1-s_{ij}/s_i s_j$　（1−相関係数）

順位相関　　$d_{ij}=1-\tilde{s}_{ij}/\tilde{s}_i\tilde{s}_j$　（1−順位相関係数）

ここに、$\tilde{s}_i$ 及び $\tilde{s}_{ij}$ は、データの代わりに変数別に付与された順位データを用いて求めた標準偏差と共分散である。

カテゴリデータに対しては、まず以下のような変数 i, j に対する統計量 χ_{ij}^2 を求める。

$$\chi_{ij}^2=\sum_{k=1}^{r_i}\sum_{l=1}^{r_j}\frac{\left(\left|n_{kl}-n_{k\bullet}n_{\bullet l}/n\right|-1/2\right)^2}{n_{k\bullet}n_{\bullet l}/n}$$

ここに、r_i は変数 i の分類数、n_{kl} は変数 i の k 番目の分類と変数 j の l 番目の分類に含まれるデータ数及び、$n_{k\bullet}$ と $n_{\bullet l}$ はそれぞれ n_{kl} の l についての和と k についての和である。

これを用いて以下のように距離を定義する。

平均平方根一致係数　　$d_{ij}=\sqrt{\chi_{ij}^2/n}$

一致係数　　$d_{ij}=\sqrt{\chi_{ij}^2/(\chi_{ij}^2+n)}$

クラメールの V　　$d_{ij}=\sqrt{(\chi_{ij}^2/n)/\min(r_i-1,r_j-1)}$

次にクラスター構成法について述べる。ここではクラスター f とクラスター g を結合してクラスター h を作り、他のクラスター l との距離を求める場合を考える。クラスター h とクラスター l の距離を D_{hl} で表わすと、これらの関係は以下のように与えられる。

最短距離法　　$D_{hl}=\frac{1}{2}D_{fl}+\frac{1}{2}D_{gl}-\frac{1}{2}\left|D_{fl}-D_{gl}\right|$

最長距離法　　$D_{hl}=\frac{1}{2}D_{fl}+\frac{1}{2}D_{gl}+\frac{1}{2}\left|D_{fl}-D_{gl}\right|$

メジアン法　　$D_{hl}=\frac{1}{2}D_{fl}+\frac{1}{2}D_{gl}-\frac{1}{4}D_{fg}$

重心法　　$D_{hl}^2=\frac{n_f}{n_h}D_{fl}^2+\frac{n_g}{n_h}D_{gl}^2-\frac{n_f n_g}{n_h^2}D_{fg}^2$

群平均法　　$D_{hl}^2=\frac{n_f}{n_h}D_{fl}^2+\frac{n_g}{n_h}D_{gl}^2$

ウォード法　　$D_{hl}^2=\frac{1}{n_h+n_l}\left[(n_f+n_l)D_{fl}^2+(n_g+n_l)D_{gl}^2-n_l D_{fg}^2\right]$

７．正準相関分析

7.1　正準相関分析とは

変数間の相関は、相関係数や順位相関係数で与えられるが、これを拡張して、変数の組と変数の組の間の相関を考える手法を正準相関分析という。それぞれの変数の組で正準変数と呼ばれる線形関数を作り、その正準変数の相関が最も高くなるように、線形結合の係数を決定する。例えば以下のように、変数 1、変数 2、変数 3 で 1 つの正準変数 y を作り、変数 4 と変数 5 でもう 1 つの正準変数 z を作って、それらの相関係数が最大となるように係数を決める。

$y = a_1$ 変数 $1 + a_2$ 変数 $2 + a_3$ 変数 3

$z = b_1$ 変数 $4 + b_2$ 変数 5

この y と z の相関係数を正準相関係数というが、少ない方の変数の個数だけ（この場合 2 個）正準相関係数が求められる。複数の正準相関係数に対応する正準変数が、どの程度の相関を説明するかを見るためには、寄与率の値を見る。異なる正準相関係数に属する正準変数の相関は 0 である。

正準変数と同じ組の変数との相関係数を正準負荷量、正準変数と違う組の変数との相関係数を交差負荷量という。正準負荷量は正準変数の解釈に用いられる。

以下の例を用いて正準相関分析の解釈をまとめてみよう。

例　表 1 の正準相関分析 1.txt のデータを用いて、身長と座高及び、体重と胸囲の間で相関の高い特徴的な量を求めよ。

表 1　正準相関データ

身長	座高	体重	胸囲
148	78	41	72
160	86	49	77
159	86	45	80
153	83	43	76
:	:	:	:
148	78	38	70

まとめ

体形に関する変数、身長、体重、胸囲、座高について、体の縦方向の大きさを表す変数、身長・座高と体の横方向の大きさを表す変数、体重・胸囲に分け、それらでどのような特徴的な量が作れ、またそれらの間に最大どれだけの相関があるか正準相関分析を用いて調べる。

これらの変数の 1 次関数が最大の相関を持つようにするには、2 つの正準変数を以下のようにおけばよい。その際の正準相関係数の値は、0.894 である。これは 1 次元目の正準相関係数と呼ばれる。

y = 0.402 身長 ＋0.618 座高

z = 1.111 体重 －0.125 胸囲

さらに高い次元での正準相関係数を考えるかどうかを見るために、1 次元の正準相関係数

の寄与率を見ると値が 0.959 であるので、１次元だけで十分な寄与率が与えられたことが分かる。

7.2　プログラムの利用法

メニュー［分析－多変量解析他－関係分析手法－正準相関分析］を選択すると、正準相関分析の実行画面が図１のように表示される。

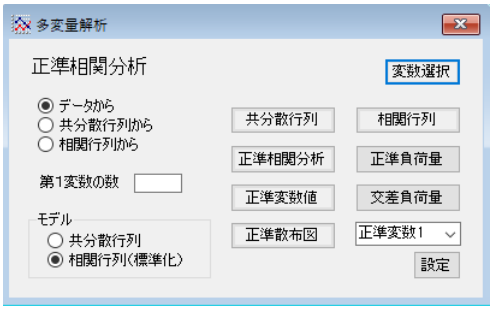


図１　正準相関分析実行画面

分析は、主成分分析等と同様、元データ、分散共分散行列、相関行列から実行できるが、正準変数の値と正準変数の散布図については、当然元データがないと求められない。計算のモデルは、データをそのまま利用する場合と、標準化して相関行列を用いて計算する場合のどちらかを選ぶようになっている。直感的に分り易いのはそのままの値を利用するものであるが、変数の大きさが大きく異なる場合や係数から重要性を読み取ろうとする場合には標準化した方がよい。図２は５つの変数を、３つと２つに分け、「正準相関分析」ボタンをクリックした出力画面（正準相関分析 2.txt）である。

図２　正準相関分析出力画面

この場合、変数を図２の並びの順に選択し、「第１変数」に含まれる変数の数として３を指定する。結果は２つの正準変数の値と２つの正準相関係数の値を表示する。

次に図３に「正準変数値」ボタンをクリックした場合の実行結果を示す。

図３　正準変量値出力画面

ここでは各個体ごとの正準変数の値が表示されているが、これは標準化されたデータを元にしており、結果は標準化された値となっている。これらのデータから第 1 正準変数について散布図を作ったものが図 4 である。正準変数の選択は「設定」ボタンでできる。

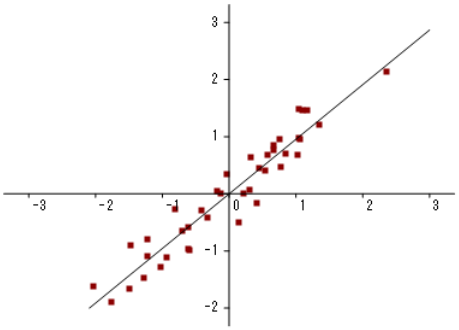


図 4　正準変数の散布図

第 1 正準変数のうちの一方を横軸に、もう一方を縦軸にとっているが、相当高い正準相関係数になることが見て取れる。

正準変数と、それと同じ組の変数との間の相関係数を正準負荷量という。「正準負荷量」ボタンをクリックすると、正準負荷量と各正準変量の寄与率が図 5 のように表示される。

正準負荷量

	正準負荷量1	正準負荷量2
1群		
寄与率	0.7944	0.0973
英語	0.9953	-0.0694
国語	0.9025	-0.4291
社会	0.7604	-0.3207
2群		
寄与率	0.8659	0.1341
数学	0.9793	-0.2025
理科	0.8791	0.4766

図 5　正準負荷量

正準変数と、それと違う組の変数との間の相関係数を交差負荷量という。「交差負荷量」ボタンをクリックすると、交差負荷量の値が図 6 のように表示される。

交差負荷量

	交差負荷量	交差負荷量
正準変数1		
冗長性係数	0.7914	0.0121
数学	0.9362	-0.0608
理科	0.8404	0.1432
正準変数2		
冗長性係数	0.7261	0.0088
英語	0.9515	-0.0208
国語	0.8628	-0.1289
社会	0.7270	-0.0963

図 6 交差負荷量

問題（正準相関分析 2.txt）

正準相関分析 2.txt について、文系科目（英語・国語・社会）と理系科目（数学・理科）に分け、正準相関分析を実行し、以下の問いに答えよ。但し、相関行列を用いたモデルで、第 1 正準変数について考えること。

１）文系科目と理系科目の正準相関係数はいくらか。［　　］

２）文系科目と理系科目の正準変数はそれぞれどのように表されるか。

文系正準変数＝［　　］英語＋［　　］国語＋［　　］社会

理系正準変数＝［　　］数学＋［　　］理科

３）各変数の正準負荷量の値はいくらか。

英語	国語	社会	数学	理科

４）各変数の交差負荷量の値はいくらか。

数学	理科	英語	国語	社会

５）各正準変数と最も相関のある同じ組の科目は何か。

文系正準変数では［英語・国語・社会］、理系正準変数では［数学・理科］

６）各正準変数と最も相関のある違う組の科目は何か。

文系正準変数へは［数学・理科］、理系正準変数へは［英語・国語・社会］

７）各科目の平均と標準偏差（不偏分散からのもの）を求め、

標準化変数＝（値－平均）／標準偏差

の式によって、英語 60、国語 72、社会 66、数学 58、理科 55 の人の標準化変数値を求めよ。

科目	英語	国語	社会	数学	理科
標準化変数値					

８）上の標準化値を利用して、この人の正準変数の値を求めよ。

文系正準変数値［　　］　理系正準変数値［　　］

問題解答（正準相関分析 2.txt）

１）文系科目と理系科目の正準相関係数はいくらか。［ 0.956 ］

２）文系科目と理系科目の正準変数はそれぞれどのように表されるか。

文系正準変数＝［ 1.193 ］英語＋［ −0.081 ］国語＋［ −0.149 ］社会

理系正準変数＝［ 0.739 ］数学＋［ 0.314 ］理科

３）各変数の正準負荷量の値はいくらか。

英語	国語	社会	数学	理科
0.995	0.903	0.760	0.979	0.879

４）各変数の交差負荷量の値はいくらか。

数学	理科	英語	国語	社会
0.936	0.840	0.952	0.863	0.727

５）各正準変数と最も相関のある同じ組の科目は何か。

文系正準変数では［英語・国語・社会］、理系正準変数では［数学・理科］

６）各正準変数と最も相関のある違う組の科目は何か。

文系正準変数へは［数学・理科］、理系正準変数へは［英語・国語・社会］

７）各科目の平均と標準偏差（不偏分散からのもの）を求め、
標準化変数＝（値－平均）／標準偏差　の式によって、英語 60、国語 72、社会 66、数学 58、理科 55 の人の標準化変数値を求めよ。

科目	英語	国語	社会	数学	理科
標準化変数値	−0.404	−0.229	−0.303	−0.006	−0.145

８）上の標準化値を利用して、この人の正準変数の値を求めよ。
文系正準変数値［ −0.418 ］　理系正準変数値［ −0.050 ］

7.3　正準相関分析の理論

正準相関分析は変数 $x_1, x_2, \cdots, x_r$ と変数 $y_1, y_2, \cdots, y_s$ を含む２群間の相関係数を、これらの変数を用いた 1 次関数間の相関係数と定義し、この相関係数が最大となるように係数を決める手法である。

まず、以下のような線形結合により、新しい変数 u，v を考える。

$$u = {}^t\mathbf{ax}, \quad v = {}^t\mathbf{by},$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_r \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_s \end{pmatrix}$$

ここに、$\mathbf{a}$，$\mathbf{b}$ は係数ベクトルである。

変数 $x_1, x_2, \cdots, x_r$ と変数 $y_1, y_2, \cdots, y_s$ の分散共分散行列をそれぞれ $\mathbf{S}_{xx}$，$\mathbf{S}_{yy}$ とし、2 組の変数間の分散共分散行列を $\mathbf{S}_{xy}$（$\mathbf{S}_{yx} = {}^t\mathbf{S}_{xy}$）とすると、$u$ と v の相関係数 r_{uv} は以下となる。

$$r_{uv} = {}^t\mathbf{aS}_{xy}\mathbf{b}$$

但し係数ベクトルは u，v の分散が 1 になるように ${}^t\mathbf{aS}_{xx}\mathbf{a} = 1$，${}^t\mathbf{bS}_{yy}\mathbf{b} = 1$ と規格化している。

制約条件 ${}^t\mathbf{aS}_{xx}\mathbf{a} = 1$，${}^t\mathbf{bS}_{yy}\mathbf{b} = 1$ を入れ、Lagrange の未定定数法を用いて r_{uv} が最大となるように係数を求めると、以下の固有値問題に帰着する。

$$\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{a} = \rho^2\mathbf{a}, \quad {}^t\mathbf{aS}_{xx}\mathbf{a} = 1,$$

$$\mathbf{b} = \frac{1}{\rho}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{a}$$

ここに ρ は未定定数であるが、r_{uv} に等しいことが上の計算過程から分かっており、最大の相関係数の 2 乗は最大の固有値に等しい。この固有値に対応する固有ベクトル $\mathbf{a}$，$\mathbf{b}$ で決まる変数 u，v を（第１）正準変数、その時の相関係数を（第１）正準相関係数という。これに倣って α 番めに大きい固有値に対応する固有ベクトルから同様に求まるものをそれぞれ第 α 正準変数、第 α 正準相関係数という。

個体（レコード）λ について、変数 x_i のデータを $x_{i\lambda}$，変数 y_j のデータを $y_{j\lambda}$ とするとこの個体の正準変数値 u_λ，v_λ は以下のように与えられる。

$$u_\lambda = \sum_{i=1}^{r} a_i x_{i\lambda}, \quad v_\lambda = \sum_{j=1}^{s} b_j y_{j\lambda}$$

ここでは元のデータから分散共分散行列を用いて求める方法を示したが、変数の大きさ（ばらつき）に極端な差があるときは、各変数を標準化して相関行列から同様の計算を進める。

正準変数uと変数x_iとの相関係数r_{ui}、正準変数vと変数y_jとの相関係数r_{vj}を正準負荷量という。正準負荷量を使った以下の定義を寄与率P_u, P_vという。

$$P_u = \sum_{i=1}^{r} r_{ui}^2 / r\ , \quad P_v = \sum_{j=1}^{s} r_{vj}^2 / s$$

正準変数uと変数y_jとの相関係数r_{uj}、正準変数vと変数x_iとの相関係数r_{vi}を交差負荷量という。公差負荷量を使った以下の定義を冗長性係数Q_u, Q_vという。

$$Q_u = \sum_{j=1}^{s} r_{uj}^2 / s\ , \quad Q_v = \sum_{i=1}^{r} r_{vi}^2 / r$$

８．数量化Ｉ類

8.1　数量化Ｉ類とは

ここでは数量化Ｉ類について、以下の例を用いて説明する。

例

２つの地域（1：都市部、2：山村部）と、３つの気候（1：温暖、2：平均的、3：寒冷）の条件で、ある商品の販売率を求めたところ表 1 に示す数量化Ｉ類 1.txt のデータを得た。販売率（目的変数）を地域と気候で予測する式を作り、それがどの程度有効か検討せよ。

表 1　数量化Ｉ類データ

販売率	地域	気候
3.0	1	2
1.8	2	1
1.5	2	2
:	:	:
2.3	1	3

この例題の解答はこの節の最後に示す。

この地域と気候の項目をアイテムと呼び、その中の分類をカテゴリと呼ぶ。数量化Ｉ類ではこれらのアイテムのデータから、アイテムを複数のカテゴリに分けて、表 2 の形の 0/1 データを作る。

表 2　数量化Ｉ類 0/1 データ

販売率	地域 1	地域 2	気候 1	気候 2	気候 3
3.0	1	0	0	1	0
1.8	0	1	1	0	0
1.5	0	1	0	1	0
:	:	:	:	:	:
2.3	1	0	0	0	1

このカテゴリ i、アイテム j として分けたデータを x_{ij} として、以下の式で目的変数を予測する。

$$Y = a_{11}x_{11} + a_{12}x_{12} + a_{21}x_{21} + a_{22}x_{22} + a_{23}x_{23} + a_{00} \qquad (1)$$

但し、このデータはアイテム毎のカテゴリの和が 1 になっているため、すべてのデータが独立ではなく、アイテム毎の独立な変数の数は 1 つ減る。そのため、カテゴリウェイトと呼ばれる係数 a_{ij} も完全に決まるのではなく、アイテム毎に値をずらして与えることができる。その自由度を用いて、カテゴリウェイトもいくつかの定義がある。

まず、普通にカテゴリウェイトと呼ばれるものは、2 番目以降のアイテムの第 1 カテゴリの係数を 0 にして、定数項が 0 になるように定義されている。次に重回帰カテゴリウェイトと呼ばれるものは、すべてのアイテムの第 1 カテゴリを 0 にしている。また、基準化カテゴリウェイトと呼ばれるものは、定数項が目的変数の平均値になるようにしている。基準化カテゴリウェイトでは、係数の符号で、あるカテゴリが目的変数を上げる方に効くのか、下げる方に効くのか分かるので、目的変数への影響を見る際には非常に便利である。

重回帰ウェイトは、各アイテムの第 1 カテゴリを取り除いて、0/1 データを使って重回帰

分析にかけたものと同じ結果を与える。これを利用すると、重回帰分析に質的な目的変数を加えることも可能で、その際には 0/1 データに変換し、第１カテゴリを取り除いてやればよいことが分かる。

重回帰分析に 5 段階評価の結果を数値として加えるべきか、数量化Ｉ類の方法で加えるべきか考えることがあるかも知れないが、数量化Ｉ類の方法で加える方が、変数数が増える分だけ、一般に良い結果が得られる。しかし、そのメリットと扱いが厄介になるデメリットを考えれば、著者なら数値で加える方を選ぶだろう。

例題の予測式の係数値を表 3 として示す。寄与率は 0.937 と非常に高く、良い分析結果を与えている。

表３　例題の係数値

	地域:1	地域:2	気候:1	気候:2	気候:3	定数項
カテゴリウェイト	3.517	1.892	0.000	-0.375	-1.467	0.000
重回帰　ウェイト	0.000	-1.625	0.000	-0.375	-1.467	3.517
基準化　ウェイト	0.488	-1.138	0.699	0.324	-0.767	2.330
重相関係数	0.968					
寄与率	0.937					

8.2　プログラムの利用法

メニュー［分析－多変量解析他－予測手法－数量化Ｉ類］を選択して表示される実行画面を図１に与える。

図１　数量化Ｉ類実行画面

入力にはアイテム毎にカテゴリ名が記されているものとアイテム内をカテゴリ数に分け 0/1 で回答を表わしたものの 2 種類のデータが利用できる。もちろん 0/1 で表わされたデータには、アイテム毎のカテゴリ数を与える必要があり、テキストボックス内にカンマ区切りで入力する。コマンドボタン「0/1 型への変換」ではカテゴリ名データからもう１つの入力型である 0/1 型データに変換する。出力結果を図 2 に示す。

データ基本型

	販売率	地域:1	地域:2	気候:1	気候:2	気候:3
1	3.0	1	0	0	1	0
2	1.8	0	1	1	0	0
3	1.5	0	1	0	1	0
4	3.3	1	0	0	1	0
5	2.2	1	0	0	0	1
6	2.0	1	0	0	0	1
7	3.5	1	0	1	0	0
8	2.0	0	1	1	0	0
9	1.7	1	0	0	0	1
10	2.3	1	0	0	0	1

図２　0/1 型データへの変換

カテゴリウェイトの値はコマンドボタン「カテゴリウェイト」をクリックすることによって得られる。また、これらの値による予測値から得られる重相関係数と寄与率も与えられる。出力画面を図３に示す。

カテゴリウェイト

	カテゴリウェイ	重回帰ウェイ	基準化ウェイ
地域:1	3.5167	0.0000	0.4875
地域:2	1.8917	-1.6250	-1.1375
気候:1	0.0000	0.0000	0.6992
気候:2	-0.3750	-0.3750	0.3242
気候:3	-1.4667	-1.4667	-0.7675
定数項	0.0000	3.5167	2.3300
重相関R	0.968	調整済R	0.951
寄与率R^2	0.937	調整済R^2	0.905
有効性F値	29.621	自由度	3,6
参考p値	0.0005		

図３　カテゴリウェイト

ここでは定数項を 0 としたカテゴリウェイトの他に、各アイテムのカテゴリの影響の正負がはっきり分かる基準化カテゴリウェイトや、各アイテムの第 1 カテゴリを 0 とした重回帰ウェイトが求められる。重回帰ウェイトは 0/1 データから、第 1 カテゴリを 0 として、重回帰分析を実行した結果と同じになる。有効性 F 値は、回帰式の有効性の検定の検定値である。これを F 分布と仮定した場合の上側確率の値を参考 p 値として与えてある。

目的変数とアイテム間の相関行列、目的変数とアイテム間の偏相関係数、ウェイト範囲、変数の重要性の F 値等は「アイテム重要性」ボタンをクリックすることにより図 4 のように表示される。重要性 F 値についても参考のため F 分布の際の上側確率を与えている。

アイテム重要性

	販売率	地域	気候
販売率	1.0000	0.5584	0.3152
地域	0.5584	1.0000	-0.5843
気候	0.3152	-0.5843	1.0000
ウェイト範囲		1.6250	1.4667
偏相関係数		0.9642	0.9529
重要性F値		76.5861	29.6388
自由度		1,6	2,6
参考p値		0.0001	0.0008

図４　アイテム重要性

目的変数に対する予測値と残差は「予測値と残差」ボタンで図 5 のように与えられ、その「散布図」を図 6 のように与えられる。


予測値と残差

	観測値	予測値	残差
1	3.0	3.142	-0.142
2	1.8	1.892	-0.092
3	1.5	1.517	-0.017
4	3.3	3.142	0.158
5	2.2	2.050	0.150
6	2.0	2.050	-0.050
7	3.5	3.517	-0.017
8	2.0	1.892	0.108
9	1.7	2.050	-0.350
10	2.3	2.050	0.250

図５　予測値と残差

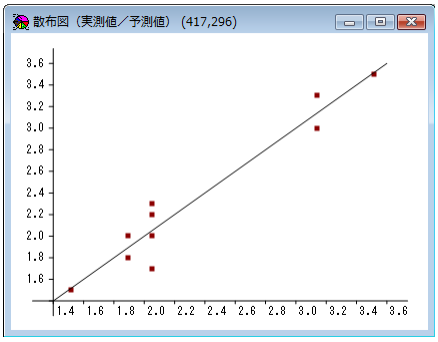


図６　予測値と実測値の散布図

予測について

機械学習などでは､訓練データに対して独立なテストデータを用いて目的変数の予測を行う。図 1 の分析実行画面の一番下に「新データ予測」ボタンがあるが、これは一度数量化Ｉ類を実行した後、新たなデータを選択して予測を行うためのボタンである。テストデータに目的変数を含む場合は「目的変数付き」チェックボックスにチェックを入れて、「新データ予測」ボタンをクリックする。テストデータで予測値を計算し、適合性を R^2 の値で示してくれる。これはデータ形式として「分類名データから」のとき有効である。

参考文献

[1] 河口至商，多変量解析入門Ⅱ，森北出版, 1978.
[2] 永田靖・棟近雅彦，サイエンス社, 2001.

問題（数量化Ｉ類 2.txt）

数量化Ｉ類 2.txt は店舗の売り上げを立地、人通り、競合の 3 段階分類データで予測しようとするものである。

１）カテゴリウェイト（定数項を 0）を用いた予測式を表せ。

予測売り上げ＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］競合３

２）重回帰カテゴリウェイト（各先頭アイテムを基準）を用いた予測式を表せ。

予測売り上げ＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］競合３＋［　　］

３）基準化カテゴリウェイトを用いた（目的変数の平均値を基準）予測式を表せ。

予測売り上げ＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］競合３＋［　　］

４）予測式は実測値の変動を何％予測できるか。［　　］％

５）立地：2，人通り：2，競合：2 の店舗の売り上げを予測せよ。［　　］

６）ウェイト範囲で見る場合、予測値に最も大きな影響を与えるアイテムは何か。
［立地・人通り・競合］

７）数量化Ｉ類と同じ分析を 0/1 データを用いた重回帰分析で行った。但し、各アイテムの第１カテゴリは係数が 0 として、変数から外した。そのときの重回帰式を示せ。
予測売り上げ＝［　　］立地２＋［　　］立地３
＋［　　］人通り２＋［　　］人通り３
＋［　　］競合２＋［　　］競合３＋［　　］

８）このことから上の重回帰分析と数量化Ｉ類は［同じ・異なる］ものと考えられる。

問題解答（数量化Ｉ類 2.txt）

１）カテゴリウェイト（定数項を 0）を用いた予測式を表せ。
予測売り上げ＝［ 3745 ］立地１＋［ 4171 ］立地２＋［ 4187 ］立地３
＋［ 0 ］人通り１＋［ 237 ］人通り２＋［ 612 ］人通り３
＋［ 0 ］競合１＋［ -457 ］競合２＋［ -454 ］競合３

２）重回帰カテゴリウェイト（各先頭アイテムを基準）を用いた予測式を表せ。
予測売り上げ＝［ 0 ］立地１＋［ 425 ］立地２＋［ 441 ］立地３
＋［ 0 ］人通り１＋［ 237 ］人通り２＋［ 612 ］人通り３
＋［ 0 ］競合１＋［-457 ］競合２＋［-454］競合３＋［ 3745 ］

３）基準化カテゴリウェイトを用いた（目的変数の平均値を基準）予測式を表せ。
予測売り上げ＝［ -290 ］立地１＋［ 135 ］立地２＋［ 151 ］立地３
＋［ -290 ］人通り１＋［ -53 ］人通り２＋［ 322 ］人通り３
＋［ 334 ］競合１＋［ -123 ］競合２＋［ -120 ］競合３＋［ 3991 ］

４）予測式は実測値の変動を何%予測できるか。［ 73.8 ］%

５）立地：2，人通り：2，競合：2 の店舗の売り上げを予測せよ。［ 3951 ］

６）ウェイト範囲で見る場合、予測値に最も大きな影響を与えるアイテムは何か。
［立地・人通り・競合］

７）数量化Ｉ類と同じ分析を 0/1 データを用いた重回帰分析で行った。但し、各アイテムの第１カテゴリは係数が 0 として、変数から外した。そのときの重回帰式を示せ。
予測売り上げ＝［ 425 ］立地２＋［ 441 ］立地３
＋［ 237 ］人通り２＋［ 612 ］人通り３
＋［ -457 ］競合２＋［ -454 ］競合３＋［ 3745 ］

８）このことから上の重回帰分析と数量化Ｉ類は［同じ・異なる］ものと考えられる。

8.3　数量化Ｉ類の理論

数量化Ｉ類は、目的変数をカテゴリデータから推測する手法で、量的データの重回帰分析に相当する。数量化Ｉ類の変数は目的変数とアイテム毎に複数個含まれるカテゴリ変数からなる。データの基本的な形は表 1 に示される。カテゴリデータは各アイテム中の１つのカテゴリを選択するようになっており、選択された値が 1 で、他の値が 0 であるように定められている。これはデータの一般的な書式 $x_{ij\lambda}$ を用いて以下のように表わすこともできる。

$$x_{ij\lambda} \in \{0,1\},\ \sum_{j=1}^{r_i} x_{ij\lambda} = 1$$

表１　数量化Ｉ類のデータ

目的変数	アイテム１				アイテム p		
	カテゴリ１	…	カテゴリ r_1	…	カテゴリ１	…	カテゴリ r_p
y_1	x_{111}	…	$x_{1r_1 1}$	…	x_{p11}	…	$x_{pr_p 1}$
y_2	x_{112}	…	$x_{1r_1 2}$	…	x_{p12}	…	$x_{pr_p 2}$
：	：		：		：		：
y_n	x_{11n}	…	$x_{1r_1 n}$	…	x_{p1n}	…	$x_{pr_p n}$

これより全カテゴリ数 r_c は以下で与えられる。

$$r_c = \sum_{i=1}^{p} r_i$$

目的変数は第２アイテム以降の第１カテゴリを除いた、以下の式で予測される。

$$Y_\lambda = \sum_{j=1}^{r_1} \hat{a}_{1j} x_{1j\lambda} + \sum_{i=2}^{p} \sum_{j=2}^{r_i} \hat{a}_{ij} x_{ij\lambda}$$

ここに、係数 $\hat{a}_{ij}$ は以下の残差変動 EV を最小化するように求める。後に述べるが係数はすべて独立ではない。アイテム内の係数の１つは他の係数で求めることができる。それにより係数の数 r_d は以下で与えられる。

$$r_d = r_c - p$$

残差変動 EV の係数 $\hat{a}_{ij}$ についての微分を０として、以下の解を得る。

$$EV = \sum_{\lambda=1}^{n} (y_\lambda - Y_\lambda)^2 \qquad \rightarrow \qquad \hat{\mathbf{a}} = ({}^t\mathbf{X}\mathbf{X})^{-1}{}^t\mathbf{X}\mathbf{y}$$

ここに、各行列やベクトルは以下のように定義されるが、第２アイテム以降の第１カテゴリを外しているのは、行列 ${}^t\mathbf{X}\mathbf{X}$ の正則性を失わせないためである。

$${}^t\hat{\mathbf{a}} = (\hat{a}_{11} \quad \cdots \quad \hat{a}_{1r_1} \quad \hat{a}_{22} \quad \cdots \quad \hat{a}_{2r_2} \quad \cdots \quad \hat{a}_{p2} \quad \cdots \quad \hat{a}_{pr_p})$$

$${}^t\mathbf{y} = (y_1 \quad y_2 \quad \cdots \quad y_n)$$

$$\mathbf{X} = \begin{pmatrix} x_{111} & \cdots & x_{1r_1 1} & x_{221} & \cdots & x_{2r_2 1} & \cdots & x_{p21} & \cdots & x_{pr_p 1} \\ x_{112} & \cdots & x_{1r_1 2} & x_{222} & \cdots & x_{2r_2 2} & \cdots & x_{p22} & \cdots & x_{pr_p 2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{11n} & \cdots & x_{1r_1 n} & x_{22n} & \cdots & x_{2r_2 n} & \cdots & x_{p2n} & \vdots & x_{pr_p n} \end{pmatrix}$$

また、この係数は、

$$Y_\lambda = \sum_{i=1}^{p} \sum_{j=2}^{r_i} \hat{a}_{ij} x_{ij\lambda} + \hat{a}_0$$

として通常の重回帰分析の手法で求めることもできる。もちろん値は前のものと異なる。

ここで係数の自由度について考えてみる。

アイテム数を p 個、第 i のアイテムのカテゴリ数を r_i 個とし、第 i アイテムの第 k カテゴリ、レコード λ のデータを $x_{i(k)\lambda}=\{0,1\}$ とし、数量化Ｉ類の予測式が以下で与えられたとする。

$$y_\lambda=\sum_{i=1}^{p}\sum_{k=1}^{r_i}b_{i(k)}x_{i(k)\lambda}+b_0\ ,\quad \sum_{k=1}^{r_i}x_{i(k)\lambda}=1$$

この式から、以下の関係も与えられる。

$$\bar{y}=\sum_{i=1}^{p}\sum_{k=1}^{r_i}b_{i(k)}\bar{x}_{i(k)}+b_0$$

この係数（カテゴリウェイト）には以下の自由度が存在する。

$$b'_{i(k)}=b_{i(k)}-c_i\ ,\ \ b'_0=b_0+\sum_{i=1}^{p}c_i$$

なぜなら、

$$\sum_{i=1}^{p}\sum_{k=1}^{r_i}b'_{i(k)}\,x_{i(k)\lambda}+b'_0=\sum_{i=1}^{p}\sum_{k=1}^{r_i}\left(b_{i(k)}-c_i\right)x_{i(k)\lambda}+b_0+\sum_{i=1}^{p}c_i=\sum_{i=1}^{p}\sum_{k=1}^{r_i}b_{i(k)}x_{i(k)\lambda}+b_0$$

この解に対して代表的なカテゴリウェイトを作ってみる。

重回帰ウェイト

$c_i=b_{i(0)}$

これにより、$b'_{i(0)}=0$ となる。

通常のカテゴリウェイト

$$c_1=-b_0-\sum_{i=2}^{p}c_i\ ,\ c_i=b_{i(0)}\ \left(i\neq1\right)$$

これにより、$b'_0=0,\ b'_{i(0)}=0\ \left(i\neq1\right)$ となる。

基準化ウェイト（これが最も重要である）

$$c_i=\sum_{k=1}^{r_i}b_{i(k)}\bar{x}_{i(k)}$$

これにより、

$$\sum_{i=1}^{p}c_i=\sum_{i=1}^{p}\sum_{k=1}^{r_i}b_{i(k)}\bar{x}_{i(k)}=\bar{y}$$

となり、予測式は以下となる。

$$y_\lambda=\sum_{i=1}^{p}\sum_{k=1}^{r_i}b'_{i(k)}\,x_{i(k)\lambda}+\bar{y}$$

これは $b'_{i(k)}$ が目的変数を平均より上げるか下げるか分かるようになる。

分析の寄与率 R^2（重相関係数 R）、自由度調整済み寄与率 R^{*2}（自由度調整済み重相関係数 R^*）は、以下のように全変動 SV、回帰変動 RV、残差変動 EV、係数の数 r_d（全

カテゴリ数－全アイテム数）を用いて以下のように与えられる。

$$SV=\sum_{\lambda=1}^{n}(y_\lambda-\bar{y})^2=\sum_{\lambda=1}^{n}(y_\lambda-Y_\lambda)^2+\sum_{\lambda=1}^{n}(Y_\lambda-\bar{y})^2=EV+RV$$

$$R^2=RV/SV=1-EV/SV\,,\quad R^{*2}=1-\frac{EV/(n-r_d-1)}{SV/(n-1)}$$

回帰式の有効性の F 値は回帰変動と残差変動を比べて、回帰変動が十分大きいことが重要で、この検定には、以下の性質が利用される。

$$F=\frac{RV/r_d}{EV/(n-r_d-1)}\sim F_{r_d,n-r_d-1}\text{分布}$$

各アイテムと目的変数の共分散行列s_{ij},s_{iy},s_{yy}を以下で定義する。

$$s_{ij}=\frac{1}{n-1}\sum_{\lambda=1}^{n}(X_{i\lambda}-\bar{X}_i)(X_{j\lambda}-\bar{X}_j)\,,\quad s_{iy}=\frac{1}{n-1}\sum_{\lambda=1}^{n}(X_{i\lambda}-\bar{X}_i)(y_\lambda-\bar{y})\,,$$

$$s_{yy}=\frac{1}{n-1}\sum_{\lambda=1}^{n}(y_\lambda-\bar{y})^2$$

ここに、アイテムiの予測値$X_{i\lambda}$及びその平均$\bar{X}_i$は以下で与えられる。

$$X_{i\lambda}=\sum_{j=1}^{r_i}\tilde{a}_{ij}x_{ij\lambda}\,,\quad \bar{X}_i=\frac{1}{n}\sum_{\lambda=1}^{n}X_{i\lambda}$$

上で定義した共分散行列を用いた相関行列$\mathbf{R}$の逆行列$\mathbf{R}^{-1}$の成分r^{ij},r^{iy},r^{yy}から、アイテムiと目的変数との偏相関係数$\tilde{r}_{iy}$は以下のように求められる。

$$\tilde{r}_{iy}=-r^{iy}\Big/\sqrt{r^{ii}r^{yy}}$$

アイテムの重要性を調べるために、p個のアイテムに 1 つ付け加える場合を考える。全変動SV、p個のアイテムの回帰変動RV、p個のアイテムの残差変動EV、係数の数r_d、$p+1$個のアイテムの回帰変動RV'、残差変動EV'、係数の数r'_dを用いて、付け加えるアイテムの重要性の F 値は以下となる。

$$F=\frac{(EV-EV')/(r'_d-r_d)}{EV'/(n-r'_d-1)}\qquad \text{自由度}\ \ r'_d-r_d\,,\ n-r'_d-1$$

また、p個のアイテムの数量化 I 類による式の有効性の F 値は以下となる。

$$F=\frac{RV/r_d}{EV/(n-r_d-1)}\qquad \text{自由度}\ \ r_d\,,\ n-r_d-1$$

９．数量化Ⅱ類

9.1　数量化Ⅱ類とは

数量化Ⅱ類は判別分析と同じく、個体（レコード）の判別を 2 群の場合は判別の分点で、3 群以上の場合は、判別関数の大きさで分けるようにする分析手法である。ここでは数量化Ⅱ類について、以下の例を用いて説明する。

例

顧客が車を購入する際、3 種類の特性について検討し、a かｂの車種を購入した（数量化Ⅱ類 1.txt)。顧客がどのような選択を行うかでどちらの車を購入するか判別する式を作り、判別の程度を検討せよ。

表１　数量化Ⅱ類データ

群	価格	外観	性能
A	1	1	2
A	2	1	1
A	1	2	1
A	2	2	1
B	1	1	3
:	:	:	:
B	2	1	3

数量化Ⅱ類でも、数量化Ｉ類と同様、上のようなアイテムのデータから、表 2 の形のカテゴリごとに分けた 0/1 データを作る。

表２　数量化Ⅱ類 0/1 データ

群	価格:1	価格:2	外観:1	外観:2	性能:1	性能:2	性能:3
A	1	0	1	0	0	1	0
A	0	1	1	0	1	0	0
A	1	0	0	1	1	0	0
A	0	1	0	1	1	0	0
B	1	0	1	0	0	0	1
:	:	:	:	:	:	:	:
B	0	1	1	0	0	0	1

このデータのアイテム i、その中のカテゴリ j のデータを x_{ij} として、どちらの群に属するかの判別式を以下のように作成する。

$$y = a_{11}x_{11} + a_{12}x_{12} + a_{21}x_{21} + a_{22}x_{22} + a_{31}x_{31} + a_{32}x_{32} + a_{33}x_{33} + a_{00}$$

ここで、a_{ij} はカテゴリウェイトと呼ばれるパラメータである。このデータには、数量化Ｉ類のところでも述べたように、アイテム毎のカテゴリを加えると 1 になるという性質があり、すべての変数が独立ではない。そのため、カテゴリウェイトにもアイテム毎に値をずらす自由度が存在する。この自由度を用いていくつかの種類のカテゴリウェイトが定義されている。

通常のカテゴリウェイトは、2 群の場合でも 3 群以上の場合でも、各アイテムの第 1 カテゴリの係数を 0 においたもので、第 1 カテゴリの変数を除いて判別分析を行った結果と一致するように定義している。また、基準化カテゴリウェイトは、各カテゴリが判別に対して

正に働くか負に働くかをはっきりと示せるようにしたものである。

数量化Ⅱ類ではその他に、判別関数を用いて判別を行った場合の誤判別確率も表示できるようにしておく必要がある。これは、実際にデータに対して判別を行って計算する。

9.2　プログラムの利用法

メニュー［分析－多変量解析他－判別手法－数量化Ⅱ類］を選択すると、数量化Ⅱ類の分析実行画面が図１のように表示される。

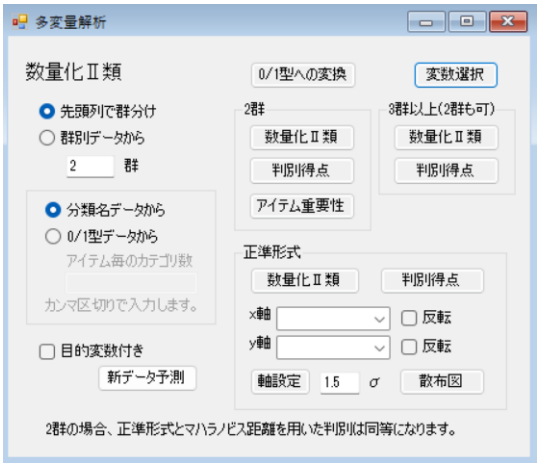


図１　数量化Ⅱ類分析実行画面

データは先頭列で群分けを行なう場合と既に群別になっている場合が取り扱えるが、群別データからの場合は群の数を入力する必要がある。データの形式は各アイテムについてカテゴリ名を与える場合とカテゴリが既に 0/1 データとして分けられている場合があるが、0/1 データの場合、各アイテムのカテゴリ数をカンマ区切りで入力しなければならない。また、計算方式としては、図 1 の上部に示されたマハラノビス形式と下部に示された正準形式のどちらかを選択できる。

マハラノビス形式の結果は、各カテゴリの第１アイテムを除いた変数で判別分析を行った結果と一致する。我々はまず、2 群の場合の結果を比較して、3 群の場合の違いを見ることにする。2 群の場合、2 つの形式で「数量化Ⅱ類」コマンドボタンをクリックした結果を比較する。マハラノビス形式の結果を図 2a に、正準形式の結果を図 2b に与える。

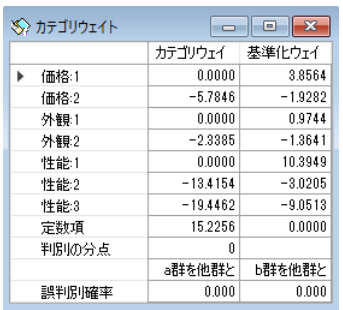


図 2a　マハラノビス形式

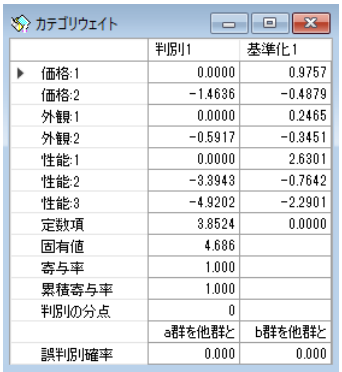


図 2b　正準形式

ここではカテゴリウェイト、基準化されたカテゴリウェイト、判別の分点、誤判別確率が表示される。判別の分点は 0 にしている。正準形式の場合は、固有値と寄与率、累積寄与率が表示されるが、2 群の場合、寄与率と累積寄与率は 1 になる。マハラノビス形式と正準形式の 2 つのカテゴリウェイトはそれぞれ比例している。

2 群の場合、2 つの方法は同等であるので、以後はマハラノビス形式の結果のみを表示する。「アイテム重要性」ボタンをクリックすると、図 3 のような結果が表示される。

アイテム重要性

	サンプル	価格	外観	性能
▶ サンプル	1.0000	0.1905	−0.0891	0.4541
価格	0.1905	1.0000	0.0891	0.0685
外観	−0.0891	0.0891	1.0000	−0.3607
性能	0.4541	0.0685	−0.3607	1.0000
ウェイト範囲		5.7846	2.3385	13.4154
偏相関係数		0.1703	0.0696	0.4441
F値		2.2656	0.3703	9.3462
自由度		1,5	1,5	2,5
参考p値		0.1926	0.5694	0.0205

図 3 アイテム重要性

ここでは、各個体の 0/1 のデータにカテゴリウェイトを掛けてアイテムごとに足した値をアイテムの値とし、その合計の判別関数値と合わせて、これらのアイテムの値と判別関数値間の相関係数を計算した相関行列が一番上に表示されている。また、その相関行列とそれを元に計算される偏相関係数（判別関数値と、他のアイテムの影響を除いたあるアイテムの相関係数）及びアイテム毎のカテゴリウェイトの最大と最小の差であるウェイト範囲が表示される。ウェイト範囲は各アイテムの重要性を見るのに用いられる。またアイテムの重要性を示す F 値等も表示される。データに正規性がないために、F 値の確率は参考 p 値として表示してある。

図 4 は「判別得点」をクリックした場合の結果を表わしている。各個体が元々所属する群とその個体の数量化された値が示される。判別の助けとなるように各群の判別得点の平均や 2 群の場合は判別の分点も示されている。

判別得点

	所属群	判別得点	予測群
▶ 1	a	1.8103	a
2	a	9.4410	a
3	a	12.8872	a
4	a	7.1026	a
5	b	−4.2205	b
6	b	−12.3436	b
7	b	−6.3128	b
8	b	−10.0051	b
9	b	−3.9744	b
10	b	−10.0051	b
群別得点平均	a	7.8103	
	b	−7.8103	
判別の分点		0	

図 4　判別得点

次に 3 群以上の場合について説明する。3 群の場合、正準形式とマハラノビス形式ではかなり異なる。マハラノビス形式では群別の判別関数が出力されるのに対して、正準形式では

固有値に対応する判別関数が出力される。前者はどの判別関数の値が大きいかによって判別結果を決めるが、後者は判別結果を多次元上に表示するためのものである。結果を比較して示しておく。それぞれ、図 5a と図 5b のように結果が表示される。

カテゴリウェイト

	1群判別関	2群判別関	1群基準化	2群基準化
吐き気:0	0.0000	0.0000	−0.6471	−1.9608
吐き気:1	1.4118	2.8235	0.7647	0.8627
吐き気:2	0.7059	5.4118	0.0588	3.4510
頭痛:0	0.0000	0.0000	−0.2353	−1.1373
頭痛:1	0.9412	1.8824	0.7059	0.7451
頭痛:2	−0.2353	3.5294	−0.4706	2.3922
定数項	−0.2941	−2.5098	0.5882	0.5882
	1群を他群と	2群を他群と		
誤判別確率	0.000	0.167		

図 5a　マハラノビス形式

カテゴリウェイト

	判別1	基準化1
吐き気:0	0.0000	0.7282
吐き気:1	−0.7825	−0.0543
吐き気:2	−2.6084	−1.8802
頭痛:0	0.0000	0.4999
頭痛:1	−0.5217	−0.0217
頭痛:2	−2.0867	−1.5868
定数項	1.2281	0.0000
固有値	0.976	
寄与率	1.000	
累積寄与率	1.000	
判別の分点	0	
	1群を他群と	2群を他群と
誤判別確率	0.000	0.167

図 5b　正準形式

それぞれの方法の「判別得点」をクリックした結果を図 6a と図 6b に示す。

判別得点

	所属群	1群判別得	2群判別得	3群判別得	予測群
1	1	1.5297	−3.7739	−5.5614	1
2	1	2.7328	3.2730	0.4542	2
3	1	−0.5328	−12.7114	−15.8739	1
4	1	−0.5328	−12.7114	−15.8739	1
5	1	−0.5328	−12.7114	−15.8739	1
6	2	4.4516	13.3400	13.7623	3
7	2	3.2484	7.8645	6.1752	2
8	2	4.7953	12.2105	10.7667	2
9	2	4.4516	13.3400	13.7623	3
10	2	5.3109	16.8020	16.4877	2
11	3	4.4516	13.3400	13.7623	3
12	3	4.9672	17.9315	19.4833	3
13	3	4.7953	12.2105	10.7667	2
14	3	4.9672	17.9315	19.4833	3

図 6a　マハラノビス形式の判別得点

判別得点

	所属群	判別得点1	判別得点2
1	1	2.1112	0.6742
2	1	0.9115	−1.6372
3	1	3.8454	0.3633
4	1	3.8454	0.3633
5	1	3.8454	0.3633
6	2	−1.3902	0.5801
7	2	−0.1558	−1.0850
8	2	−0.8227	−1.3263
9	2	−1.3902	0.5801
10	2	−1.8900	−0.7741
11	3	−1.3902	0.5801
12	3	−2.4575	1.1324
13	3	−0.8227	−1.3263
14	3	−2.4575	1.1324
群別得点平均	1	2.9118	0.0254
	2	−1.1298	−0.4050
	3	−1.7820	0.3796

図 6b　正準形式の判別得点

マハラノビス形式では、判別関数の値の最も大きい群に判別されることが示されているが、正準形式では判別結果は明確に示されていない。正準形式では複数の次元の判別点を見て判断を下すため。2 次元上に散布図を描画する機能が付けられている。メニューの「軸設定」で表示する次元を設定し、「散布図」ボタンにより、図 7 のように判別得点を平面上に表示する。図中の楕円は1.5σ を表す楕円である。重なった点が多いため、散布図はあまり見易いとは言えない。

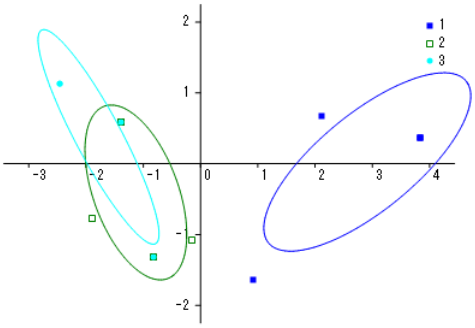


図７　判別得点による散布図

予測について

機械学習などでは、訓練データに対して独立なテストデータを用いて目的変数の予測を行う。図 1 の分析実行画面の一番下に「新データ予測」ボタンがあるが、これは一度数量化Ⅱ類を実行した後、新たなデータを選択して予測を行うためのボタンである。テストデータに目的変数を含む場合は「目的変数付き」チェックボックスにチェックを入れて、「新データ予測」ボタンをクリックする。テストデータで予測値を計算し、適合性を R^2 の値で示してくれる。これはデータ形式として「先頭列で群分け」で「分類名データから」のとき有効である。

問題（数量化Ⅱ類 1b.txt）

数量化Ⅱ類 1b.txt は店舗の成功か失敗かを立地、人通り、競合の 3 段階分類データで予測しようとするものである。

１）成否はある売り上げを境に成功と失敗の 2 つに分けたものであるが（成功 1、失敗 2)、
　カテゴリウェイトを用いた判別関数を表せ。
　判別関数＝［　　］立地１＋［　　］立地２＋［　　］立地３
　＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
　＋［　　］競合１＋［　　］競合２＋［　　］競合３＋［　　］

２）基準化カテゴリウェイトを用いた判別関数を表せ。
　判別関数＝［　　］立地１＋［　　］立地２＋［　　］立地３
　＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
　＋［　　］競合１＋［　　］競合２＋［　　］競合３＋［　　］

３）判別の分点はいくらか。　［　　］

４）誤判別確率はいくらか。
　成功を失敗と誤判別［　　］　失敗を成功と誤判別［　　］

５）成功するためにはどちらが有利か。（ヒント　判別得点が大きな値ほど成功になる）
　［立地１・立地３］　［人通り１・人通り３］　［競合１・競合３］

６）ウェイト範囲で見る場合、判別に最も影響を与えるアイテムは何か。
［立地・人通り・競合］

７）１番の店舗の判別得点はいくらか。それはどう判別されたか。
判別得点［　　］　判定［成功・失敗］

問題解答（数量化Ⅱ類 1b.txt）

１）成否はある売り上げを境に成功と失敗の２つに分けたものであるが（成功 1、失敗 2)、カテゴリウェイトを用いた判別関数を表せ。
判別関数＝［ 0 ］立地１＋［ 3.006 ］立地２＋［ 3.586 ］立地３
＋［ 0 ］人通り１＋［ 1.142 ］人通り２＋［ 2.315 ］人通り３
＋［ 0 ］競合１＋［ −2.749 ］競合２＋［ −4.105 ］競合３＋［ −1.340 ］

２）基準化カテゴリウェイトを用いた判別関数を表せ。
判別関数＝［ −2.429 ］立地１＋［ 0.577 ］立地２＋［ 1.157 ］立地３
＋［ −1.310 ］人通り１＋［ −0.168 ］人通り２＋［ 1.005 ］人通り３
＋［ 2.399 ］競合１＋［ −0.350 ］競合２＋［ −1.707 ］競合３＋［ 0 ］

３）判別の分点はいくらか。　［ 0 ］

４）誤判別確率はいくらか。
成功を失敗と誤判別［ 0.167 ］　失敗を成功と誤判別［ 0.111 ］

５）成功するためにはどちらが有利か。（ヒント　判別得点が大きな値ほど成功になる）
［立地１・立地３］　［人通り１・人通り３］　［競合１・競合３］

６）ウェイト範囲で見る場合、判別に最も影響を与えるアイテムは何か。
［立地・人通り・競合］

７）１番の店舗の判別得点はいくらか。それはどう判別されたか。
判別得点［ −0.717］　判定［成功・失敗］

9.3　数量化Ⅱ類の理論

数量化Ⅱ類はカテゴリデータに関する線形判別関数を定義し、個体を分類することが狙いであり、判別分析に相当する。カテゴリデータで群分類を行なう数量化Ⅱ類は、群の数を m、群 α のデータ数を n_α、アイテム数を p、アイテム i のカテゴリ数を r_i として、表１のデータ形式を元にする。

表１　数量化Ⅱ類のデータ

	アイテム１				アイテム p		
	カテゴリ 1	…	カテゴリ r_1	…	カテゴリ 1	…	カテゴリ r_p
群１	x^1_{111} ⋮ $x^1_{11n_1}$	… …	$x^1_{1r_11}$ ⋮ $x^1_{1r_1n_1}$	…	x^1_{p11} ⋮ $x^1_{p1n_1}$	… …	$x^1_{pr_p1}$ ⋮ $x^1_{pr_pn_1}$
⋮	⋮		⋮		⋮		⋮
群 m	x^m_{111} ⋮ $x^m_{11n_m}$	… …	$x^m_{1r_11}$ ⋮ $x^m_{1r_1n_m}$	…	x^m_{p11} ⋮ $x^m_{p1n_m}$	… …	$x^m_{pr_p1}$ ⋮ $x^m_{pr_pn_m}$

一般にデータを $x^{\alpha}_{ij\lambda} \in \{0,1\}$ の形で表わすと、$\alpha\,(1,2,\cdots,m)$ は群、$\lambda\,(1,2,\cdots,n_\alpha)$ は個体、$i\,(1,2,\cdots,p)$ はアイテム、$j\,(1,2,\cdots,r_i)$ はアイテム毎のカテゴリである。変数には次の関係がある。

$$\sum_{j=1}^{r_i} x^{\alpha}_{ij\lambda} = 1 \qquad (1)$$

このため、アイテムごとに独立なカテゴリの数は 1 つ少なくなる。通常は第 1 カテゴリを除いた変数を用いて分析を実行する。

ここで、$x^{\alpha}_{ij\lambda}$ の表式を判別分析と類似のものとするため、新しい表記として $x^{\alpha}_{I\lambda}$ を導入する。この大文字の $\boldsymbol{I}$ はアイテム i 、その中のカテゴリ $j\,(=2,\cdots,r_i)$ について、順番にアイテム 1 から並べた数で、以下で定義される。

$$I \equiv \sum_{k=1}^{i-1}(r_k - 1) + (j-1), \qquad I = 1,2,\cdots,P \equiv \sum_{k=1}^{p}(r_k - 1)$$

この変数表記法を用いると第 1 カテゴリを除いた数量化Ⅱ類は判別分析と同等であることが理解し易い。以後は以下のように置き換えることによって、両者の書式を使い分けることにする。

$$\sum_{I=1}^{P} f_I \quad \Leftrightarrow \quad \sum_{i=1}^{p}\sum_{j=1}^{r_i} f_{ij}$$

1) マハラノビスの距離に基づく方法

新しい変数表記法 $x^{\alpha}_{I\lambda}$ でデータを見ると 0,1 型のデータであっても、判別分析と同等に扱うことができる。よってデータの判別はマハラノビスの距離に基づく方法を用いて、判別分析と同じように行うことができる。但し、データの分布は正規分布でないので、判別分析の

最初のところで述べた分布関数による判別の理由付けはできない。しかし、2 群の場合は正準形式と同等であるので、判別関数による群間分散の最大化の方法による理由付けは説得力がある。3 群以上の場合は、群間の 1 対比較によって判別を行うものと解釈すると、判別の問題は判別分析と全く同等に考えることができる。

2 群の場合、判別分析と同じように作られた係数を用いて判別関数は以下のように与えられる。ここでは判別関数との類似性を強調するため、新しい変数表示法を用いている。

$$z=\sum_{I=1}^{P}a_I x_I-\frac{1}{2}\sum_{I=1}^{P}(\overline{x}_I^1+\overline{x}_I^2)a_I \quad , \quad a_I=\sum_{J=1}^{P}(\mathbf{S}^{-1})_{IJ}(\overline{x}_J^1-\overline{x}_J^2) \qquad (2)$$

また、3 群以上の場合、群 α の判別関数は以下のように与えられる。

$$z^{\alpha}=\sum_{I=1}^{P}a_I^{\alpha}x_I-\frac{1}{2}\sum_{I=1}^{P}\overline{x}_I^{\alpha}a_I^{\alpha} \quad , \quad a_I^{\alpha}=\sum_{J=1}^{P}(\mathbf{S}^{-1})_{IJ}\overline{x}_J^{\alpha} \qquad (3)$$

2 群の場合も 3 群以上の場合も、係数ベクトル a_{ij} は各アイテムの第 1 カテゴリを除いたものであるので、以下のような基準化された係数 d_{ij} $(i=1,\cdots,p,\ j=1,2,\cdots,r_i)$ も計算しておく。

2 群の場合　　$d_{ij}=\hat{a}_{ij}-\sum_{k=1}^{r_i}\tilde{x}_{ik}\hat{a}_{ik}$,　　$\hat{a}_{ij}=\begin{cases}0 & j=1\\ a_{ij} & j\neq 1\end{cases}$

3 群以上の場合　$d_{ij}^{\alpha}=\hat{a}_{ij}^{\alpha}-\sum_{k=1}^{r_i}\tilde{x}_{ik}\hat{a}_{ik}^{\alpha}$,　　$\hat{a}_{ij}^{\alpha}=\begin{cases}0 & j=1\\ a_{ij}^{\alpha} & j\neq 1\end{cases}$

ここに基準化ウェイトの意味がカテゴリの影響が判別に正に働くか負に働くかを見ることであると考えて、以下のように、$\tilde{x}_{ik}$ はアイテム i カテゴリ k における群平均の単純平均とした。

$$\tilde{x}_{ik}=\frac{1}{m}\sum_{\alpha=1}^{m}\overline{x}_{ik}^{\alpha}$$

基準化されたカテゴリウェイトを用いると、判別関数値は以下のように与えられる。

2 群の場合　　$z=\sum_{i=1}^{p}\sum_{j=1}^{r_p}d_{ij}x_{ij}$ (4)

3 群以上の場合　$z^{\alpha}=\sum_{i=1}^{p}\sum_{j=1}^{r_i}d_{ij}^{\alpha}x_{ij}+\sum_{i=1}^{p}\sum_{j=1}^{r_i}\tilde{x}_{ij}\hat{a}_{ij}^{\alpha}-\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{r_i}\overline{x}_{ij}^{\alpha}\hat{a}_{ij}^{\alpha}$ (5)

判別分析は変数一つひとつが独立であったが、数量化Ⅱ類の場合は、1 つのアイテムが判別分析の 1 つの変数に対応する。その中にはいくつかのカテゴリが含まれているために、アイテムの重要性は複数のカテゴリをまとめた重要性と解釈される。そのため、アイテムの重要性をみるには、カテゴリによる判別関数値の変化幅であるウェイト範囲や以下に述べるアイテムと判別関数値との相関係数、アイテムと判別関数値との偏相関係数の値などが参照される。

アイテムと判別関数間の相関係数を次のように与える。

$$r_{ij}=s_{ij}\Big/\sqrt{s_{ii}s_{jj}}\ ,\quad r_{iz}=s_{iz}\Big/\sqrt{s_{ii}s_{zz}}$$

ここに、アイテムと判別関数間の共分散 s_{ij}, s_{iz}, s_{zz} は以下のように定義される。

$$s_{ij}=\frac{1}{n-1}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}(x_{i\lambda}^{\alpha}-\overline{x}_i)(x_{j\lambda}^{\alpha}-\overline{x}_j)\ ,\quad s_{iz}=\frac{1}{n-1}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}(x_{i\lambda}^{\alpha}-\overline{x}_i)(z_{\lambda}^{\alpha}-\overline{z})\ ,$$

$$s_{zz}=\frac{1}{n-1}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}(z_{\lambda}^{\alpha}-\overline{z})^2$$

但し、$x_{i\lambda}^{\alpha}=\sum_{j=1}^{r_i}\hat{a}_{ij}x_{ij\lambda}^{\alpha}$,　$\overline{x}_i=\frac{1}{n}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}x_{i\lambda}^{\alpha}$,　$\overline{z}=\frac{1}{n}\sum_{\alpha=1}^{m}n_\alpha z^{\alpha}$ である。

変更点を明らかにするために、プログラム変更以前の定義も与えておく。

$$s_{iz}=\frac{1}{n-1}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}(x_{i\lambda}^{\alpha}-\overline{x}_i)(\overline{z}^{\alpha}-\overline{z})\ ,\quad s_{zz}=\frac{1}{n-1}\sum_{\alpha=1}^{m}n_\alpha(\overline{z}^{\alpha}-\overline{z})^2\ ,\quad \overline{z}^{\alpha}=\frac{1}{n_\alpha}\sum_{\lambda=1}^{n_\alpha}z_{\lambda}^{\alpha}$$

アイテム i と判別関数との偏相関係数 $\tilde{r}_{iy}$ は、上の相関係数を用いた相関行列 $\mathbf{R}$ の逆行列 $\mathbf{R}^{-1}$ の成分 r^{ij}, r^{iz}, r^{zz} を用いて、以下のように与えられる。

$$\tilde{r}_{iz}=-r^{iz}\Big/\sqrt{r^{ii}r^{zz}}$$

数量化Ⅱ類では 2 群の判別の場合、各アイテムについて判別分析と同様にその有効性の F 値を求めることができる。アイテム i の有効性の F 値は以下となる。最後の分布形は仮に変数の正規性が成り立つ場合の性質であるが、当然数量化Ⅱ類のデータでは成り立たない。参考までの仮の表示である。

$$F_i=\frac{(n_1+n_2-p-1)n_1n_2(D^2-D_i^2)}{(n_1+n_2)(n_1+n_2-2)+n_1n_2D_i^2}\sim F_{r_i-1,n_1+n_2-p-1}\ \text{分布}$$

ここに、D_i^2 は両群のカテゴリ i を除いたマハラノビス距離である。

2）正準形式に基づく方法

マハラノビス形式と同様に、判別関数は係数 $a_{ij}\ (i=1,\cdots,p,\ j=2,\cdots,r_i)$ と定数 z_0 を用いて以下のように与える。

$$z_\lambda=\sum_{i=1}^{p}\sum_{j=2}^{r_i}a_{ij}x_{ij\lambda}+z_0$$

この判別関数は新しい変数表記法では以下となる。

$$z_\lambda=\sum_{I=1}^{P}a_Ix_I+z_0$$

この表記法では、第 1 カテゴリを除いた数量化Ⅱ類と判別分析が同等である。

我々は z_λ^{α} の群間の変動 s_B^2 と群別変動の合計 s^2 を以下のように定義し、群間の変動を際立たせるために、これらの分散比 $\rho=s_B^2/s^2$ を最大化することを考える。

$$s_B^2=\sum_{\alpha=1}^{m}n_\alpha(\overline{z}^{\alpha}-\overline{z})^2\ ,\quad s^2=\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha}(z_\lambda^{\alpha}-\overline{z}^{\alpha})^2$$

ここに、$\bar{z}^{\alpha}=\dfrac{1}{n_{\alpha}}\sum_{\lambda=1}^{n_{\alpha}}z_{\lambda}^{\alpha}$ ，　$\bar{z}=\dfrac{1}{n}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_{\alpha}}z_{\lambda}^{\alpha}$ ，　$n=\sum_{\alpha=1}^{m}n_{\alpha}$ である。

この分散比を係数で微分することにより、判別分析と同様に以下の方程式が得られる。

$$\mathbf{Ba}=\rho\mathbf{Sa} \tag{6}$$

この方程式はデータを以下のようにまとめ、

$$\mathbf{X}=\begin{pmatrix} x_{121}^{1} & \cdots & x_{1r_1 1}^{1} & \cdots & x_{p21}^{1} & \cdots & x_{pr_p 1}^{1} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{12n_1}^{1} & \cdots & x_{1r_1 1}^{1} & \cdots & x_{p2n_1}^{1} & \cdots & x_{pr_p n_1}^{1} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{121}^{m} & \cdots & x_{121}^{m} & \cdots & x_{p21}^{m} & \cdots & x_{pr_p 1}^{m} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{12n_m}^{m} & \cdots & x_{12n_m}^{m} & \cdots & x_{p2n_m}^{m} & \cdots & x_{pr_p n_m}^{m} \end{pmatrix}$$

$$\bar{\mathbf{X}}_B=\begin{pmatrix} \bar{x}_{12}^{1} & \cdots & \bar{x}_{1r_1}^{1} & \cdots & \bar{x}_{p2}^{1} & \cdots & \bar{x}_{pr_p}^{1} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{12}^{1} & \cdots & \bar{x}_{1r_1}^{1} & \cdots & \bar{x}_{p2}^{1} & \cdots & \bar{x}_{pr_p}^{1} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{12}^{m} & \cdots & \bar{x}_{12}^{m} & \cdots & \bar{x}_{p2}^{m} & \cdots & \bar{x}_{pr_p}^{m} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{12}^{m} & \cdots & \bar{x}_{12}^{m} & \cdots & \bar{x}_{p2}^{m} & \cdots & \bar{x}_{pr_p}^{m} \end{pmatrix}\begin{matrix} \left.\begin{matrix}\\\\\end{matrix}\right\}n_1 \\ \vdots \\ \left.\begin{matrix}\\\\\end{matrix}\right\}n_m \end{matrix}$$

$$\bar{\mathbf{X}}=\begin{pmatrix} \bar{x}_{12} & \cdots & \bar{x}_{1r_1} & \cdots & \bar{x}_{p2} & \cdots & \bar{x}_{pr_p} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{12} & \cdots & \bar{x}_{1r_1} & \cdots & \bar{x}_{p2} & \cdots & \bar{x}_{pr_p} \end{pmatrix}\left.\begin{matrix}\\\\\end{matrix}\right\}n$$

方程式中の行列を以下のように定義することによって得られる。

$${}^{t}\mathbf{a}=\begin{pmatrix} a_{12} & \cdots & a_{1r_1} & \cdots & a_{p2} & \cdots & a_{pr_p} \end{pmatrix}$$

$$\mathbf{S}=\frac{1}{n-m}{}^{t}(\mathbf{X}-\bar{\mathbf{X}}_B)(\mathbf{X}-\bar{\mathbf{X}}_B),\quad \mathbf{B}=\frac{1}{n-m}{}^{t}(\bar{\mathbf{X}}_B-\bar{\mathbf{X}})(\bar{\mathbf{X}}_B-\bar{\mathbf{X}})$$

ここにnはすべての群のデータ数の合計、mは群の数である。

方程式 (6) は正準判別分析と同様の方法で変形され、以下となる。

$$\mathbf{Au}=\rho\mathbf{u} \tag{7}$$

ここに、$\mathbf{A}=\mathbf{F}^{-1}\mathbf{B}\,{}^{t}\mathbf{F}^{-1}$，　$\mathbf{u}={}^{t}\mathbf{Fa}$、また$\mathbf{F}$は$\mathbf{S}=\mathbf{F}\,{}^{t}\mathbf{F}$となる下三角行列である。

(7) 式の第r固有値に対する規格化された固有ベクトル$\mathbf{u}^{(r)}$を使って、係数は$\mathbf{a}^{(r)}={}^{t}\mathbf{F}^{-1}\mathbf{u}^{(r)}$となり、これにより判別関数は以下となる。

$$z^{(r)}=\sum_{I=1}^{P}a_I^{(r)}x_I-\sum_{I=1}^{P}a_I^{(r)}\tilde{x}_I \tag{8}$$

ここで定数項については、正準判別分析と同様に、各固有値に対応する判別関数の群別平均の単純平均が 0 になるようにしている。

係数$a_{ij}^{(r)}$は各アイテムの第 1 カテゴリを除いたものであるので、以下のような基準化した係数$d_{ij}^{(r)}$ $(i=1,\cdots,p,\ j=1,2,\cdots,r_i)$も計算しておく。

$$d_{ij}^{(r)}=\hat{a}_{ij}^{(r)}-\sum_{k=1}^{r_i}\hat{a}_{ik}^{(r)}\tilde{x}_{ik}\ ,\qquad \hat{a}_{ij}^{(r)}=\begin{cases}0 & j=1\\ a_{ij}^{(r)} & j\neq 1\end{cases}$$

ここに基準化ウェイトの意味を考えて、$\tilde{x}_{ik}$は判別関数のときと同様に、アイテムi カテゴリk における群平均の単純平均とした。

$$\tilde{x}_{ik}=\frac{1}{m}\sum_{\alpha=1}^{m}\bar{x}_{ik}^{\alpha}$$

基準化されたカテゴリウェイトを用いると、判別関数は以下のように与えられる。

$$z^{(r)}=\sum_{i=1}^{p}\sum_{j=1}^{r_i}d_{ij}^{(r)}x_{ij} \qquad (9)$$

１０．数量化Ⅲ類

10.1 数量化Ⅲ類とは

数量化Ⅲ類は、与えられたカテゴリ（変数）に反応するかどうかで１か０の数値を与えて、カテゴリや個体の類似性を解明しようとするものである。

例えば各人が表 1 の食品について、それぞれの好み（1：好物、0：それほどでも）を与えた（数量化Ⅲ類 1.txt）。これから好みの特徴を表す式を求め、人と食品を分類する。

表１　数量化Ⅲ類のデータ

ご飯	パン	うどん	そば	ラーメン	スパ
1	0	1	1	1	0
1	0	1	0	0	0
0	1	0	0	1	1
:	:	:	:	:	:
0	1	0	1	1	0

この分割表のデータ $x_{i\lambda}$ （i：カテゴリ，λ：個体）を元に、行と列別々に類似性を見る分析が数量化Ⅲ類である。類似性は、2 つの分類変数にそれぞれ、特徴的な量 u_i^a と v_λ^a を考え、それらの量が最大の相関係数を持つようにして考える。

カテゴリの類似度は、殆どの場合 $\alpha=1,2$ のカテゴリウェイトと呼ばれる 2 次元の量 u_i^a の近さで与えられる。結果は平面上に点を打ってその近さで見ることが多い。個体の類似度は同じく $\alpha=1,2$ の個体ウェイトと呼ばれる 2 次元の量 v_λ^a の近さで与えられる。これも平面上に点を打ってその近さで見る。結果の精度は $\alpha=1,2$ を使う場合、2 次元までの累積寄与率で与えられる。このデータの具体的な分析については次節で述べる。

表 1 のデータの中には、穀類と全く関係のないデータを含めることができる。例えば性別を男性と女性に分けて加えると、男性の好みや女性の好みの分類ができる。但し、これらの変数の影響により、元の穀類の配置がかなり変わることがあるので、注意を要する（実際、これは問題がないのか調査中）。

10.2　プログラムの利用法

メニュー「分析－多変量解析他－分類手法－数量化Ⅲ類」を選択すると図 1 に示される実行画面が表示される。

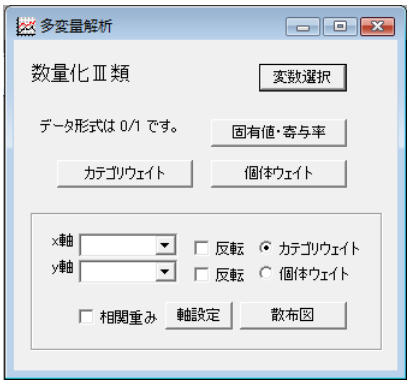
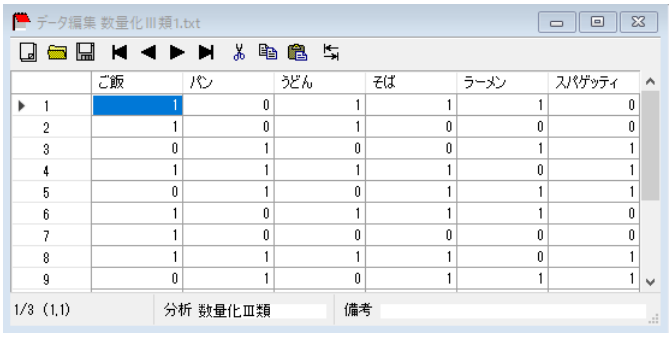


図 1　数量化Ⅲ類実行画面

分析は図２のような {0,1} の値を持つデータから実行される。



データ編集 数量化Ⅲ類1.txt

	ご飯	パン	うどん	そば	ラーメン	スパゲッティ
1	1	0	1	1	1	0
2	1	0	1	0	0	0
3	0	1	0	0	1	1
4	1	1	1	1	0	1
5	0	1	0	1	1	1
6	1	0	1	1	1	0
7	1	0	0	0	0	0
8	1	1	1	1	0	1
9	0	1	0	1	1	1

1/3 (1,1)　分析 数量化Ⅲ類　備考

図２　分割表データ（数量化Ⅲ類 1.txt）

変数を選択して、「固有値・寄与率」ボタンをクリックすると図３のような結果が示される。

固有値・寄与率

	第1次元	第2次元	第3次元	第4次元	第5次元
固有値	0.351	0.156	0.095	0.060	0.025
相関係数	0.592	0.395	0.308	0.246	0.159
寄与率	0.510	0.227	0.138	0.088	0.037
累積寄与率	0.510	0.737	0.875	0.963	1.000

図３　固有値・寄与率画面

ここで表示される固有値は、個体に与えたパラメータと変数に与えたパラメータの相関係数 ρ の値の２乗である。

分析メニューで「カテゴリウェイト」ボタンをクリックすると図４のような結果が表示される。

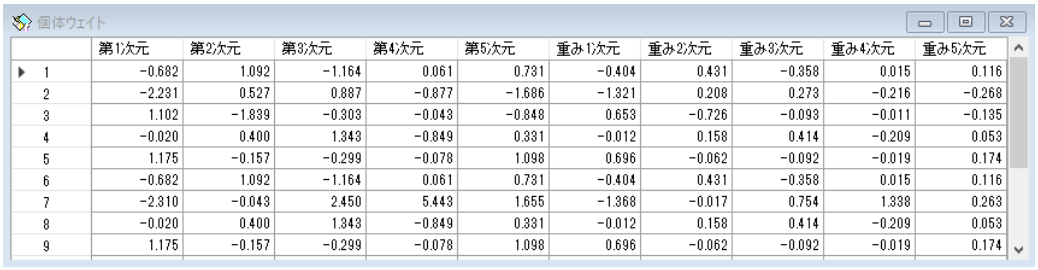
カテゴリウェイト

	第1次元	第2次元	第3次元	第4次元	第5次元	重み1次元	重み2次元	重み3次元	重み4次元	重み5次元
ご飯	-1.368	-0.017	0.754	1.338	0.263	-0.810	-0.007	0.232	0.329	0.042
パン	1.200	-0.062	0.853	0.316	-1.618	0.711	-0.024	0.263	0.078	-0.257
うどん	-1.274	0.433	-0.208	-1.770	-0.799	-0.755	0.171	-0.064	-0.435	-0.127
そば	0.826	1.929	-0.089	-0.045	1.102	0.489	0.761	-0.027	-0.011	0.175
ラーメン	0.201	-0.622	-1.891	0.536	-0.101	0.119	-0.245	-0.582	0.132	-0.016
スパゲッティ	0.556	-1.494	0.758	-0.884	1.315	0.329	-0.589	0.233	-0.217	0.209

図４　カテゴリウェイト画面

ここでは自明な解に対応する結果は表示されていない。

分析メニューの「個体ウェイト」ボタンをクリックすると、図５の個体ウェイト画面が表示される。



個体ウェイト

	第1次元	第2次元	第3次元	第4次元	第5次元	重み1次元	重み2次元	重み3次元	重み4次元	重み5次元
1	-0.682	1.092	-1.164	0.061	0.731	-0.404	0.431	-0.358	0.015	0.116
2	-2.231	0.527	0.887	-0.877	-1.686	-1.321	0.208	0.273	-0.216	-0.268
3	1.102	-1.839	-0.303	-0.043	-0.848	0.653	-0.726	-0.093	-0.011	-0.135
4	-0.020	0.400	1.343	-0.849	0.331	-0.012	0.158	0.414	-0.209	0.053
5	1.175	-0.157	-0.299	-0.078	1.098	0.696	-0.062	-0.092	-0.019	0.174
6	-0.682	1.092	-1.164	0.061	0.731	-0.404	0.431	-0.358	0.015	0.116
7	-2.310	-0.043	2.450	5.443	1.655	-1.368	-0.017	0.754	1.338	0.263
8	-0.020	0.400	1.343	-0.849	0.331	-0.012	0.158	0.414	-0.209	0.053
9	1.175	-0.157	-0.299	-0.078	1.098	0.696	-0.062	-0.092	-0.019	0.174

図５　個体ウェイト画面

カテゴリウェイトや個体ウェイトを図で表示するには、まずどちらを表示するかをラジオボタンで選択し、「軸設定」ボタンをクリックしてｘ軸とｙ軸の成分を選択する。その後、「散布図」ボタンをクリックすると図６や図７のような散布図が表示される。

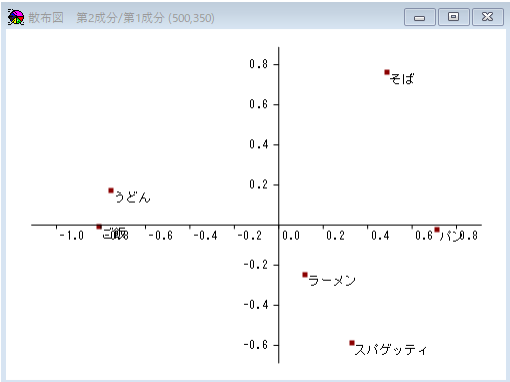


図６　カテゴリウェイトの散布図

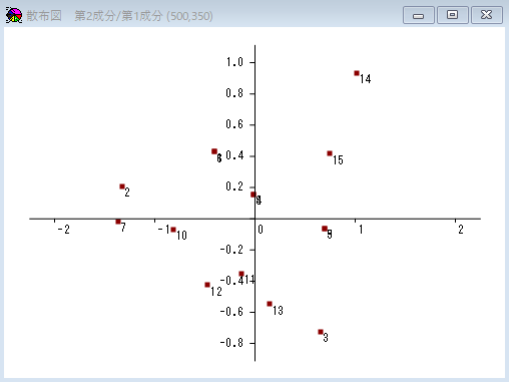


図７　個体ウェイトの散布図

デフォルトでは図１の「相関重み」チェックボックスにチェックを入れてあるので、これは散布図の各成分に相関係数をかけて表示する方法になっている。固有値の寄与を図で表すにはこの方法の方がよい。場合によって成分を反転させて表示する方がよい場合もあるが、そのためには反転チェックボックスにチェックを入れて「散布図」ボタンをクリックする。

問題（数量化Ⅲ類 2.txt）

数量化Ⅲ類 2.txt は高校生、大学生、会社員の好きなブランドを調査した結果である。1の項目は回答者が○を付けた項目である。数量化Ⅲ類を用いて以下の問いに答えよ。
データの与え方によって、このように違った項目を混在させることもできる。

１）３次元までの寄与率と累積寄与率を求めよ。

	１次元	２次元	３次元
寄与率			
累積寄与率			

２）このうち分析に２次元まで利用するとして、カテゴリウェイトの値を求めよ。

	第１次元	第２次元
A		
B		
C		
D		
高校生		
大学生		
会社員		

３）カテゴリウェイトの散布図から以下の問いに答えよ。

高校生に最も人気のあるブランドは　　［A・B・C・D］

大学生に最も人気のあるブランドは　　［A・B・C・D］

会社員に最も人気のあるブランドは　　［A・B・C・D］

４）先頭３人の２次元までの個体ウェイトの値を求めよ。

	第１次元	第２次元
1		
2		
3		

５）個体ウェイトの散布図から高校生、大学生、会社員はグループになっているか。

［なっている・なっていない］

問題解答（数量化Ⅲ類 2.txt）

１）３次元までの寄与率と累積寄与率を求めよ。

	１次元	２次元	３次元
寄与率	0.524	0.315	0.091
累積寄与率	0.524	0.839	0.930

２）このうち分析に２次元まで利用するとして、カテゴリウェイトの値を求めよ。

	第１次元	第２次元
A	-1.137	0.205
B	-0.159	-1.010
C	1.108	0.445
D	0.508	0.080
高校生	1.519	1.127
大学生	-0.123	-2.031
会社員	-1.730	1.492

３）カテゴリウェイトの散布図から以下の問いに答えよ。

高校生に最も人気のあるブランドは　　［A・B・[C]・D］

大学生に最も人気のあるブランドは　　［A・[B]・C・D］

会社員に最も人気のあるブランドは　　［[A]・B・C・D］

４）先頭３人の２次元までの個体ウェイトの値を求めよ。

	第１次元	第２次元
1	1.325	0.902
2	-0.289	-1.128
3	-1.279	0.375

５）個体ウェイトの散布図から高校生、大学生、会社員はグループになっているか。

［[なっている]・なっていない］

10.3　数量化Ⅲ類の理論

数量化Ⅲ類は個体及びカテゴリにそれぞれ数値を与えて、データの持つ類似性を解明しようとするものである。個々のデータはカテゴリに反応した場合 1、反応しない場合は 0 で与えられる。

$$x_{i\lambda} \in \{0,1\}$$

ここに、i はカテゴリ、λ は個体を表わす。また、カテゴリ数を p、データ数を n（$p \le n$）とする。

この分析では、カテゴリと個体に対してカテゴリウェイトと個体ウェイトと呼ばれる特徴的な点数 u_i と v_λ を与える。そのようにすると λ 番目の個体の i 番目のカテゴリの回答に対して、数値の組 $(u_i x_{i\lambda}, v_\lambda x_{i\lambda})$ が割り当てられる。即ち、各回答の反応した位置には数値の組 (u_i, v_λ) が割り当てられる。この反応した点を 1 つのデータ点と考えると、カテゴリと個体に割り当てられた数値間の散布図が得られる。各カテゴリや個体への数値の与え方によって散布図の形状は変わってくる。与えられた数値の順にカテゴリや個体を並べ替えると考えると、並べ替えによって大まかに散布図の形状を変えていると考えてもよい。似た回答をされたカテゴリや個体に属するデータ点を近くにまとめ、それと異なる回答をしたカテゴリや個体に属するデータ点を遠く離すには、この散布図の相関係数が最大になるように（データ点が直線状に並ぶように）点数を与えるとよい。数量化Ⅲ類では、このような考え方に基づき議論を進めて行く。

まず、各点の平均について考え、これが 0 になるように変数の原点を決める。即ち、以下とする。

$$\bar{u} = \frac{1}{T}\sum_{i=1}^{p}\sum_{\lambda=1}^{n} x_{i\lambda}u_i = \frac{1}{T}\sum_{i=1}^{p} c_i u_i = 0 \quad ,$$

$$\bar{v} = \frac{1}{T}\sum_{i=0}^{p}\sum_{\lambda=1}^{n} x_{i\lambda}v_\lambda = \frac{1}{T}\sum_{\lambda=1}^{n} d_\lambda v_\lambda = 0$$

$$c_i = \sum_{\lambda=1}^{n} x_{i\lambda} \quad , \qquad d_\lambda = \sum_{i=1}^{p} x_{i\lambda} \quad , \qquad T = \sum_{i=1}^{p}\sum_{\lambda=1}^{n} x_{i\lambda} \, ,$$

これによって、2 変量 (u_i, v_λ) の分散、共分散は以下で与えられる。

$$S_u^2 = \frac{1}{T}\sum_{i=1}^{p}\sum_{\lambda=1}^{n} x_{i\lambda}^2 u_i^2 = \frac{1}{T}\sum_{i=1}^{p} c_i u_i^2 \quad ,$$

$$S_v^2 = \frac{1}{T}\sum_{i=1}^{p}\sum_{\lambda=1}^{n} x_{i\lambda}^2 v_\lambda^2 = \frac{1}{T}\sum_{\lambda=1}^{n} d_\lambda v_\lambda^2$$

$$S_{uv} = \frac{1}{T}\sum_{i=1}^{p}\sum_{\lambda=1}^{n} x_{i\lambda}u_i x_{i\lambda}v_\lambda = \frac{1}{T}\sum_{i=1}^{p}\sum_{\lambda=1}^{n} x_{i\lambda}u_i v_\lambda$$

これからカテゴリと個体の相関係数を $\rho = S_{uv} / S_u S_v$ と表わす。点数の分散を 1 とする制約条件を付けて、この相関係数 ρ を最大にする点数を求めるために、Lagrange の未定乗数法を用いる。

$$L = S_{uv} - \eta\left(S_u^2 - 1\right) - \mu\left(S_v^2 - 1\right)$$

ここに η と μ は未定乗数である。これを u_i と v_λ で微分して、以下の方程式を得る。

$$\sum_{\lambda=1}^{n} x_{i\lambda} v_\lambda - 2\eta c_i u_i = 0\,,\quad \sum_{i=1}^{p} x_{i\lambda} u_i - 2\mu d_\lambda v_\lambda = 0$$

これらの式を行列で表示すると以下のようになる。

$$\mathbf{Xv} = 2\eta\mathbf{Cu}\,,\quad \mathbf{X'u} = 2\mu\mathbf{Dv} \tag{1}$$

ここに

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pn} \end{pmatrix},\quad \mathbf{C} = \begin{pmatrix} c_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c_p \end{pmatrix},\quad \mathbf{D} = \begin{pmatrix} d_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_n \end{pmatrix},$$

$$\mathbf{u'} = \begin{pmatrix} u_1 & \cdots & u_p \end{pmatrix},\quad \mathbf{v'} = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix}$$

これらの行列を用いると、以下の関係も示される。

$$\mathbf{u'Cu} = TS_u^2 = T\,,\quad \mathbf{v'Dv} = TS_v^2 = T\,,\quad \mathbf{u'Xv} = \mathbf{v'X'u} = TS_{uv} = T\rho$$

(1) の方程式で、左式に左から $\mathbf{u'}$ を掛けると上の関係から、$\rho = 2\eta$、同様に右式に左から $\mathbf{v'}$ を掛けると $\rho = 2\mu$ を得る。右式を $\mathbf{v}$ について解いて左式に代入すると以下となる。

$$\mathbf{C}^{-1}\mathbf{XD}^{-1}\mathbf{X'u} = \rho^2\mathbf{u}\,,\quad \text{また、}\ \mathbf{v} = \rho^{-1}\mathbf{D}^{-1}\mathbf{X'u} \tag{2}$$

また $\mathbf{v}$ についても対等に同様の関係が示されるが、ここでは省略する。

さて、ここで $S_u^2 = 1$ としたことから、$\mathbf{u}$ の規格化条件が $\frac{1}{T}\mathbf{u'Cu} = 1$ となるので、新たに以下のベクトル $\mathbf{z}$ を考える。

$$\mathbf{z} = \frac{1}{\sqrt{T}}\mathbf{C}^{1/2}\mathbf{u}\,,\quad \text{ここに}\ \ \mathbf{C}^{1/2} = \begin{pmatrix} \sqrt{c_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{c_p} \end{pmatrix}$$

これを用いて最終的に方程式 (2) は以下となる。

$$\mathbf{Az} = \rho^2\mathbf{z}\,,\quad \mathbf{A} = \mathbf{C}^{-1/2}\mathbf{XD}^{-1}\mathbf{X'C}^{-1/2}\,,\ \text{規格化条件}\quad \mathbf{z'z} = 1 \tag{3}$$

異なる固有値 ρ_α^2（$\alpha = 1, \cdots, p$）に対する固有ベクトルを $\mathbf{z}^\alpha$ とすると、各点数は以下のように表される。

$$\mathbf{u}^\alpha = \sqrt{T}\mathbf{C}^{-1/2}\mathbf{z}^\alpha\,,\quad \mathbf{v}^\alpha = \rho_\alpha^{-1}\sqrt{T}\mathbf{D}^{-1}\mathbf{X'C}^{-1/2}\mathbf{z}^\alpha \tag{4}$$

ここでもう一度 (2) 式について考える。この方程式を成分表示すると以下となる。

$$\sum_{\lambda=1}^{n}\sum_{j=1}^{p} \frac{1}{c_i}\, x_{i\lambda}\, \frac{1}{d_\lambda}\, x_{j\lambda} u_j = \rho^2 u_i$$

ここで、$u_j = 1$ とすると。上式は以下となる。

$$\rho^2 = \sum_{\lambda=1}^{n}\sum_{j=1}^{p}\frac{1}{c_i}x_{i\lambda}\frac{1}{d_\lambda}x_{j\lambda} = \frac{1}{c_i}\sum_{\lambda=1}^{n}x_{i\lambda}\frac{1}{d_\lambda}\sum_{j=1}^{p}x_{j\lambda} = \frac{1}{c_i}\sum_{\lambda=1}^{n}x_{i\lambda} = 1$$

$$v_\lambda = \sum_{j=1}^{p}\frac{1}{d_\lambda}x_{j\lambda} = 1$$

即ち(2) 式には $\rho^2 = 1$，$\mathbf{u} = \mathbf{1}$，$\mathbf{v} = \mathbf{1}$ の自明な解が存在するが、この解は

$$\bar{u} = \frac{1}{T}\sum_{i=1}^{p}c_i u_i = 1 \neq 0 \quad , \qquad \bar{v} = \frac{1}{T}\sum_{\lambda=1}^{n}d_\lambda = 1 \neq 0$$

であるから、除外する。

点数 $\mathbf{u}$，$\mathbf{v}$ の与え方には、以下のように相関係数を掛ける方法もある。

$\tilde{\mathbf{u}}^\alpha = \rho_\alpha \mathbf{u}^\alpha$，　$\tilde{\mathbf{v}}^\alpha = \rho_\alpha \mathbf{v}^\alpha$

ここで $p \le n$ を仮定してきたが、$p > n$ の場合、先に $\mathbf{v}$ について求め、後で $\mathbf{u}$ について求めるが、方法は同様であるので省略する。

このカテゴリウェイト $\mathbf{u}^\alpha$ と個体ウェイト $\mathbf{v}^\alpha$ を用いてカテゴリ得点 $\mathbf{y}^\alpha$ と個体得点 $\mathbf{w}^\alpha$ をそれぞれ以下のように定義する場合もあるが、ここでは省略する。

$\mathbf{y}^\alpha = \mathbf{X}\mathbf{v}^\alpha$，$\mathbf{w}^\alpha = \mathbf{X}'\mathbf{u}^\alpha$

各成分の重要性を表すために、自明な解に対する固有値を ρ_p^2 として、これを除いて寄与率 λ_α を以下のように定義する。

$$\lambda_\alpha = \rho_\alpha^2 / \sum_{\beta=1}^{p-1}\rho_\beta^2 \qquad (\alpha \neq p)$$

１１．コレスポンデンス分析

11.1　コレスポンデンス分析とは

コレスポンデンス分析は、カテゴリに分けられた 2 次元分割表を元にして、類似のカテゴリを見つけ出す手法である。類似のカテゴリは 2 次元座標上で近い距離に表示され、利用者の判断により分類される。ここでは、以下の例を元にコレスポンデンス分析について説明する。

例（高橋信, Excel で学ぶコレスポンデンス分析, オーム社, 2005）

各年代の学生に好きな歌手を選んでもらったところ、以下の集計結果が得られた。それぞれの歌手はどの世代に支持されているか。コレスポンデンス分析で検討せよ。

表１　2 次元分割表

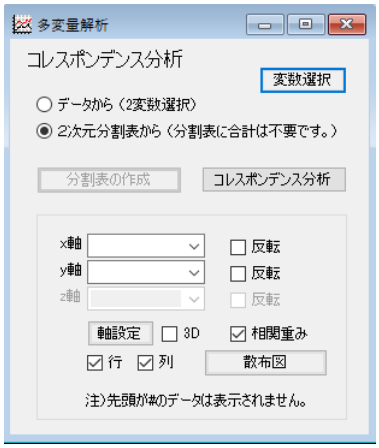
	A	B	C	D	計
中学生	10	19	13	5	47
高校生	13	8	15	16	52
大学生	18	11	14	8	51
計	41	38	42	29	150

この分割表で、例えば、中学生に支持されている歌手はと考え、10/41,19/38,13/42, 5/29 と考えた結果と、A を支持する学生はと考え、10/47,13/52,18/51 とした結果は一致するか？このような考え方では、どの学生同士、どの商品同士が近いか？など、答えがすぐに見えない場合もある。

この分割表のデータを元に、行と列すべての項目について類似性を見る分析がコレスポンデンス分析である。類似性は、2 つの分類変数にそれぞれ、特徴的な量u_i^aとv_λ^aを考え、それらの量が最大の相関係数を持つようにして考える。この考え方は数量化Ⅲ類に似ている。カテゴリの近さはこれらの値による距離で見る。コレスポンデンス分析は 2 次元分割表の 2 つの変数が入り混じることが特徴である。各カテゴリは 2 次元平面上に散布図として表示され、類似性は見た目で判断できる。

11.2　プログラムの利用法

メニュー［分析－多変量解析他－分類手法－コレスポンデンス分析］を選択すると図 1 に示されるコレスポンデンス分析実行画面が表示される。分析は通常の質的データと図 2 のような分割表の 2 通りから選択できる。



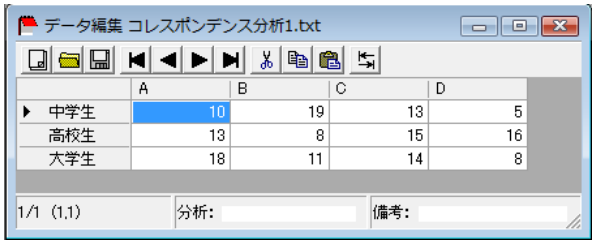
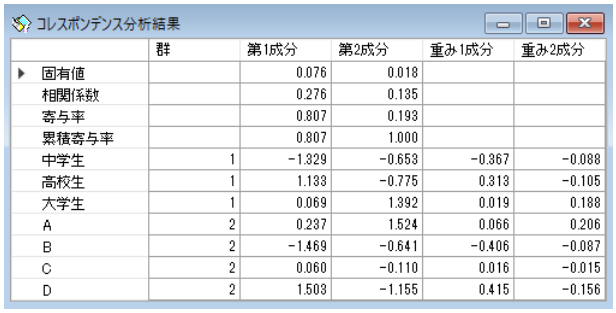


図１　コレスポンデンス分析実行画面　　図２　分割表データ

変数を選択して、「コレスポンデンス分析」ボタンをクリックすると図３のような分析結果が表示される。

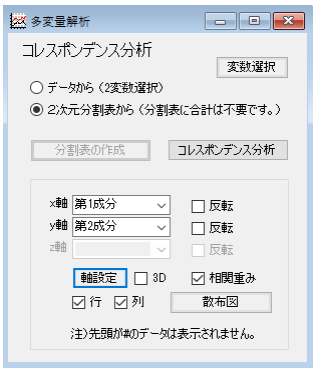


	群	第1成分	第2成分	重み1成分	重み2成分
固有値		0.076	0.018		
相関係数		0.276	0.135		
寄与率		0.807	0.193		
累積寄与率		0.807	1.000		
中学生	1	−1.329	−0.653	−0.367	−0.088
高校生	1	1.133	−0.775	0.313	−0.105
大学生	1	0.069	1.392	0.019	0.188
A	2	0.237	1.524	0.066	0.206
B	2	−1.469	−0.641	−0.406	−0.087
C	2	0.060	−0.110	0.016	−0.015
D	2	1.503	−1.155	0.415	−0.156

図３　コレスポンデンス分析実行結果

出力される成分数は２つの変数のカテゴリ数の小さい方から自明な固有値の数の１を引いた数であり、この例の場合２である。重み成分はそれぞれの成分に相関係数をかけたものである。

この結果を図の上で表示するには、まず「軸設定」ボタンをクリックし、図４のようにｘ軸とｙ軸に表示される成分の中で適切なものを選択する。通常はｘ軸に第１成分、ｙ軸に第２成分を表示する。「散布図」ボタンをクリックすると図５のような結果が表示される。



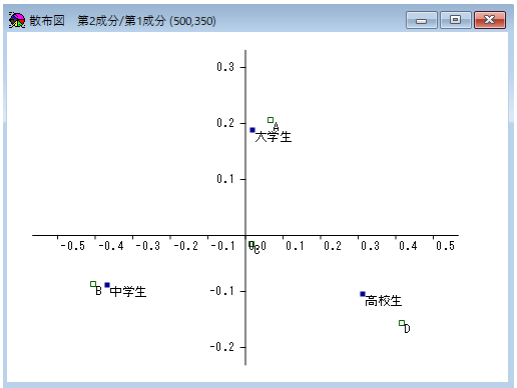


図４　軸設定された実行画面　　図５　散布図画面

相関係数の重みを付ける場合は、「相関重み」チェックボックスにチェックを入れ、軸を反転させて表示したい場合は、それぞれの軸の「反転」チェックボックスにチェックを入れて散布図を表示する。図 5 は相関重みにチェックを入れたグラフである。

問題（コレスポンデンス分析 2.txt）

コレスポンデンス分析 2.txt は 3 つの地域で 4 つの商品の売れ筋を調べた結果である。コレスポンデンス分析を用いて以下の問いに答えよ。

１）地域と商品に関する 2 次元分割表を描け。（合計は不要）

	商品１	商品２	商品３	商品４
地域１				
地域２				
地域３				

２）2 つの変数に付けたパラメータの相関係数、寄与率、累積寄与率を求めよ。

	第１成分	第２成分
相関係数		
寄与率		
累積寄与率		

３）2 つの変数に付けたパラメータの値を求めよ。

	第１成分	第２成分
地域１		
地域２		
地域３		
商品１		
商品２		
商品３		
商品４		

４）散布図を見て地域で売れ筋の商品を選択せよ。（複数選択）

地域１　［商品１・商品２・商品３・商品４］

地域２　［商品１・商品２・商品３・商品４］

地域３　［商品１・商品２・商品３・商品４］

５）2 次元分割表で見た場合、商品比率の最も高い商品はどれか。（単数選択）

地域１　［商品１・商品２・商品３・商品４］

地域２　［商品１・商品２・商品３・商品４］

地域３　［商品１・商品２・商品３・商品４］

以上から、コレスポンデンス分析の結果は、単純に比率だけから見たものとは異なる。

問題解答（コレスポンデンス分析 2.txt）

１）地域と商品に関する 2 次元分割表を描け。（合計は不要）

	商品１	商品２	商品３	商品４
地域１	14	12	13	16
地域２	21	13	17	11
地域３	30	20	16	17

２）2 つの変数に付けたパラメータの相関係数、寄与率、累積寄与率を求めよ。

	第１成分	第２成分
相関係数	0.121	0.082
寄与率	0.683	0.317
累積寄与率	0.683	1.000

３）2 つの変数に付けたパラメータの値を求めよ。

	第１成分	第２成分
地域１	−1.623	−0.038
地域２	0.647	−1.344
地域３	0.592	1.029
商品１	1.099	0.396
商品２	0.143	0.717
商品３	−0.112	−1.817
商品４	−1.652	0.581

４）散布図を見て地域で売れ筋の商品を選択せよ。（複数選択）

地域１　［商品１・商品２・商品３・**商品４**］
地域２　［商品１・商品２・**商品３**・商品４］
地域３　［**商品１**・**商品２**・商品３・商品４］

５）2 次元分割表で見た場合、商品比率の最も高い商品はどれか。（単数選択）

地域１　［商品１・商品２・商品３・**商品４**］
地域２　［**商品１**・商品２・商品３・商品４］
地域３　［**商品１**・商品２・商品３・商品４］

11.3　コレスポンデンス分析の理論

今 2 つの質的な変数、変数 1 と変数 2 があるとする。変数 1 のカテゴリ数を p、変数 2 のカテゴリ数を q（一般性を失わず $p \le q$）とする。この 2 つの変数に対して p 行 q 列の 2 次元分割表を考え、変数 1 のカテゴリ i、変数 2 のカテゴリ j に属するデータ数を n_{ij} とする。またデータ数の合計を以下のように定義する。

$$n_{i\cdot} \equiv \sum_{j=1}^{q} n_{ij}, \quad n_{\cdot j} \equiv \sum_{i=1}^{p} n_{ij}, \quad n \equiv \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}$$

次に変数 1 のカテゴリ i のデータに点数 u_i、変数 2 のカテゴリ j のデータに点数 v_j を与え、これらの点数の値によって各カテゴリ間の特徴的な関係を考えることとする。但し、これらの関係は変数 1 の点数と変数 2 の点数との相関係数を最大にするものとして与える。

これらの点数に対して、２つの変数の相関係数 ρ は以下のように与えられる。

$$\rho = \frac{S_{uv}}{S_u S_v} \quad ,$$

$$S_{uv} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} u_i v_j \,, \quad S_u^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i\cdot} u_i^2 \,, \quad S_v^2 = \frac{1}{n}\sum_{j=1}^{q} n_{\cdot j} v_j^2$$

ここに、S_{uv} は共分散、S_u^2 と S_v^2 は分散であり、2 つの変数の点数について平均は 0 としている。

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{p} n_{i\cdot} u_i = 0, \quad \bar{v} = \frac{1}{n}\sum_{j=1}^{q} n_{\cdot j} v_j = 0$$

この相関係数 ρ について、点数の分散を 1 とする制約条件を付けて最大値を求めるために Lagrange の未定乗数法を用いる。

$$L = S_{uv} - \lambda\left(S_u^2 - 1\right) - \mu\left(S_v^2 - 1\right)$$

ここに λ と μ は未定乗数である。これを u_i と v_j で微分して、以下の方程式を得る。

$$\sum_{k=1}^{q} n_{ik} v_k - 2\lambda n_{i\cdot} u_i = 0, \qquad \sum_{k=1}^{p} n_{kj} u_k - 2\mu n_{\cdot j} v_j = 0$$

これらの式を行列で表示すると上式は以下のようになる。

$$\boldsymbol{N}\boldsymbol{v} = 2\lambda \boldsymbol{D}_r \boldsymbol{u}, \quad \boldsymbol{N}'\boldsymbol{u} = 2\mu \boldsymbol{D}_c \boldsymbol{v}$$

ここに

$$\boldsymbol{N} = \begin{pmatrix} n_{11} & \cdots & n_{1q} \\ \vdots & \ddots & \vdots \\ n_{p1} & \cdots & n_{pq} \end{pmatrix}, \quad \boldsymbol{D}_r = \begin{pmatrix} n_{1\cdot} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & n_{p\cdot} \end{pmatrix}, \quad \boldsymbol{D}_c = \begin{pmatrix} n_{\cdot 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & n_{\cdot q} \end{pmatrix},$$

$$\boldsymbol{u}' = \begin{pmatrix} u_1 & \cdots & u_p \end{pmatrix}, \quad \boldsymbol{v}' = \begin{pmatrix} v_1 & \cdots & v_q \end{pmatrix}$$

上の方程式で、左式に左から $\boldsymbol{u}'$ を掛けると $\rho = 2\lambda$、同様に右式に左から $\boldsymbol{v}'$ を掛けると $\rho = 2\mu$ を得る。右式を $\boldsymbol{v}$ について解いて左式に代入すると以下となる。

$$\boldsymbol{D}_r^{-1}\boldsymbol{N}\boldsymbol{D}_c^{-1}\boldsymbol{N}'\boldsymbol{u} = \rho^2 \boldsymbol{u}, \quad \text{また、} \boldsymbol{v} = \rho^{-1}\boldsymbol{D}_c^{-1}\boldsymbol{N}'\boldsymbol{u} \qquad (1)$$

また $\boldsymbol{v}$ についても同様の関係が示されるが、ここでは省略する。

ここで $S_u^2 = 1$ としたことから、$\boldsymbol{u}$ の規格化条件を $\frac{1}{n}\boldsymbol{u}'\boldsymbol{D}_r\boldsymbol{u} = 1$ として、新たに以下のベクトル $\mathbf{z}$ を考える。

$$\mathbf{z} \equiv \frac{1}{\sqrt{n}}\boldsymbol{D}_r^{1/2}\boldsymbol{u}, \quad \text{ここに} \quad \boldsymbol{D}_r^{1/2} = \begin{pmatrix} \sqrt{n_{1\cdot}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{n_{p\cdot}} \end{pmatrix}$$

これを用いて(1)式は最終的に以下となる。

$$\boldsymbol{A}\mathbf{z} = \rho^2 \mathbf{z}, \quad \mathbf{z}'\mathbf{z} = 1, \quad \boldsymbol{A} \equiv \boldsymbol{D}_r^{-1/2}\boldsymbol{N}\boldsymbol{D}_c^{-1}\boldsymbol{N}'\boldsymbol{D}_r^{-1/2} \qquad (2)$$

異なる固有値 ρ_α^2（$\alpha = 1, \cdots, p$）に対する固有ベクトルを $\mathbf{z}^\alpha$ とすると、各点数は以下の

ように表される。

$$\boldsymbol{u}^{\alpha}=\sqrt{n}\boldsymbol{D}_r^{-1/2}\mathbf{z}^{\alpha},\quad \boldsymbol{v}^{\alpha}=\rho_{\alpha}^{-1}\sqrt{n}\boldsymbol{D}_c^{-1}\boldsymbol{N}'\boldsymbol{D}_r^{-1/2}\mathbf{z}^{\alpha}$$

ところで、(1) 式には $\rho^2=1$，$\boldsymbol{u}=1$ の自明な解が存在し、それに基づく固有値と固有ベクトルが得られるが、この解は除外される。

その他、点数 $\boldsymbol{u}$，$\boldsymbol{v}$ の与え方には、以下のように相関係数を掛ける方法もある。

$$\tilde{\boldsymbol{u}}^{\alpha}=\rho_{\alpha}\boldsymbol{u}^{\alpha},\ \tilde{\boldsymbol{v}}^{\alpha}=\rho_{\alpha}\boldsymbol{v}^{\alpha}$$

各成分の重要性を表すために、自明な解に対する固有値を ρ_p^2 として、以下で与えられる寄与率 λ_{α} を考える場合もある。

$$\lambda_{\alpha}=\rho_{\alpha}^2/\sum_{\beta=1}^{p-1}\rho_{\beta}^2 \qquad (\alpha\neq p)$$

11.4　数量化Ⅲ類とコレスポンデンス分析の関係

ここでは、数量化Ⅲ類とコレスポンデンス分析の関係を述べておく。例えば数量化Ⅲ類の図１のデータを用いて２つの分析でカテゴリウェイトと個体ウェイトについての散布図を示しておく。理論のところでも見たように、これらは同じ答えが期待される。

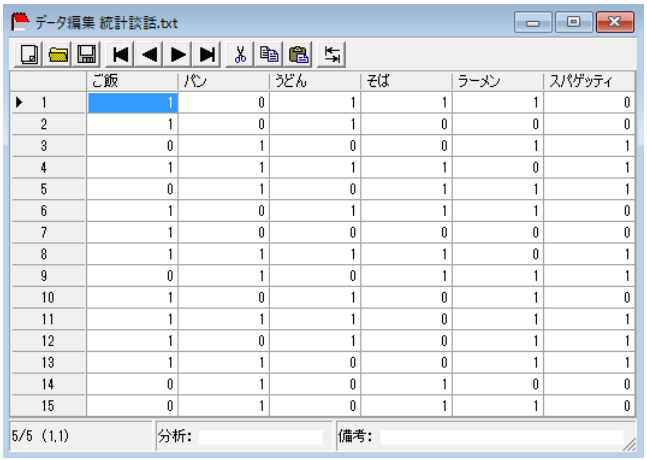


図１　数量化Ⅲ類データ

図２はカテゴリウェイトについて、左が数量化Ⅲ類の結果、右がコレスポンデンス分析の結果である。同様に図３は個体ウェイトについての同じ結果である。

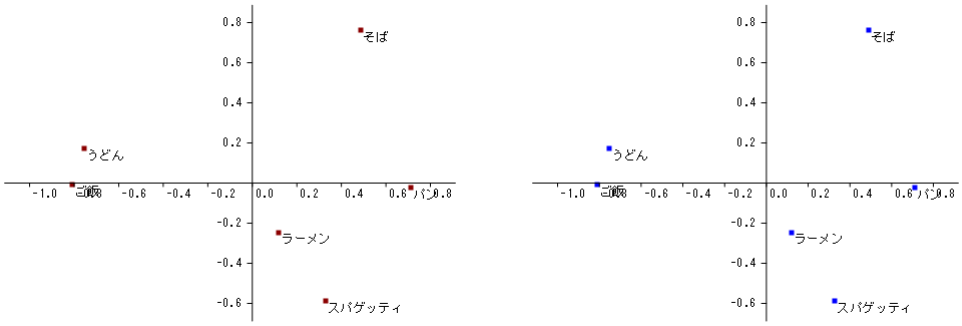


図２　数量化Ⅲ類とコレスポンデンス分析のカテゴリウェイト

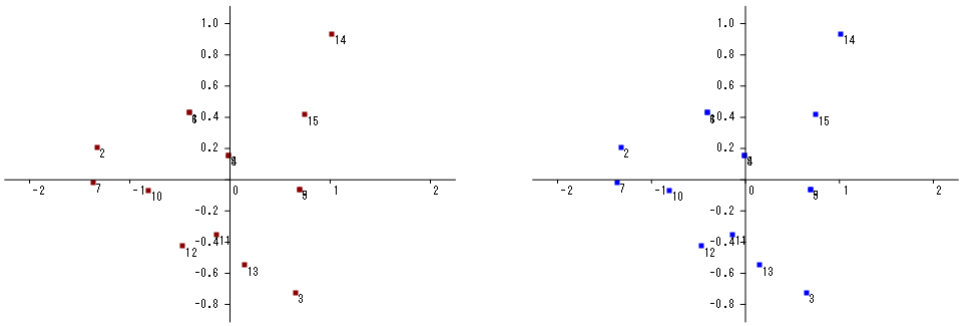


図３　数量化Ⅲ類とコレスポンデンス分析の個体ウェイト

これらは明らかに同じものであり、0/1 データを用いる限り両者は同じである。コレスポンデンス分析ではこれ以外の分割表データも扱えるので、コレスポンデンス分析は数量化Ⅲ類の拡張になっている。ただ、コレスポンデンス分析で、分割表を作る前のデータから処理することを考えると、コレスポンデンス分析はカテゴリ数自由の 2 変数のデータ、数量化Ⅲ類はカテゴリ数 2 の一般の複数変数のデータが対象である。変数という意味では数量化Ⅲ類はコレスポンデンス分析の一部を含んでいる。

ここで、数量化Ⅲ類についてもコレスポンデンス分析についても、「相関重み」にデフォルトでチェックが入っている。これは各次元のベクトルの長さの 2 乗が固有値の大きさ（寄与率に比例する）になるように調整されたものである。これにより、通常のユークリッド距離が、寄与率を考慮した類似度（距離）となっていることが分かる。これは主成分分析や因子分析の散布図で、因子負荷量を用いていることと同じである。

変数や個体を 2 次元平面上に散布させて類似性を見る方法は、他の次元の寄与を考えないということで、近似的な見方である。これに対して、これらの分析で出力された結果のすべての次元について、平方ユークリッド距離を考えたクラスター分析を併用させることは、すべての影響を別の表現法で眺める方法として興味深い。