College Analysis レファレンスマニュアル

－ 多変量解析 －

目次

１．実験計画法

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。このプログラムではこれらの分析の関係を図１のようにまとめ、それに基づいて分析メニューが作られている。

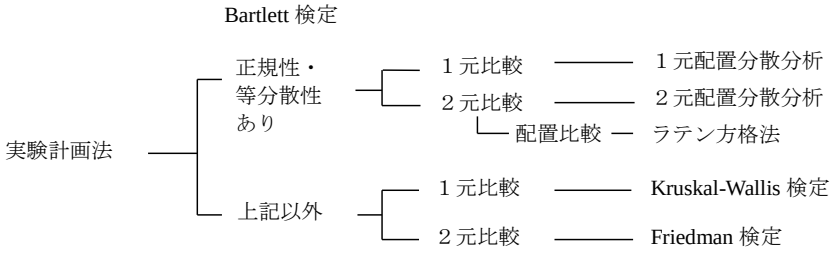


図１ 実験計画法の全体像

1.1 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。

$$x_{i\lambda} = \mu + \alpha_i + \varepsilon_{i\lambda}, \qquad \varepsilon_{i\lambda} \sim N(0,\sigma^2)$$ 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。

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

1.2 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 種類の水準の中間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間値が等しいとして以下の性質を利用する。

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{p} n_i \left(\frac{w_i}{n_i} - \frac{n+1}{2} \right)^2 \sim \chi^2_{p-1} \text{ 分布}$$

1.3 Bartlett の検定

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。

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

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

1.4 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_2	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。

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

但し、各パラメータには以下の条件を付ける。

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

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。

$$S = S_P + S_Q + S_I + S_E$$

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。

$$x_{ij} = \mu + \alpha_j + \beta_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$ 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。

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

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、

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

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

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）

1.5 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。

$$D = \frac{12}{s(s+1)r}\sum_{j=1}^{s} w_j^2 - 3r(s+1) \sim \chi^2_{s-1} \text{ 分布}$$

1.6 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。

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

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。

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

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

1.8 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。

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

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。

$$Z_{ij} = \frac{\left|\frac{w_i}{n_i} - \frac{w_j}{n_j}\right| - \frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}{\sqrt{\frac{n(n+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim N(0,1) \text{ 分布}$$

実験計画法の分析画面を図 2 に示す。

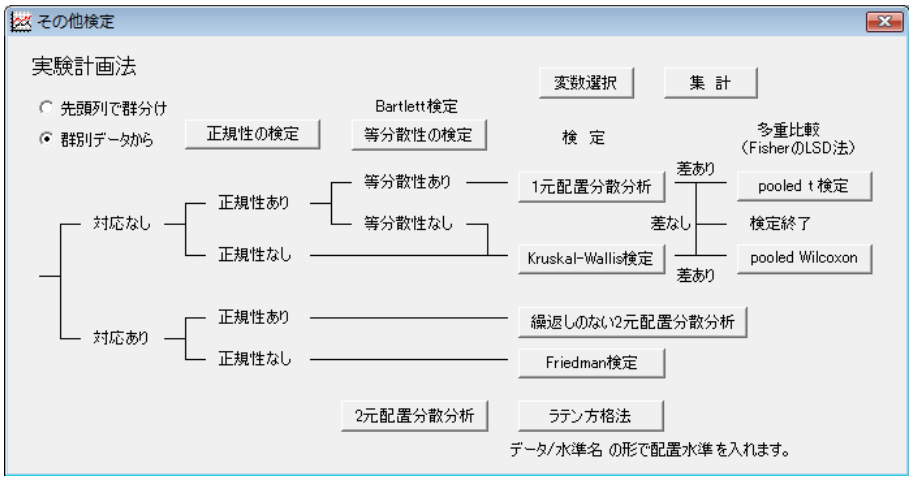


図 2 実験計画法分析画面

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。

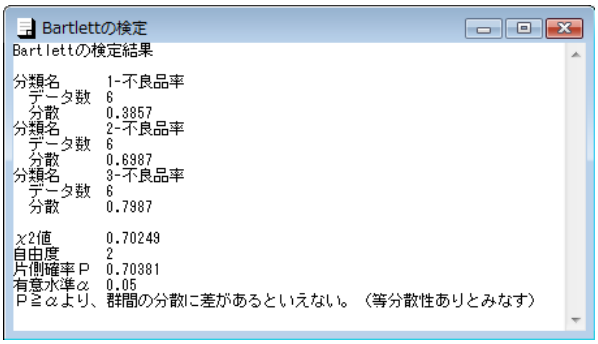


図 3 等分散の検定出力画面

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。

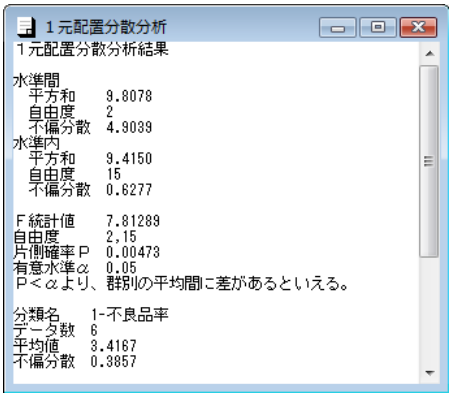


図 4a　1 元配置分散分析出力画面

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。

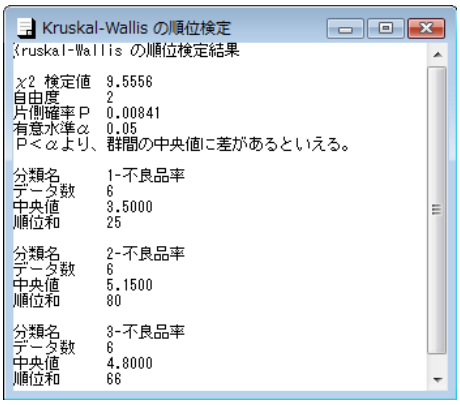


図 5 Kruskal-Wallis 検定出力画面

「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。

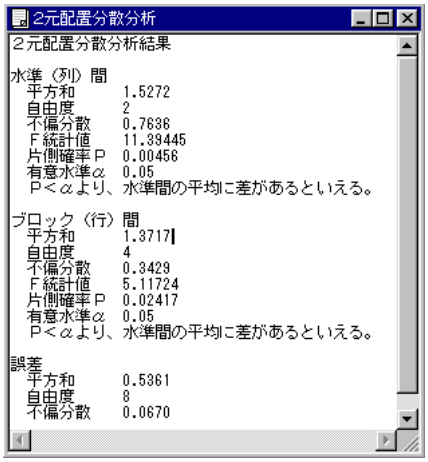
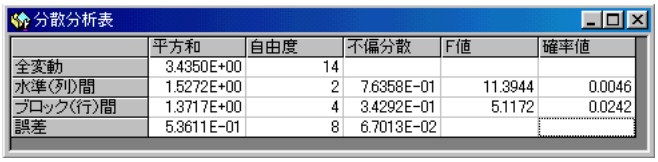


図 6a　2 元配置分散分析（繰り返しなし）



	平方和	自由度	不偏分散	F値	確率値
全変動	3.4350E+00	14			
水準(列)間	1.5272E+00	2	7.6358E-01	11.3944	0.0046
ブロック(行)間	1.3717E+00	4	3.4292E-01	5.1172	0.0242
誤差	5.3611E-01	8	6.7013E-02		

図 6b　2 元配置分散分析表（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。

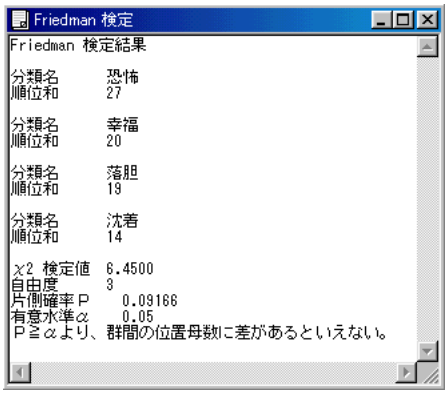


図 7 Friedman 検定出力画面

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。

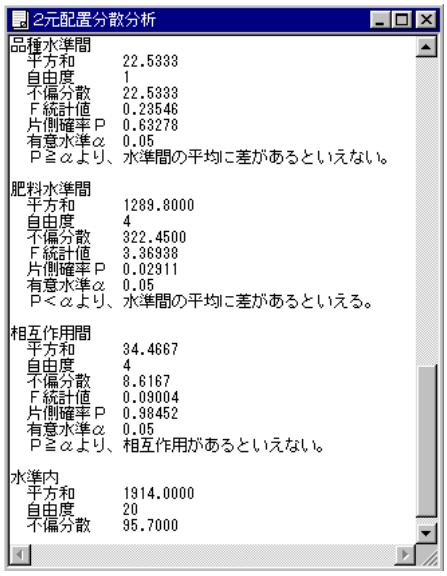
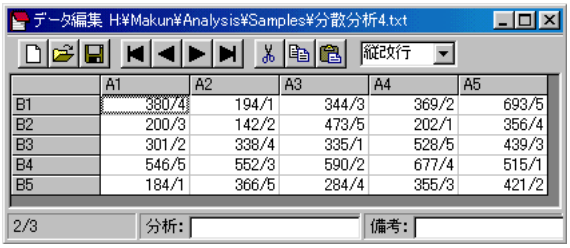


図 8a　2 元配置分散分析（繰り返しあり）

	平方和	自由度	不偏分散	F値	確率値
全変動	3.2608E+03	29			
品種水準間	2.2533E+01	1	2.2533E+01	0.2355	0.6328
肥料水準間	1.2898E+03	4	3.2245E+02	3.3694	0.0291
相互作用間	3.4467E+01	4	8.6167E+00	0.0900	0.9845
水準内	1.9140E+03	20	9.5700E+01		

図 8b　2 元配置分散分析表（繰り返しあり）

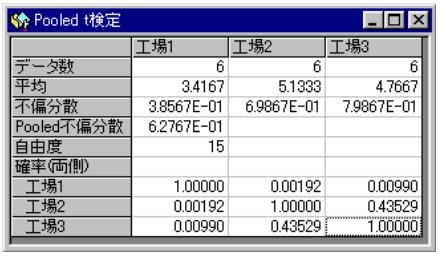
データの処理順序の差も検出したい場合、ラテン方格法を利用する。これには処理順序を入力しておく必要があるため、データに加えて順序を「データ/順序」のように / で区切って入力する。このデータ形式の例を図 9 に示す。出力は水準、ブロック、配置間の差を検定した結果を、図 6a と図 6b のようにテキストと分散分析表の 2 種類で表示するが、具体的な画面については省略する。



	A1	A2	A3	A4	A5
B1	380/4	194/1	344/3	369/2	693/5
B2	200/3	142/2	473/5	202/1	356/4
B3	301/2	338/4	335/1	528/5	439/3
B4	546/5	552/3	590/2	677/4	515/1
B5	184/1	366/5	284/4	355/3	421/2

図 9 ラテン方格法データ例

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 検定と結合された順位による Wilcoxon の順位和検定の出力結果をそれぞれ図 10 と図 11 に示す。



	工場1	工場2	工場3
データ数	6	6	6
平均	3.4167	5.1333	4.7667
不偏分散	3.8567E-01	6.9867E-01	7.9867E-01
Pooled不偏分散	6.2767E-01		
自由度	15		
確率(両側)			
工場1	1.00000	0.00192	0.00990
工場2	0.00192	1.00000	0.43529
工場3	0.00990	0.43529	1.00000

図 10 pooled t 検定出力結果

	工場1	工場2	工場3
データ数	6	6	6
順位和	25.000	80.000	66.000
確率(両側)			
工場1	1.00000	0.00350	0.03055
工場2	0.00350	1.00000	0.48208
工場3	0.03055	0.48208	1.00000

図 11 pooled Wilcoxon 検定出力結果

２．重回帰分析

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。

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

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。

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

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

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

具体的な分析画面を図 1、データを図 2 に示す。変数選択で、全てのデータを選択する。

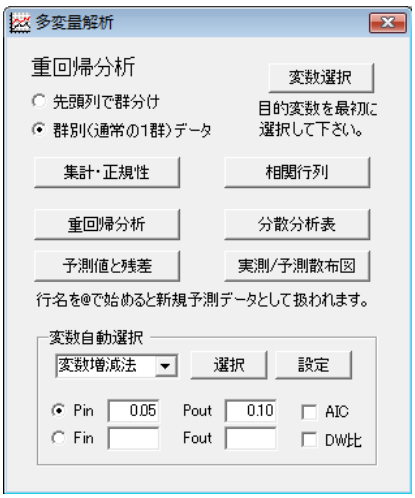


図 1 重回帰分析メニュー画面

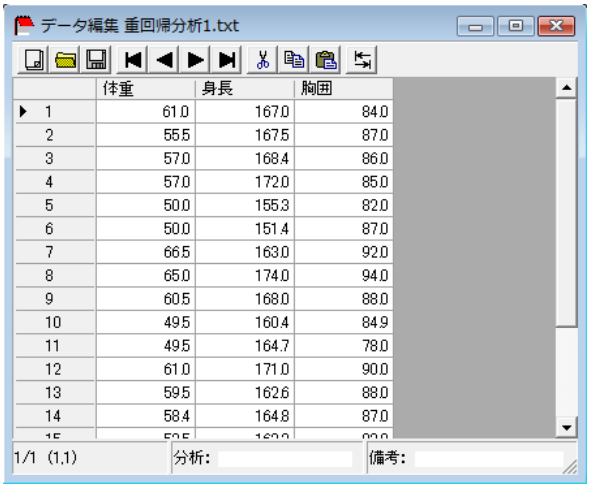


図 2 重回帰分析データ

「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。

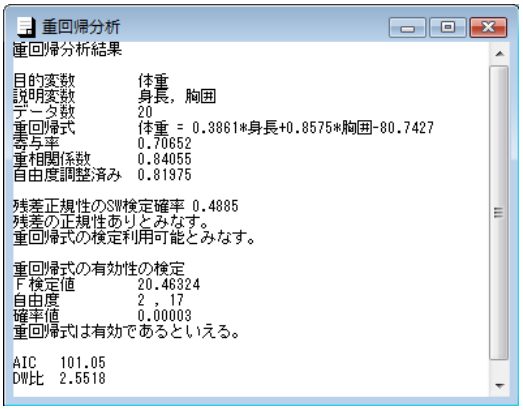


図 3a 重回帰分析出力画面 1

図 3b 重回帰分析出力画面 2

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。

	平方和	自由度	不偏分散	F検定値
▶ 全変動	462.4055	19		20.4632
回帰変動	326.7009	2	163.3504	確率値
残差変動	135.7046	17	7.9826	0.0000

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、残差が示される。

	実測値	予測値	残差
▶ 1	61.0	55.762	5.238
2	55.5	58.528	-3.028
3	57.0	58.018	-1.018
4	57.0	58.550	-1.550
5	50.0	49.530	0.470
6	50.0	52.312	-2.312
7	66.5	61.078	5.422
8	65.0	67.040	-2.040
9	60.5	59.579	0.921
10	49.5	53.986	-4.486

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。

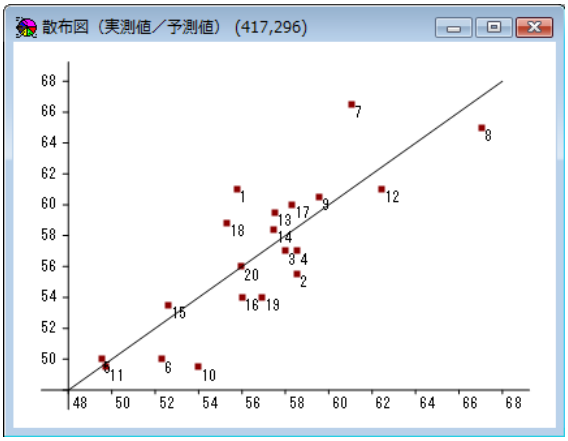
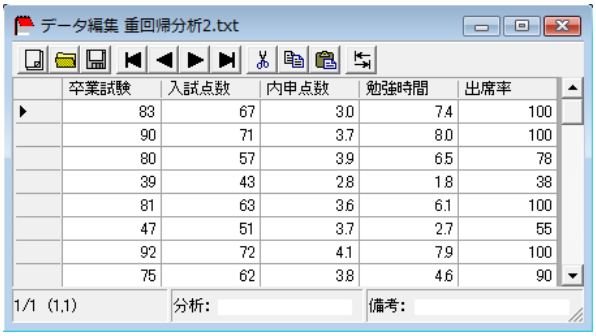


図 6 実測値と予測値の散布図

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。

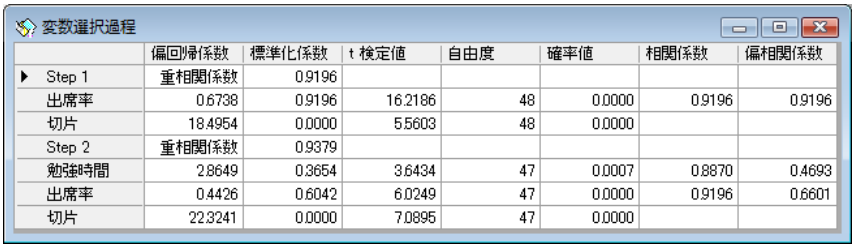


	卒業試験	入試点数	内申点数	勉強時間	出席率
▶	83	67	3.0	7.4	100
	90	71	3.7	8.0	100
	80	57	3.9	6.5	78
	39	43	2.8	1.8	38
	81	63	3.6	6.1	100
	47	51	3.7	2.7	55
	92	72	4.1	7.9	100
	75	62	3.8	4.6	90

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」左側のラジオボックスをチェックすると検定確率で指定し「Fin」左側のラジオボックスをチェックすると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。

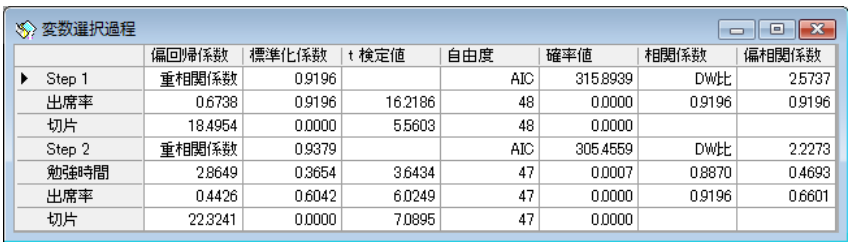
変数の選択法として、変数増加法、変数減少法、変数増減法のどれかを選び、「選択」ボタンをクリックすると図８のように選択過程での種々の統計量が表示される。



	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

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じ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。我々はこの問題について、１変数の回帰分析では分類別に回帰分析を行うプログラムを開発していたが、多変数の重回帰分析では今回新たに機能を追加した。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。ここで群による違いを明確にするために、故意に説明変数は両群同じ値にしている。

	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

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。

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

	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。

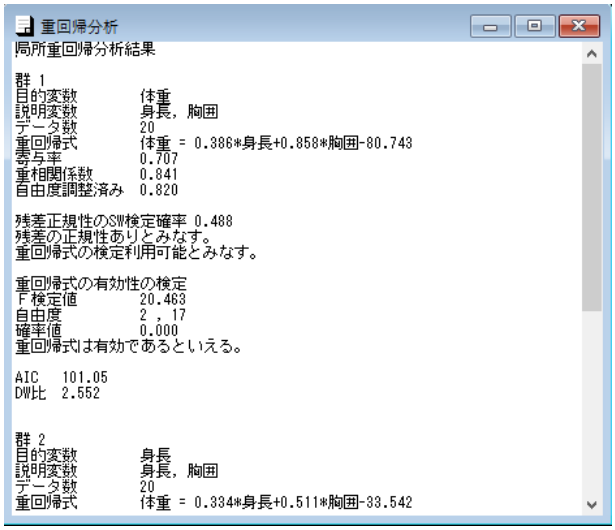
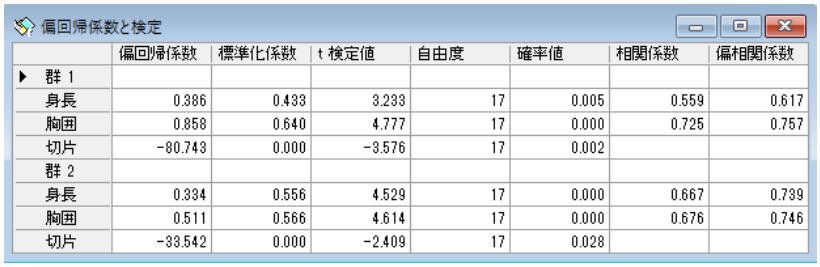


図 12a 群分けした重回帰分析結果１



	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
▶ 群 1							
身長	0.386	0.433	3.233	17	0.005	0.559	0.617
胸囲	0.858	0.640	4.777	17	0.000	0.725	0.757
切片	-80.743	0.000	-3.576	17	0.002		
群 2							
身長	0.334	0.556	4.529	17	0.000	0.667	0.739
胸囲	0.511	0.566	4.614	17	0.000	0.676	0.746
切片	-33.542	0.000	-2.409	17	0.028		

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。

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、上の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一致しており、重なって表示されている。

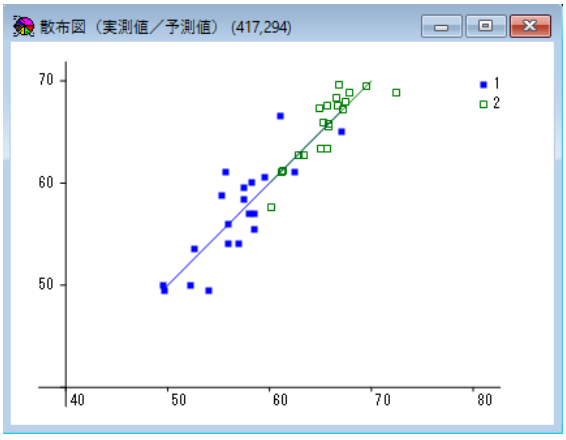


図 15 群分けされた実測値／予測値散布図

３．判別分析

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。

3.1 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。

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

これより、損失を最小にするためには 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。

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

ここに、判別の分点は 0 である。

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

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

これらを成分とする平均ベクトル $\overline{\mathbf{x}}^{\alpha}$ 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。

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

これを判別関数に代入して以下の線形判別関数を得る。

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

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

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

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。

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

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 $\overline{\mathbf{x}}^{\alpha}$ 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。

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

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 $(\overline{\mathbf{x}}^1 + \overline{\mathbf{x}}^2)/2$ 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。

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

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

$$F_i=\frac{(n_1+n_2-p-1)n_1n_2(D^2-D_i^2)}{(n_1+n_2)(n_1+n_2-2)+n_1n_2D_i^2}\sim F_{1,n_1+n_2-p-1}\text{分布}$$

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。

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

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。

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

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

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。

3.2 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。

$$\mathbf{Ba} = \rho\mathbf{Sa} \tag{5}$$

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。

$$\mathbf{Au} = \rho\mathbf{u} \tag{6}$$

${}^t\mathbf{uu} = 1$ 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。

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

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。

$$z^{(r)} = {}^t\mathbf{u}\mathbf{c}^{(r)} + {}^t\bar{\mathbf{x}}\mathbf{a}^{(r)} - {}^t\tilde{\mathbf{x}}\mathbf{a}^{(r)}, \qquad c_i^{(r)} = a_i^{(r)} s_i \tag{8}$$

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。

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

3.3 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。

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

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

これより、解 $\mathbf{a}$ を求めると以下となる。

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

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。

3.4　ソフトウェアの利用法

メニュー［分析－多変量解析等－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。

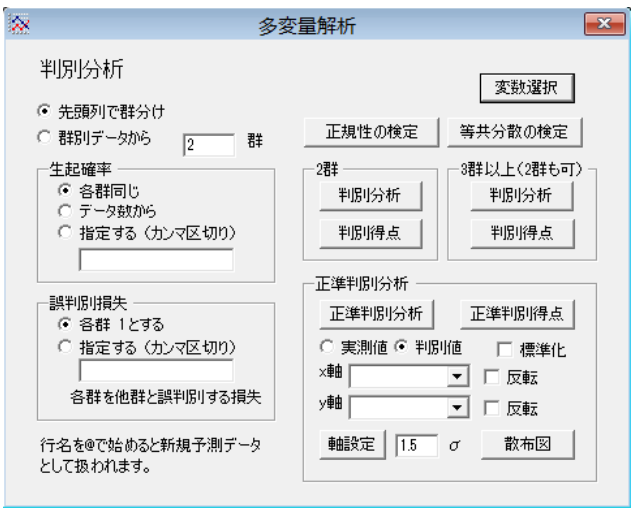


図　1 判別分析画面

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。

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。

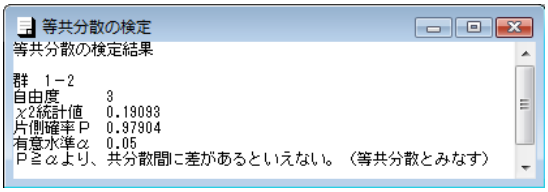


図 2　等共分散の検定

	勉強時間	平均点	定数項
判別関数	2.2461	0.2007	-23.0187
標準化係数	2.6210	2.2787	-0.3788
F検定値	19.8822	15.0274	
自由度	1,27	1,27	
確率	0.0001	0.0006	
マハラノビスの距離	5.6823		
誤判別確率	1群を2群と	2群を1群と	
理論から	0.1167	0.1167	
実測から	0.0769	0.0588	
判別数（実＼予）	1群	2群	
1群	12	1	
2群	1	16	
判別確率(実＼予)	1群	2群	
1群	0.9231	0.0769	
2群	0.0588	0.9412	

図 3　判別分析実行画面（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。

	所属群	判別得点	判別群
6	1	0.3280	1
7	1	-0.7743	2
8	1	4.9054	1
9	1	1.3153	1
10	1	1.8934	1
11	1	1.0704	1
12	1	4.0450	1
13	1	2.2301	1
14	2	-4.8682	2
15	2	-0.0469	2
16	2	-0.9540	2
17	2	-2.1784	2

図 4　判別得点（2 群の形式）

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

	勉強時間	平均点	定数項
1群判別関数	8.7369	1.0833	-61.8513
2群判別関数	6.4908	0.8826	-38.8327
1群標準化係数	10.1951	12.2975	47.1974
2群標準化係数	7.5741	10.0189	47.5762
マハラノビスの距離	1群	2群	
1群	0.0000	5.6823	
2群	5.6823	0.0000	
誤判別確率	1群を他群と	2群を他群と	
実測から	0.0769	0.0588	
判別数 (実＼予)	1群	2群	
1群	12	1	
2群	1	16	
判別確率(実＼予)	1群	2群	
1群	0.9231	0.0769	
2群	0.0588	0.9412	

図 5　判別分析実行画面（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）

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。同じデータを用いた結果を図 7 に示す。

正準判別分析

	勉強時間	平均点	定数項
判別1	0.9423	0.0842	-9.6565
標準化1	1.0995	0.9559	-0.1589
	固有値	寄与率	累積寄与率
判別1	1.4950	1.0000	1.0000
判別の分点	0		
	1群を他群と	2群を他群と	
誤判別確率	0.0769	0.0588	

図 7　正準相関分析

生起確率が同じで誤判別損失が 1 の場合、2 群の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。

正準判別分析の判別得点では、図 8 のように最後に群別得点平均が付く。これは 3 群以上の場合でも同様である。

判別得点

	所属群	判別得点1	判別群
25	2	-1.8352	2
26	2	-2.3991	2
27	2	-2.4203	2
28	2	-1.8778	2
29	2	-0.4873	2
30	2	-2.0510	2
群別得点平均	1	1.1919	
	2	-1.1919	

図 8　正準判別分析の判別得点

次に 3 群以上の正準判別分析の結果を図 9 に示す。

正準判別分析

	がくの長さ	がくの幅	花弁の長さ	花弁の幅	定数項
▶ 判別1	0.8294	1.5345	-2.2012	-2.8105	2.1051
判別2	0.0241	2.1645	-0.9319	2.8392	-6.6615
標準化1	0.6868	0.6688	-3.8858	-2.1422	0.0000
標準化2	0.0200	0.9434	-1.6451	2.1641	0.0000
	固有値	寄与率	累積寄与率		
判別1	32.1919	0.9912	0.9912		
判別2	0.2854	0.0088	1.0000		

図 9　正準判別分析結果

ここでは標準化係数が 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偶然で、一般には 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。

判別得点

	所属群	判別得点1	判別得点2
▶ 1	1	8.0618	0.3004
2	1	7.1287	-0.7867
3	1	7.4898	-0.2654
4	1	6.8132	-0.6706
5	1	8.1323	0.5145
6	1	7.7019	1.4617
7	1	7.2126	0.3558
8	1	7.6053	-0.0116
9	1	6.5606	-1.0152
10	1	7.3431	-0.9473

図 10　正準判別分析の判別得点

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。

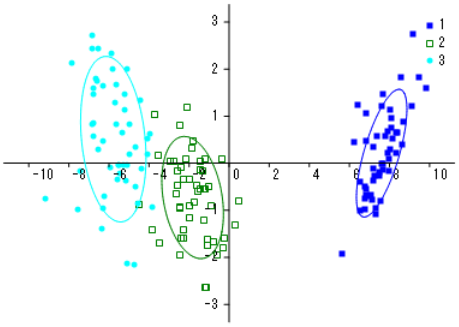


図 11　判別得点散布図

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。

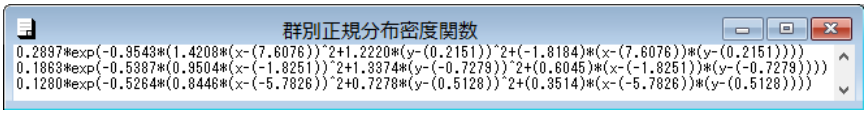


図 12　2 変量正規分布密度関数式

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]）、（範囲を設定、分割数を増加、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。

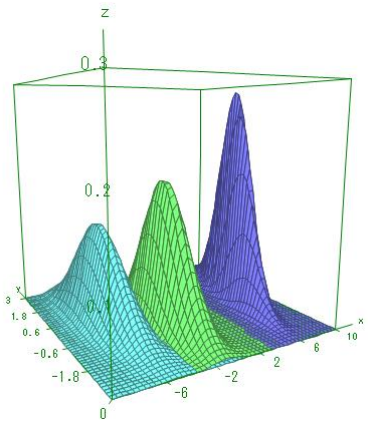


図 13　確率密度関数の視覚化

謝辞

正準判別分析とその表示方法については、岩村忠昭氏にいろいろと助言をいただきました。有難うございました。

４．主成分分析

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。

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

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 a 固有値 λ_a に対する固有ベクトル $\mathbf{u}^a$ の成分を以下のように表わす。

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

固有値 λ_a は第 a 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 a 主成分の分散の割合 c_a は以下で与えられ、寄与率と呼ばれる。

$$ c_a = \lambda_a \Big/ \sum_{i=1}^{p} \lambda_i $$

因子負荷量 r_{ai} は第 a 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。

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

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^a は個体毎の第 a 主成分の値として以下のように定義される。

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

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。

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

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。

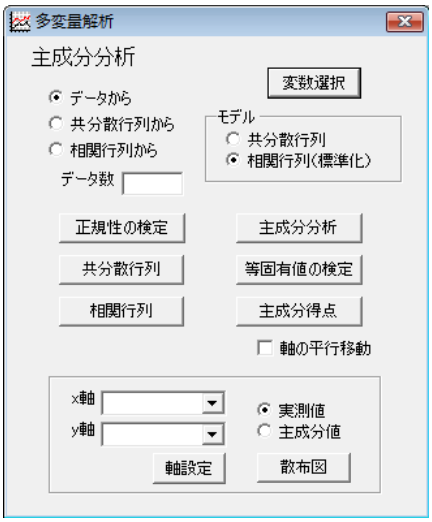


図 1 主成分分析のメニュー

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	3.5411	0.3134	0.0794	0.0661
寄与率	0.8853	0.0783	0.0199	0.0165
累積寄与率	0.8853	0.9636	0.9835	1.0000
固有ベクトル				
身長	0.4970	-0.5432	0.4496	0.5057
体重	0.5146	0.2102	0.4623	-0.6908
胸囲	0.4809	0.7246	-0.1752	0.4615
座高	0.5069	-0.3683	-0.7439	-0.2323
因子負荷量				
身長	0.9352	-0.3041	0.1267	0.1300
体重	0.9683	0.1177	0.1303	-0.1776
胸囲	0.9049	0.4056	-0.0494	0.1187
座高	0.9539	-0.2062	-0.2096	-0.0597

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
χ2値	67.0395	10.1275	0.1093
自由度	9	5	2
等固有値確率	0.00000	0.07170	0.94683
利用可能性	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番に並べた場合、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	-0.0687	0.2341	0.3491	-0.2616
2	2.8001	-0.3830	0.0957	-0.2748
3	2.6936	-0.0169	-0.3541	0.3526
4	1.3972	0.0595	-0.2074	-0.0435
5	0.9189	0.5749	0.0867	0.1780
6	-2.7897	-0.3429	-0.0325	-0.0306
7	2.4015	0.1649	0.4613	-0.1602
8	-2.7662	0.3126	0.0324	-0.2183
9	1.5295	1.6757	0.3257	0.0074
10	2.4794	-0.9564	-0.1196	-0.3841
11	0.7829	-0.1603	-0.1257	-0.2892

図 4 主成分得点出力結果

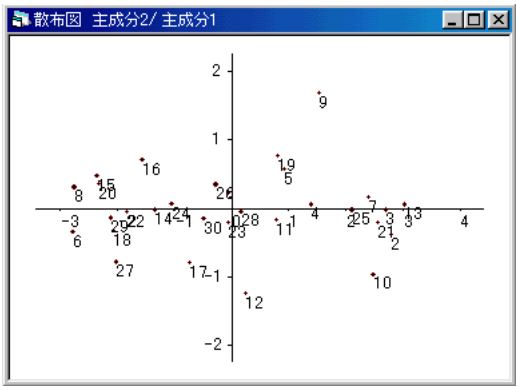


図 5 主成分得点散布図

５．因子分析

因子分析が取り扱うデータは主成分分析等と同様に p 変数、n 個体（レコード）の変量 $x_{i\lambda}$（$i=1,2,\cdots,p,\ \lambda=1,2,\cdots,n$）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 $t_i=(x_i-\overline{x}_i)/u_i$ 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 $\overline{x}_i$ 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。

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

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。

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

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。

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

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。

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

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

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

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

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、その求め方にもセントロイド法、主因子法、主成分分析法、最尤法、最小 2 乗法等種々の方法があるが [9]、ここでは歴史的に有名なセントロイド法と広く知られている主因子法、主成分分析法を取り上げる。

セントロイド法と主因子法では、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。主成分分析法では、相関行列の固有ベクトルをそのまま推定値として利用し、必要な次元までを採用する。以後詳しく見て行く。

セントロイド法は第 1 因子から逐次因子負荷量を求めていく手法で、

$$a_{i1} = \sum_{j=1}^{p} r_{ji} \Big/ \sqrt{\sum_{j=1}^{p}\sum_{k=1}^{p} r_{jk}} \qquad (r_{ii} = h_i)$$

の形で第 1 因子の因子負荷量を与える。次に $r_{ij}^{(1)} = r_{ij} - a_{i1}a_{j1}$ として新たな相関行列を定義するが、その際対角要素は各行の非対角要素の絶対値の最大値を用い、負の相関係数をできるだけ少なくするために、参考文献 8) のアルゴリズムに従い座標反転を行なう。この相関行列を利用して新たに第 2 因子の因子負荷量を同様の方法で計算する。

$$a_{i2} = \sum_{j=1}^{p} r_{ji}^{(1)} \Big/ \sqrt{\sum_{j=1}^{p}\sum_{k=1}^{p} r_{jk}^{(1)}}$$

さらに $r_{ij}^{(2)} = r_{ij}^{(1)} - a_{i2}a_{j2}$ を用いて新たな相関行列を作り、上に述べた方法で対角要素と負の相関についての処理を行ない、次の因子の因子負荷量を計算して行く。

次に主因子法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によって因子負荷量 $a_{i\alpha}$ が計算される。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、

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

のように与えられる。

主成分分析法は、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、

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

のように与える。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。そこで我々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。

$$\mathbf{F} = \mathbf{T}\mathbf{D}^{-1}\mathbf{A}({}^t\mathbf{A}\mathbf{D}^{-1}\mathbf{A})^{-1}$$

この推定法はBartlettの重みつき最小2乗推定法と呼ばれる。

この他にも回帰推定法と呼ばれるものがある。(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。

$$\hat{f}_{\alpha\lambda} = \frac{1}{\lambda_\alpha}\sum_{i=1}^{p} a_{i\alpha} t_{i\lambda} = \sum_{i=1}^{p}\sum_{j=1}^{p} r^{ij} a_{j\alpha} t_{i\lambda} \equiv \sum_{i=1}^{p} b_{\alpha i} t_{i\lambda} \qquad (3)$$

ここで、$\mathbf{R}\mathbf{a}_\alpha = \lambda \mathbf{a}_\alpha \Leftrightarrow \mathbf{R}^{-1}\mathbf{a}_\alpha = (1/\lambda_\alpha)\mathbf{a}_\alpha$ 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。

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

この関係は、(3)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。

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。

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

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。

最後に、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。

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

その後、参考文献 [1]を用いて、プロマックス回転についてプログラムを作成したので、追加しておく。

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。

ステップ１

直交回転後の因子負荷行列 $\mathbf{A}$ から始める。

$\mathbf{A}$ 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。

絶対値最大の２乗和が±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。

$\mathbf{A}$ 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

プライマリ因子軸と直交する（参考因子軸に沿う）成分である因子構造行列 $\mathbf{S}_p$、参考因子軸と直交する（プライマリ因子軸に沿う）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。

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

因子分析の実際の実行画面を図１に示す。データとしては主成分分析と同じように個体毎の元データ、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、歴史的なセントロイド法、主因子法、主成分分析が利用できる。いずれも共通性の推定の不完全さを補うために、共通性の値が一定値に近づくまで、近似計算を繰り返す。

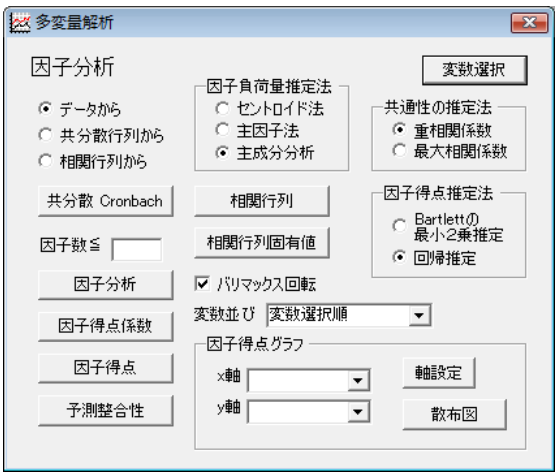


図１ 因子分析画面

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画面を示す。


因子負荷量

	因子1	因子2	共通性
国語	0.0745	-0.7869	0.6248
英語	0.2205	-0.7367	0.5914
社会	0.1566	-0.7625	0.6059
数学Ⅰ	0.8511	-0.2033	0.7657
数学Ⅱ	0.8451	-0.2964	0.8021
理科	0.8890	-0.0397	0.7919
寄与率	0.3846	0.3124	
累積寄与率	0.3846	0.6970	
符号調整済α	0.6691	0.5381	

図 2 因子分析出力画面

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、因子数を指定しない場合は、セントロイド法で累積寄与率が 0.9 を超えたところで、主因子法では固有ベクトルの値が 0.5 未満になったところで因子の出力を停止する。また、因子数を指定した場合でも、主因子法で固有値が 0 に近い負の値を取ることも見つかっており、指定した個数より少なく表示される場合もある。この原因は現在考察中である。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、Bartlett の重みつき最小 2 乗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
37	1.0797	1.3139
38	0.7078	-1.6708
39	-1.0450	0.0860
40	2.5198	0.5770
41	0.5220	-0.1223
42	-0.4754	-0.9796
43	0.2715	-1.4933
44	-0.4913	-1.1754
45	-0.3260	-0.0892
46	0.0810	-0.1131
47	-0.4149	-0.6058
48	-0.6418	0.9986
49	-1.0654	0.4957
50	0.9558	1.8029
平均	0.0000	0.0000
標準偏差	0.9899	0.9899

図 3 因子得点出力画面

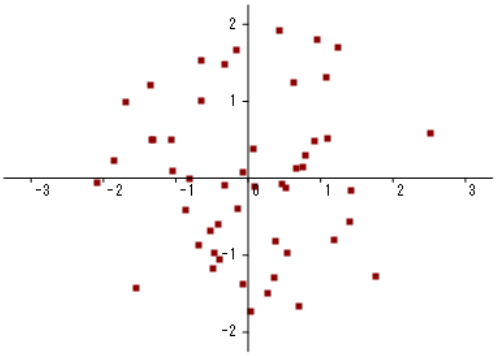


図 4 因子得点散布図

新しくバリマックス回転の機能を追加したが、それ以外に因子負荷量推定法に主成分分析を、因子得点の推定法として回帰推定も追加した。これらはよく利用されているのでデフォルトで、使うように設定している。

「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。その実行画面を図 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 3.3 実測値と予測値の比較画面

因子分析のバージョンアップで、因子負荷量推定法に主成分分析を加えたことは前に述べたが。これによって因子数を変数の数まで任意に選ぶことができるようになり、主成分分析の同じ主成分数の場合と累積寄与率が等しくなる。また、他の推定法に比べても累積寄与率の値は向上する。その他に、出力変数の並びをこれまでの変数選択順の他に、因子負荷量の大きさで２通りに並べ替える方法を加えた。これによって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るようになり、因子の解釈がより容易になる。

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，1995.

６．クラスター分析

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。

$$s_i^2 = \frac{1}{n-1} \sum_{\lambda=1}^{n} (x_{i\lambda} - \bar{x}_i)^2 \ , \ \ (\mathbf{S})_{ij} = s_{ij} = \frac{1}{n-1} \sum_{\lambda=1}^{n} (x_{i\lambda} - \bar{x}_i)(x_{j\lambda} - \bar{x}_j)$$

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　$$d_{\mu\nu} = (ad-bc)/\sqrt{(a+b)(c+d)(a+c)(b+d)}$$

ここに、a,b,c,d は以下のように与えられる。

$$a=\sum_{i=1}^{p}x_{i\mu}x_{i\nu}\ ,\quad b=\sum_{i=1}^{p}x_{i\mu}(1-x_{i\nu})\ ,\quad c=\sum_{i=1}^{p}(1-x_{i\mu})x_{i\nu}\ ,\quad d=\sum_{i=1}^{p}(1-x_{i\mu})(1-x_{i\nu})$$

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　　$d_{ij}=1-s_{ij}/s_is_j$　　（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　　$D_{hl}^2=\frac{n_f}{n_h}D_{fl}^2+\frac{n_g}{n_h}D_{gl}^2-\frac{n_fn_g}{n_h^2}D_{fg}^2$

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　　$D_{hl}^2=\frac{1}{n_h+n_l}\left[(n_f+n_l)D_{fl}^2+(n_g+n_l)D_{gl}^2-n_lD_{fg}^2\right]$

但し、重心法、群平均法、ウォード法について、距離はユークリッド距離をとるものとする。

メニュー［分析－多変量解析－クラスター分析］を選択して表示される、クラスター分析の分析画面を図 1 に示す。

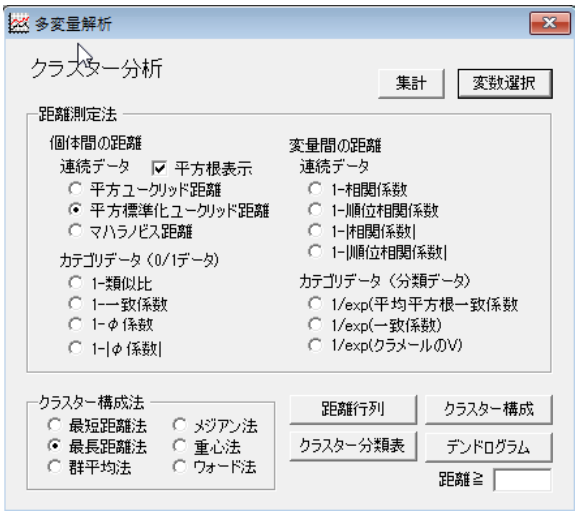


図 2 クラスター分析メニュー画面

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。

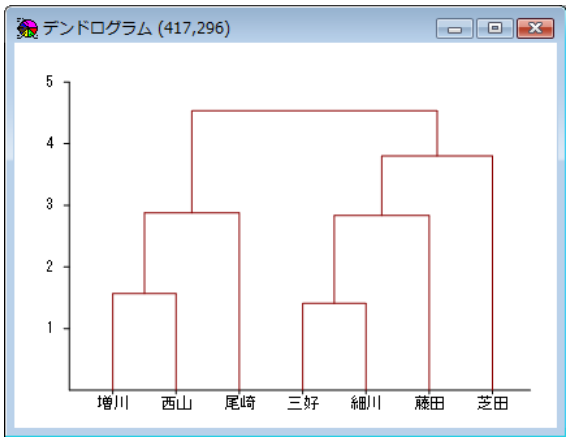


図 3 デンドログラム

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名、C の付いたものはクラスターである。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 2 分割するときに各個体がどちら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。


クラスター分類表

	並び	7	6	5	4	3	2	1
▶	増川	1	1	1	1	1	1	1
	西山	2	2	2	1	1	1	1
	三好	4	3	3	3	3	3	1
	芝田	7	4	4	4	4	4	1
	尾崎	3	5	5	5	5	1	1
	藤田	6	6	6	6	3	3	1
	細川	5	7	3	3	3	3	1

図 5 クラスター分類表

他の分析でも同様であるが、これまで予測値は欠損値データを除いて表示していたが、新しいデータを作成することを考えると欠損値を加えたままで表示し、元のデータに簡単に追加できるようにする方が賢明である。例えばこのクラスター分類表で、芝田のデータに欠損がある場合、図 6 の形式で表示すべきである。


クラスター分類表

	並び	6	5	4	3	2	1
▶	増川	1	1	1	1	1	1
	西山	2	2	2	1	1	1
	三好	3	3	3	3	3	1
	芝田						
	尾崎	6	4	4	4	3	1
	藤田	5	5	5	5	3	1
	細川	4	6	3	3	3	1

図 6 欠損値のある場合の分類表の表示

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。特に予測値の並びが変わった分析は、判別分析と数量化Ⅱ類である。これらは今まで群ごとに予測値を表示していたが、新たにデータ並びの順に表示するように作り変えた。

7．正準相関分析

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 $\mathbf{a}$，$\mathbf{b}$ で決まる変量 u，v を（第 1）正準変量、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 正準変量、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量 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。

$$Q_u = \sum_{j=1}^{s} r_{uj}^2 / s\,, \quad Q_v = \sum_{i=1}^{r} r_{vi}^2 / r$$

正準相関分析の実行画面を図１に示す。

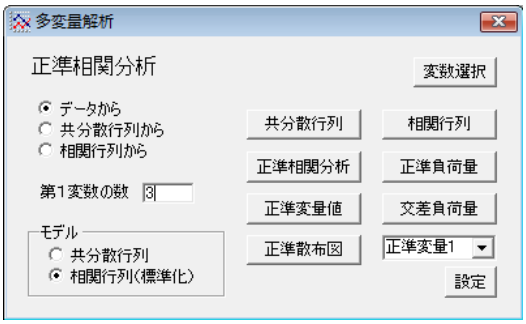


図１ 正準相関分析画面

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、正準変量の値と正準変量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、変数の大きさが相当違う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実行結果である。


正準相関分析

	正準変量 1	正準変量 2
正準相関係数	0.9560	0.3004
1群係数		
英語	1.1926	2.5235
国語	-0.0813	-2.3912
社会	-0.1494	-0.4650
2群係数		
数学	0.7392	-1.3634
理科	0.3141	1.5188

図２ 正準相関分析出力画面

この場合正準変量 u に含まれる変数の数として 3 を指定する。また、変数は同じ組の変数が並ぶように、選択順を調整する。結果は２つの正準変量の値と２つの正準相関係数の値を表示する。

次に図 3 に「正準変量の値」ボタンをクリックした場合の実行結果を示す。

	正準値1-1	正準値1-2	正準値2-1	正準値2-2
1	-0.4094	-0.2969	2.1992	1.2314
2	0.5306	0.4012	-0.4186	1.5912
3	1.0370	0.9764	-1.3427	-1.2820
4	-0.0323	0.3452	1.1427	-0.9415
5	1.0365	1.4806	-0.3974	-0.7291
6	1.0236	0.6838	-0.7758	-1.5660
7	0.5674	0.6696	-1.8596	-0.8808
8	-1.0215	-1.2806	-1.5344	0.2449
9	-1.2187	-0.7940	-0.1201	0.3359
10	-1.2154	-1.0970	-0.3800	-0.7526

図 3 正準変量の値画面

各個体毎に正準変量の値を計算して表示している。ここでは標準化されたデータから計算を進めたので、結果は標準化された値となる。これらのデータから第 1 正準変量について散布図を作ったものが、図 4 である。正準変量の選択は「設定」ボタンでできる。

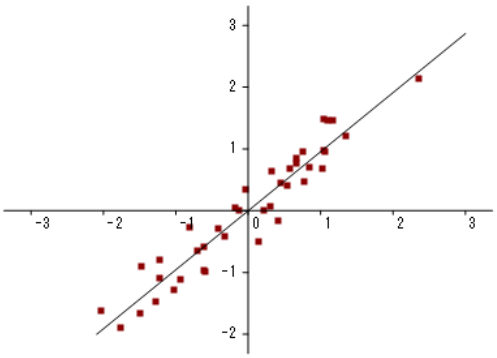


図 4 正準変量の散布図

第 1 正準変量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

８．数量化Ｉ類

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 1.1 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 1.1 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_11}	…	x_{p11}	…	x_{pr_p1}
y_2	x_{112}	…	x_{1r_12}	…	x_{p12}	…	x_{pr_p2}
:	:		:		:		:
y_n	x_{11n}	…	x_{1r_1n}	…	x_{p1n}	…	x_{pr_pn}

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。このうちの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 $\hat{a}_{ij}$ についての微係数を 0 として、以下の解を得る。

$$EV = \sum_{\lambda=1}^{n} (y_\lambda - Y_\lambda)^2 \quad \rightarrow \quad \hat{\mathbf{a}} = ({}^t\mathbf{XX})^{-1\,t}\mathbf{Xy}$$

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 ${}^t\mathbf{XX}$ 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。

$$ b'_{i(k)} = b_{i(k)} - c_i \ , \quad b'_0 = b_0 + \sum_{i=1}^{p} c_i $$

なぜなら、

$$ \sum_{i=1}^{p} \sum_{k=1}^{r_i} b'_{i(k)} x_{i(k)\lambda} + b'_0 = \sum_{i=1}^{p} \sum_{k=1}^{r_i} (b_{i(k)} - c_i) x_{i(k)\lambda} + b_0 + \sum_{i=1}^{p} c_i = \sum_{i=1}^{p} \sum_{k=1}^{r_i} b_{i(k)} x_{i(k)\lambda} + b_0 $$

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

$$ c_1 = -b_0 - \sum_{i=2}^{p} c_i \ , \quad c_i = b_{i(0)} \ (i \neq 1) $$

これにより、$b'_0 = 0,\ b'_{i(0)} = 0 \ (i \neq 1)$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 を用いて与えられる。。

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

$$R^2 = RV/SV = 1 - EV/SV, \quad R^{*2} = 1 - \frac{EV/(n - r_d - 1)}{SV/(n-1)}$$

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。

$$s_{ij} = \frac{1}{n-1} \sum_{\lambda=1}^{n} (X_{i\lambda} - \overline{X}_i)(X_{j\lambda} - \overline{X}_j), \quad s_{iy} = \frac{1}{n-1} \sum_{\lambda=1}^{n} (X_{i\lambda} - \overline{X}_i)(y_\lambda - \bar{y}),$$

$$s_{yy} = \frac{1}{n-1} \sum_{\lambda=1}^{n} (y_\lambda - \bar{y})^2$$

ここに、アイテム i の予測値 $X_{i\lambda}$ 及びその平均 $\overline{X}_i$ は以下で与えられる。

$$X_{i\lambda} = \sum_{j=1}^{r_i} \tilde{a}_{ij} x_{ij\lambda}, \quad \overline{X}_i = \frac{1}{n} \sum_{\lambda=1}^{n} X_{i\lambda}$$

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。

$$F = \frac{RV/r_d}{EV/(n - r_d - 1)} \qquad \text{自由度 } r_d,\ n - r_d - 1$$

実際の分析メニュー画面は図 1 に与える。

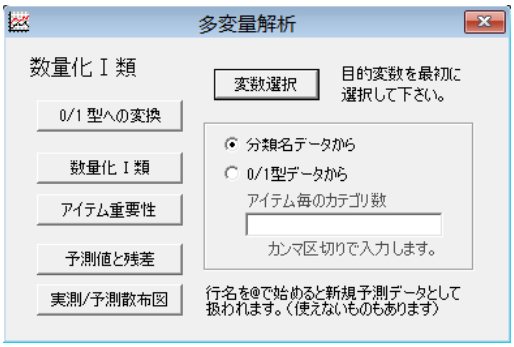


図 1 数量化 I 類メニュー画面

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

カテゴリウェイトと基準化された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。出力画面は図 3 に示す。

カテゴリウェイト

	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	3.5167	1.8917	0.0000	−0.3750	−1.4667	0.0000
重回帰 ウェイト	0.0000	−1.6250	0.0000	−0.3750	−1.4667	3.5167
基準化 ウェイト	0.4875	−1.1375	0.6992	0.3242	−0.7675	2.3300
重相関係数	0.9679	調整済	0.9514			
寄与率	0.9367	調整済	0.9051			
有効性F値	29.6205	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場合と同じ結果となる。有効性 F 値は、残差に正規性があるとは考えられないので、F 分布にはならず、p 値を求めることはできないが、参考のため F 分布の際の上側確率を与えている。

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

各アイテムが目的変数をどのように予測するかを個体毎に示すアイテムの予測値は「アイテム予測値」ボタンで図 5 のように示される。変更：この結果はカテゴリウェイトに依存するので、ボタンを削除した。

アイテム予測値

	観測値	地域	気候
1	3.0	3.517	-0.375
2	1.8	1.892	0.000
3	1.5	1.892	-0.375
4	3.3	3.517	-0.375
5	2.2	3.517	-1.467
6	2.0	3.517	-1.467
7	3.5	3.517	0.000
8	2.0	1.892	0.000
9	1.7	3.517	-1.467
10	2.3	3.517	-1.467

図 5 アイテム予測値

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 に示す。

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 6 予測値と残差

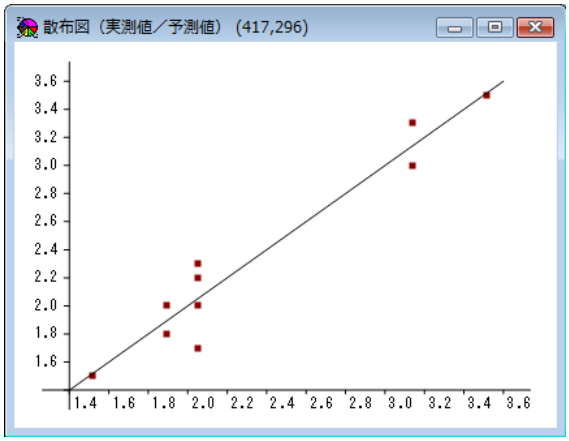


図 6 予測値と実測値の散布図

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.

９．数量化Ⅱ類

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。各変数には次の関係がある。

$$\sum_{j=1}^{r_i} x^\alpha_{ij\lambda} = 1 \tag{1}$$

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 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、$I \equiv \sum_{k=1}^{i-1}(r_k-1)+(j-1)$ で定義される。変数 I の範囲は $I=1,2,\cdots,P \equiv \sum_{k=1}^{p}(r_k-1)$ である。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

$$\sum_{I=1}^{P} f_I \quad \Leftrightarrow \quad \sum_{i=1}^{p}\sum_{j=1}^{r_i} f_{ij}$$

と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。

9.1 マハラノビスの距離に基づく方法

新しい変数表記法 $x_{I\lambda}^{\alpha}$ 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、3.3 節で述べたように、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。

$$z=\sum_{I=1}^{P}a_I x_I-\frac{1}{2}\sum_{I=1}^{P}(\bar{x}_I^1+\bar{x}_I^2)a_I \quad , \quad a_I=\sum_{J=1}^{P}(\mathbf{S}^{-1})_{IJ}(\bar{x}_J^1-\bar{x}_J^2) \tag{2}$$

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。

$$z^{\alpha}=\sum_{I=1}^{P}a_I^{\alpha}x_I-\frac{1}{2}\sum_{I=1}^{P}\bar{x}_I^{\alpha}a_I^{\alpha} \quad , \quad a_I^{\alpha}=\sum_{J=1}^{P}(\mathbf{S}^{-1})_{IJ}\bar{x}_J^{\alpha} \tag{3}$$

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。

$$\tilde{x}_{ik}=\frac{1}{m}\sum_{\alpha=1}^{m}\bar{x}_{ik}^{\alpha}$$

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　$z^{\alpha}=\sum_{i=1}^{p}\sum_{j=1}^{r_i}d_{ij}^{\alpha}x_{ij}+\sum_{i=1}^{p}\sum_{j=1}^{r_i}\tilde{x}_{ij}\hat{a}_{ij}^{\alpha}-\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{r_i}\bar{x}_{ij}^{\alpha}\hat{a}_{ij}^{\alpha}$ (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。

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

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。

9.2 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。

$$z_\lambda = \sum_{I=1}^{P} a_I x_I + z_0$$

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、$\overline{z}^\alpha = \dfrac{1}{n_\alpha}\sum_{\lambda=1}^{n_\alpha} z_\lambda^\alpha$ ， $\overline{z} = \dfrac{1}{n}\sum_{\alpha=1}^{m}\sum_{\lambda=1}^{n_\alpha} z_\lambda^\alpha$ ， $n = \sum_{\alpha=1}^{m} n_\alpha$ 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、

$$\mathbf{X} = \begin{pmatrix}
x_{121}^1 & \cdots & x_{1r_1 1}^1 & \cdots & x_{p21}^1 & \cdots & x_{pr_p 1}^1 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{12n_1}^1 & \cdots & x_{1r_1 1}^1 & \cdots & x_{p2n_1}^1 & \cdots & x_{pr_p n_1}^1 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{121}^m & \cdots & x_{121}^m & \cdots & x_{p21}^m & \cdots & x_{pr_p 1}^m \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{12n_m}^m & \cdots & x_{12n_m}^m & \cdots & x_{p2n_m}^m & \cdots & x_{pr_p n_m}^m
\end{pmatrix}$$

$$\overline{\mathbf{X}}_B = \begin{pmatrix}
\overline{x}_{12}^1 & \cdots & \overline{x}_{1r_1}^1 & \cdots & \overline{x}_{p2}^1 & \cdots & \overline{x}_{pr_p}^1 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\overline{x}_{12}^1 & \cdots & \overline{x}_{1r_1}^1 & \cdots & \overline{x}_{p2}^1 & \cdots & \overline{x}_{pr_p}^1 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\overline{x}_{12}^m & \cdots & \overline{x}_{12}^m & \cdots & \overline{x}_{p2}^m & \cdots & \overline{x}_{pr_p}^m \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\overline{x}_{12}^m & \cdots & \overline{x}_{12}^m & \cdots & \overline{x}_{p2}^m & \cdots & \overline{x}_{pr_p}^m
\end{pmatrix} \begin{matrix} \left.\begin{matrix}\\\\\\\end{matrix}\right\} n_1 \\ \vdots \\ \left.\begin{matrix}\\\\\\\end{matrix}\right\} n_m \end{matrix}$$

$$\bar{\mathbf{X}}=\begin{pmatrix} \bar{x}_{12} & \cdots & \bar{x}_{1r_1} & \cdots & \bar{x}_{p2} & \cdots & \bar{x}_{pr_p} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{12} & \cdots & \bar{x}_{1r_1} & \cdots & \bar{x}_{p2} & \cdots & \bar{x}_{pr_p} \end{pmatrix} \Bigg\} n$$

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。

$${}^t\mathbf{a}=(a_{12} \quad \cdots \quad a_{1r_1} \quad \cdots \quad a_{p2} \quad \cdots \quad a_{pr_p})$$

$$\mathbf{S}=\frac{1}{n-m}{}^t(\mathbf{X}-\bar{\mathbf{X}}_B)(\mathbf{X}-\bar{\mathbf{X}}_B),\quad \mathbf{B}=\frac{1}{n-m}{}^t(\bar{\mathbf{X}}_B-\bar{\mathbf{X}})(\bar{\mathbf{X}}_B-\bar{\mathbf{X}})$$

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。

$$z^{(r)}=\sum_{i=1}^{p}\sum_{j=1}^{r_i}d_{ij}^{(r)}x_{ij} \tag{9}$$

9.3 ソフトウェアの利用

メニュー［分析－多変量解析等－数量化Ⅱ類］を選択すると、数量化Ⅱ類のメニュー画面が図 1 のように表示される。

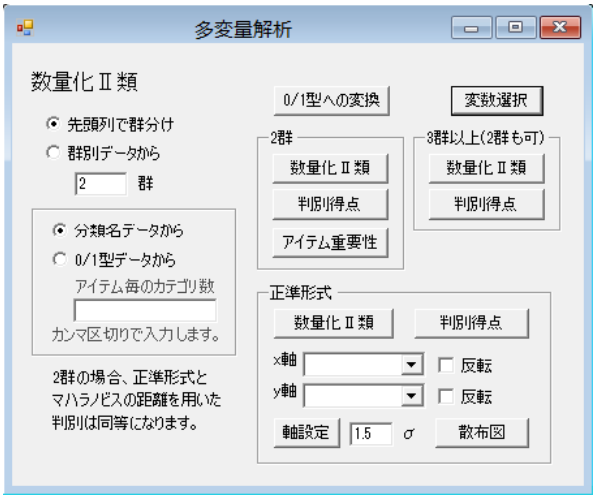
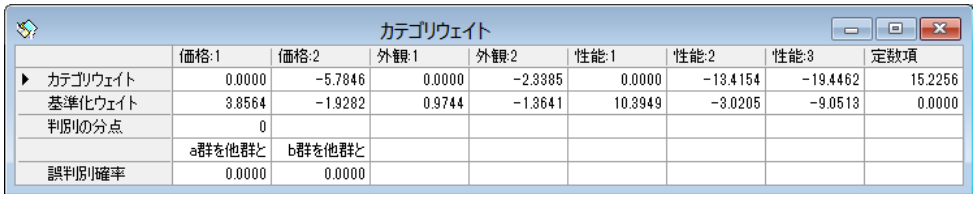


図 1 数量化Ⅱ類分析画面

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、上部に示された、参考文献 3) で与えられるマハラノビス形式と下部に示された、参考文献 4) で与えられる正準形式のどちらかを選択できる。正準形式は、これまでの計算結果を踏襲するものであるが、定義の違いから、係数について定数倍の違いがある。しかし、判別結果については同じである。マハラノビス形式は、2 群の場合、判別分析のところで示したように、正準形式と定数倍の違いを除いて同じである。しかし、3 群以上の場合では大きく異なり、判別分析と同様の結果を出力する。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。

「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。



	価格:1	価格:2	外観:1	外観:2	性能:1	性能:2	性能:3	定数項
▶ カテゴリウェイト	0.0000	-5.7846	0.0000	-2.3385	0.0000	-13.4154	-19.4462	15.2256
基準化ウェイト	3.8564	-1.9282	0.9744	-1.3641	10.3949	-3.0205	-9.0513	0.0000
判別の分点	0							
	a群を他群と	b群を他群と						
誤判別確率	0.0000	0.0000						

図 2a　マハラノビス形式のカテゴリウェイト

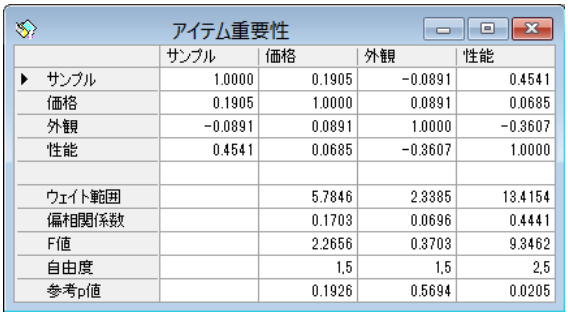
カテゴリウェイト

	価格:1	価格:2	外観:1	外観:2	性能:1	性能:2	性能:3	定数項
▶ 判別1	0.0000	-1.4636	0.0000	-0.5917	0.0000	-3.3943	-4.9202	3.8524
基準化1	0.9757	-0.4879	0.2465	-0.3451	2.6301	-0.7642	-2.2901	0.0000
	固有値	寄与率	累積寄与率					
判別1	4.6862	1.0000	1.0000					
判別の分点	0							
	a群を他群と	b群を他群と						
誤判別確率	0.0000	0.0000						

図 2b　正準形式のカテゴリウェイト

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。2 群の判別の場合、判別の分点は 0 にしている。2 つのカテゴリウェイトはそれぞれ比例している。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定義より 1 になる。

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、相関行列とそれを元に計算される偏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

以後は 3 群以上の場合を扱う。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

	吐き気:0	吐き気:1	吐き気:2	頭痛:0	頭痛:1	頭痛:2	定数項
1群判別関数	0.0000	3.2656	3.7813	0.0000	2.0625	1.7188	-0.5328
2群判別関数	0.0000	15.9844	20.5759	0.0000	8.9375	10.0670	-12.7114
3群判別関数	0.0000	16.3281	22.0491	0.0000	10.3125	13.3080	-15.8739
1群基準化関数	-2.5495	0.7161	1.2318	-1.2432	0.8193	0.4755	3.2599
2群基準化関数	-13.0993	2.8850	7.4766	-6.3913	2.5462	3.6757	6.7792
3群基準化関数	-13.6903	2.6379	8.3589	-8.0233	2.2892	5.2847	5.8397
	1群を他群と	2群を他群と	3群を他群と				
誤判別確率	0.2000	0.4000	0.2500				

図 5a マハラノビス距離を用いたカテゴリウェイト

カテゴリウェイト

	吐き気:0	吐き気:1	吐き気:2	頭痛:0	頭痛:1	頭痛:2	定数項
判別1	0.0000	-2.9339	-4.0012	0.0000	-1.7342	-2.3017	3.8454
判別2	0.0000	-2.0006	-1.4483	0.0000	0.3109	2.2173	0.3633
基準化1	2.4717	-0.4622	-1.5295	1.3737	-0.3605	-0.9280	0.0000
基準化2	1.3014	-0.6992	-0.1469	-0.9381	-0.6272	1.2793	0.0000
	固有値	寄与率	累積寄与率				
判別1	5.5682	0.9778	0.9778				
判別2	0.1263	0.0222	1.0000				

図 5b　正準形式を用いたカテゴリウェイト

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　マハラノビス距離を用いた判別得点

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　従来の方法による判別得点

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。

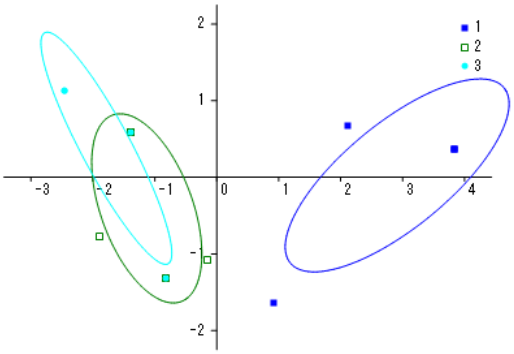


図 7　判別得点による散布図

１０．数量化Ⅲ類

数量化Ⅲ類はカテゴリと個体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。

$$\sum_{\lambda=1}^{n} x_{i\lambda} v_\lambda - 2\eta c_i u_i = 0\,,\ \sum_{i=1}^{p} x_{i\lambda} u_i - 2\mu d_\lambda v_\lambda = 0$$

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。

$$\mathbf{C}^{-1}\mathbf{XD}^{-1}\mathbf{X'u} = \rho^2\mathbf{u}\,,\ \text{また、}\ \mathbf{v} = \rho^{-1}\mathbf{D}^{-1}\mathbf{X'u} \tag{2}$$

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。

$$\mathbf{z} = \frac{1}{\sqrt{T}}\mathbf{C}^{1/2}\mathbf{u}\,,\ \text{ここに}\ \mathbf{C}^{1/2} = \begin{pmatrix} \sqrt{c_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{c_p} \end{pmatrix}$$

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

$$\bar{u}=\frac{1}{T}\sum_{i=1}^{p}c_i u_i=1\neq 0\ ,\quad \bar{v}=\frac{1}{T}\sum_{\lambda=1}^{n}d_\lambda=1\neq 0$$

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。

$$\tilde{\mathbf{u}}^{\alpha}=\rho_\alpha\mathbf{u}^{\alpha},\quad \tilde{\mathbf{v}}^{\alpha}=\rho_\alpha\mathbf{v}^{\alpha}$$

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。

$$\lambda_\alpha=\rho_\alpha^2/\sum_{\beta=1}^{p-1}\rho_\beta^2\quad(\alpha\neq p)$$

メインメニューの中の「分析−多変量解析−数量化Ⅲ類」メニューを選択すると図 1 に示される分析メニューが表示される。

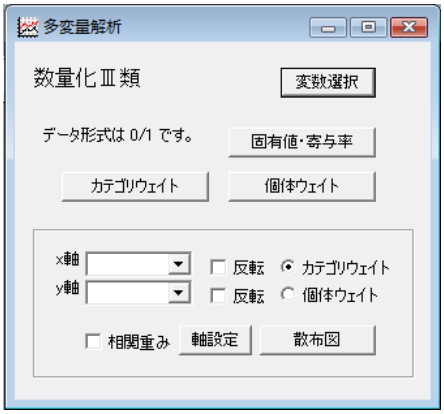
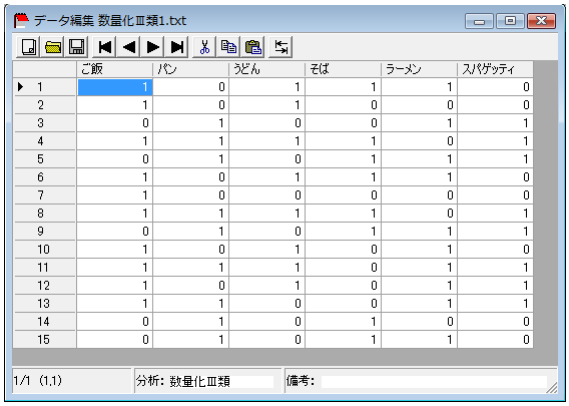


図 1 分析メニュー

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。



	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
10	1	0	1	0	1	0
11	1	1	1	0	1	1
12	1	0	1	0	1	1
13	1	1	0	0	1	1
14	0	1	0	1	0	0
15	0	1	0	1	1	0

図 2 分割表データ

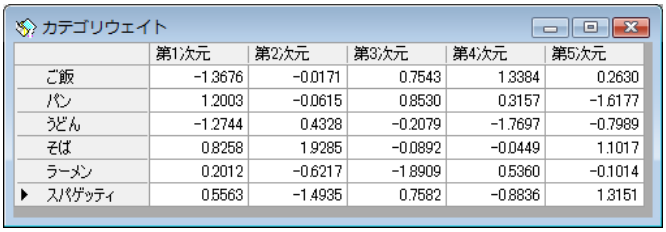
変数を選択して、「固有値・寄与率」ボタンをクリックすると図 3 のような結果が表示される。

	第1次元	第2次元	第3次元	第4次元	第5次元
固有値	0.3505	0.1556	0.0948	0.0605	0.0252
相関係数	0.5921	0.3945	0.3079	0.2459	0.1589
寄与率	0.5105	0.2267	0.1381	0.0880	0.0368
累積寄与率	0.5105	0.7371	0.8752	0.9632	1.0000

図 3 固有値・寄与率画面

ここで表示される固有値は、(3.2) 式のρ^2、相関係数は同じくρである。

図 1 の分析メニューで「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
ご飯	-1.3676	-0.0171	0.7543	1.3384	0.2630
パン	1.2003	-0.0615	0.8530	0.3157	-1.6177
うどん	-1.2744	0.4328	-0.2079	-1.7697	-0.7989
そば	0.8258	1.9285	-0.0892	-0.0449	1.1017
ラーメン	0.2012	-0.6217	-1.8909	0.5360	-0.1014
スパゲッティ	0.5563	-1.4935	0.7582	-0.8836	1.3151

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
1	-0.6819	1.0915	-1.1640	0.0608	0.7308
2	-2.2312	0.5268	0.8872	-0.8772	-1.6862
3	1.1022	-1.8392	-0.3028	-0.0432	-0.8476
4	-0.0201	0.4001	1.3434	-0.8493	0.3312
5	1.1754	-0.1573	-0.2995	-0.0780	1.0977
6	-0.6819	1.0915	-1.1640	0.0608	0.7308
7	-2.3099	-0.0435	2.4495	5.4433	1.6554
8	-0.0201	0.4001	1.3434	-0.8493	0.3312
9	1.1754	-0.1573	-0.2995	-0.0780	1.0977
10	-1.3742	-0.1741	-1.4554	0.1419	-1.3368
11	-0.2311	-0.8928	0.1732	-0.3768	-1.1830
12	-0.7957	-1.0770	-0.4760	-0.7920	1.0664
13	0.2492	-1.3903	0.3853	1.3285	-0.2218
14	1.7111	2.3661	1.2403	0.5508	-1.6237
15	1.2540	1.0521	-1.2200	1.0939	-1.2951

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。

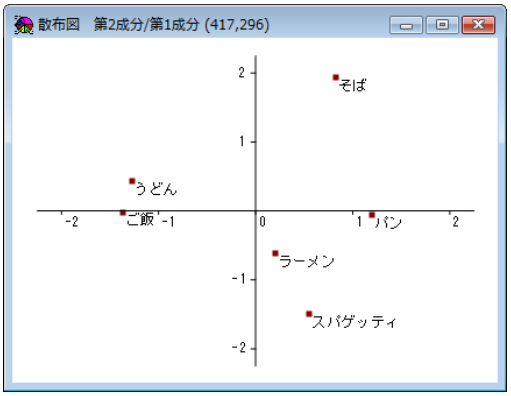


図 6 カテゴリウェイトの散布図

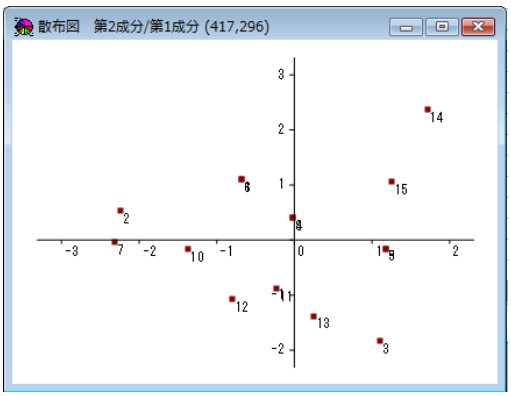


図 7 個体ウェイトの散布図

散布図の各成分には相関係数をかけて表示する場合があるが、その時には図 1 の「相関重み」チェックボックスにチェックを入れて散布図を表示する。また、成分を反転させて表示する場合は、反転チェックボックスにチェックを入れる。

１１．コレスポンデンス分析

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。

$$n_{i\bullet} \equiv \sum_{j=1}^{q} n_{ij}, \quad n_{\bullet j} \equiv \sum_{i=1}^{p} n_{ij}, \quad n \equiv \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}$$

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。

$$\rho = \frac{S_{uv}}{S_u S_v} \ ,$$

$$S_{uv} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} u_i v_j, \quad S_u^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i\bullet} u_i^2, \quad S_v^2 = \frac{1}{n}\sum_{j=1}^{q} n_{\bullet j} v_j^2$$

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。

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{p} n_{i\bullet} u_i = 0, \quad \bar{v} = \frac{1}{n}\sum_{j=1}^{q} n_{\bullet j} v_j = 0$$

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。

$$\sum_{k=1}^{q} n_{ik} v_k - 2\lambda n_{i\bullet} u_i = 0, \qquad \sum_{k=1}^{p} n_{kj} u_k - 2\mu n_{\bullet j} v_j = 0$$

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

$$\boldsymbol{N} = \begin{pmatrix} n_{11} & \cdots & n_{1q} \\ \vdots & \ddots & \vdots \\ n_{p1} & \cdots & n_{pq} \end{pmatrix}, \quad \boldsymbol{D}_r = \begin{pmatrix} n_{1\bullet} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & n_{p\bullet} \end{pmatrix}, \quad \boldsymbol{D}_c = \begin{pmatrix} n_{\bullet 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & n_{\bullet q} \end{pmatrix},$$

$$\boldsymbol{u}' = \begin{pmatrix} u_1 & \cdots & u_p \end{pmatrix}, \quad \boldsymbol{v}' = \begin{pmatrix} v_1 & \cdots & v_q \end{pmatrix}$$

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 $\boldsymbol{z}$ 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。

$$\lambda_\alpha = \rho_\alpha^2 / \sum_{\beta=1}^{p-1}\rho_\beta^2 \quad （\alpha \neq p）$$

メニュー「分析−多変量解析−コレスポンデンス分析」を選択すると図 1 に示される分析メニューが表示される。

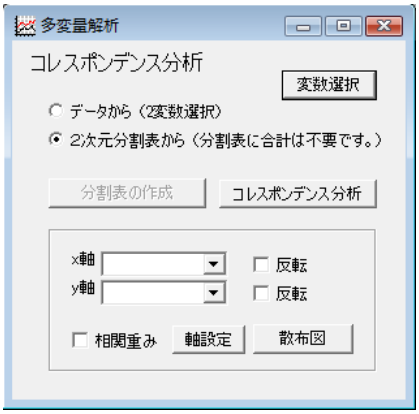


図 1 分析メニュー

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。

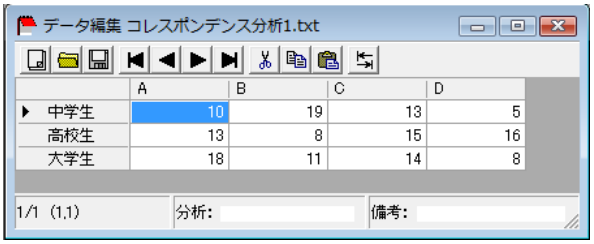


図 2 分割表データ

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		0.0763	0.0183		
相関係数		0.2762	0.1352		
中学生	1	-1.3287	-0.6528	-0.3670	-0.0883
高校生	1	1.1333	-0.7748	0.3130	-0.1048
大学生	1	0.0690	1.3916	0.0190	0.1882
A	2	0.2373	1.5238	0.0655	0.2060
B	2	-1.4691	-0.6411	-0.4058	-0.0867
C	2	0.0596	-0.1102	0.0165	-0.0149
D	2	1.5032	-1.1547	0.4152	-0.1561

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 11.4 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 11.5 のような結果が表示される。

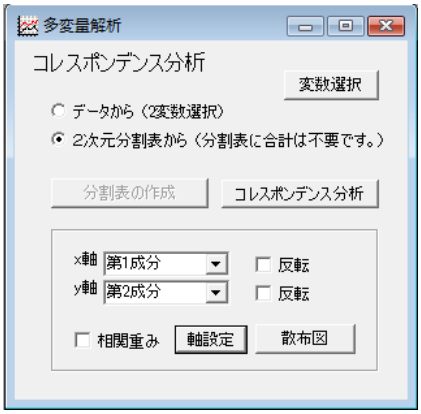


図 4 軸設定された分析メニュー

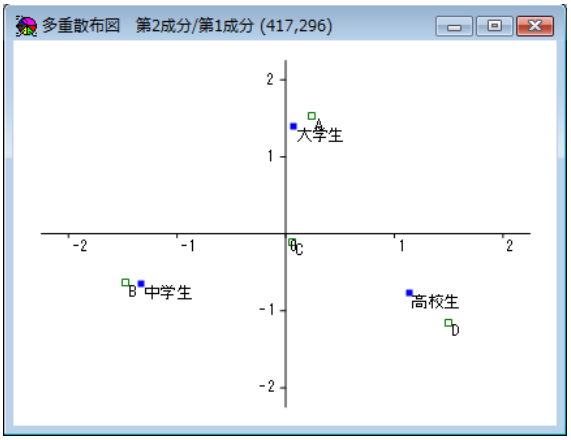


図 5 散布図画面

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。

１２．時系列分析

我々はこれまで教育用社会システム分析ソフトウェアの一環として様々なプログラムを作成してきたが、この論文では時系列分析について紹介する。時系列分析は時間の経過とともに変化する変数の過去のデータから、未来の値を予測する手法である。例えば企業の売上予測、在庫の受注予測、株価の変動など時系列的に変化するデータがこの分析の対象である。

分析方法には大きく分けて、古くから考えられてきた予測モデルという方法とデータの変動をいくつかの典型的な変動に分解する変動の分解モデルという方法がある。予測モデルには、予測値にこれまでの変動の差分を使う差の平均法、過去のデータにウェイトを付けて使う指数平滑法やブラウン法、過去の最も似た状況を探す最近隣法、重回帰分析を活用する ARIMA などがあるが、これらはデータ数が少なく周期性を見抜くことが困難なデータに適用されることが多い。

一方変動が周期性を持っているようなデータに対しては変動の分解モデルが適用される。これは変動を「傾向変動」、「季節変動」、「循環変動」、「残差」などに分け、それぞれの特徴をとらえて予測値を求めるもので、長期的な予測もある程度可能な手法である。傾向変動はデータの平均的な変動を表し、予測には移動平均や回帰を基礎とした近似モデルが利用される。一般に季節変動は周期が一定の変動で、循環変動は周期が変化する変動を表す。

本来予測モデルと変動の分解モデルは別々に考えられたものであるが、後者の傾向変動に例えば ARIMA の結果を利用するなどということも可能であるため、我々のプログラムでは２つの手法を組み合わせて使うことができるようにしている。本来変動の分解モデルの傾向変動については、移動平均や線形近似、対数近似などの近似手法が利用されることが多いので、傾向変動を２つに分けて、「傾向変動１」としてこれらの近似手法を、「傾向変動２」として先に述べた予測モデルを用いることにする。もちろんどちらか１つを選んでもよい。これらの分解の後、必要があればデータの周期的な変動の分解を行う。

周期的な変動には季節変動と循環変動があるが、循環変動についてはまだプログラムに組み込んでいない。また、季節変動を「振幅変動」と振幅が一定の「周期変動」の積に分解し、これらをまとめて以下のモデルとする。

データ変動＝傾向変動１＋傾向変動２＋振幅振動×周期変動＋残差

プログラムでは振幅変動の平均が１に近くなるように設定し、周期変動の意味を理解し易くしている。

12.1 時系列分析の方法

時間を過去から未来へ等間隔で区切ったとき、ある時点t（$t=1,\cdots,N$）でのある変数Xの値をx_tとする。時系列分析はこの変数の変化を分析し、モデルを作成して今後の予測を行うことを目的とする。以後このデータ書式を用いて予測モデルと変動の分解モデルの理論について説明する。

時系列分析では、データをそのままの形で使うより、何らかの変換を加えてから分析を進める方がよりはっきりとした結果を得られることがある。ここではよく利用されるデータの変換について述べる。

変数が値の増大とともに変動の大きさも大きくなっていくような場合は、元の変数の対数をとって新しい変数とすると分析が容易になる場合がある。また、比率や確率のように $[0,1]$ 区間の値の場合は、以下のロジット変換によって値域が $(-\infty,\infty)$ の時系列に変換できる。

対数変換　　$z_t = \log_e x_t$

ロジット変換　　$z_t = \log_e\left(\dfrac{x_t}{1-x_t}\right)$

また、時系列データの差分を使って新しい変数を作り出すことも行われる。

差分（i期）　　$z_t = x_t - x_{t-i}$

差分比（i期）　　$z_t = x_t / x_{t-i}$

12.2 予測モデル

時系列データの周期性が明らかでない場合やデータの数が周期性を見るのに十分でない場合、予測モデルと呼ばれる方法を用いて時系列データの予測が行われる。これからは図 2.1 のデータを用いて各種の予測モデルを紹介する。

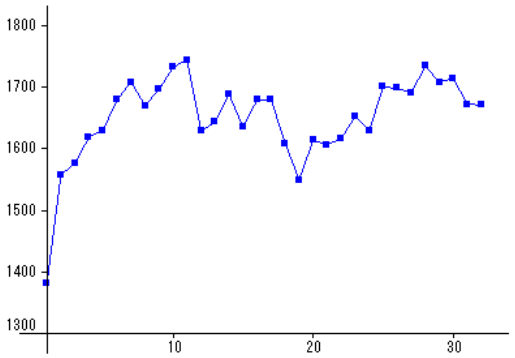


図 2.1 時系列データ

これらのモデルは基本的にt時点までのデータを元に$t+1$時点での予測値を求めるもので、長期の予測には向かない。

12.2.1 差の平均法（差分法）

これは$t+1$時点の予測値 y_{t+1} をt時点のデータ x_t とこれまでの 2 時点間の差分の平均で与えるものである。

$$y_{t+1} = x_t + A_t$$

ここに

$$A_t = \frac{(x_2 - x_1) + (x_3 - x_2) + \cdots + (x_t - x_{t-1})}{t-1} = \frac{x_t - x_1}{t-1}$$

差の平均法を用いた予測を図 2.2 に示す。これを見るとデータが上下している場合、残差の平均は相殺され、予測値は 1 期前の値と余り変わらない様子が見える。この手法はデータに上昇傾向や下降傾向が見られる場合に適用できる。

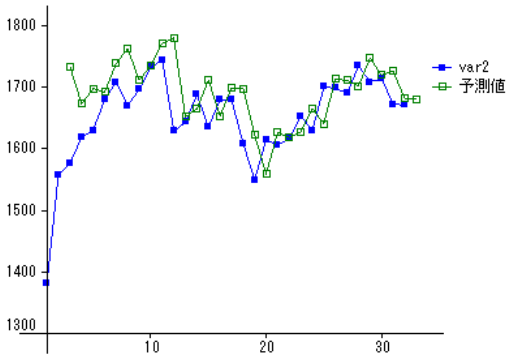


図 2.2 差の平均法を用いた予測

2 期以上の予測には実測値の代わりに予測値を使うことにすれば、予測は直線的に伸びて行く。

12.2.2 指数平滑法

この方法は$t+1$期の予測値y_{t+1}をt期の実測値x_tと予測値y_tを使って以下のように与えるものである。

$$y_{t+1} = \alpha x_t + (1-\alpha) y_t \qquad \text{但し、} \ y_1 = x_1 \ (\text{または} \ y_2 = x_1) \ \text{とする。}$$

ここにαは$0 < \alpha < 1$のパラメータである。またこの式は以下のように書き換えると、指数平滑の意味が分かり易い。

$$\begin{aligned} y_{t+1} &= \alpha x_t + (1-\alpha)[\alpha x_{t-1} + (1-\alpha) y_{t-1}] \\ &\cdots \\ &= \alpha x_t + \alpha(1-\alpha) x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \cdots + \alpha(1-\alpha)^{t-2} x_2 + (1-\alpha)^{t-1} x_1 \end{aligned}$$

これを見るとαの値が小さいほど過去からの影響を受けやすくなっていることが分かる。これは今期以前の指数平滑値を次期の予測値とするものである。この方法を用いて時系列データの変動をα=0.74 として予測した結果を図 2.3 に示す。パラメータの値は図 2.4 のようにパラメータの値を変えて残差の平均を調べ、最小値をとることによって求めた。

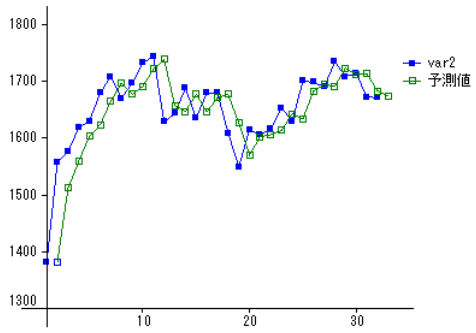


図 2.3 指数平滑法による予測（α=0.74）

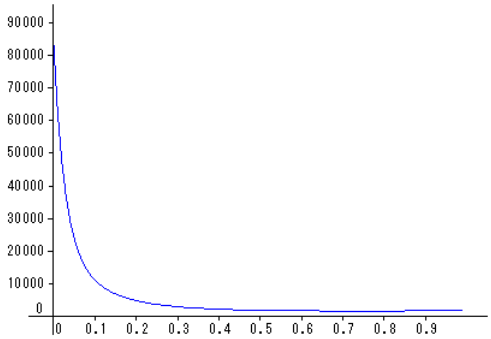


図 2.4 パラメータの推定

差の平均法と同様この場合も変動は平滑され、予測値は 1 期前の実測値に近い値になっている。また 2 期以上先の予測値は、実測データを予測データで置き換えると同じ値が続く。この予測値を見ると 1 期前の実測値にかなり引きずられていることが分かる。指数平滑法も上がり下がりのあるデータには向かない。

12.2.3 ブラウン法（ブラウンの 2 重指数平滑法）

指数平滑法は単純に今期までの指数平滑値を予測値としたものであって、予測値の精度については考慮されていない。この精度を考慮した方法がブラウン法（2 重指数平滑法）である。

ここで比較のために指数平滑法の公式を少し書き換えておく。

$$y_{t+1} = u_t$$

$$u_t = \alpha x_t + (1-\alpha)u_{t-1}$$ t 時点の x の指数平滑値（$t+1$ 時点の x の予想値）

ブラウン法は、指数平滑法で予測される $t+1$ 期の予測値 u_t に、この予測値と指数平滑法による u_t の予測値 v_{t-1} との差（の m' 倍）を足して来期を予測するものである。指数平滑を 2 度行うので 2 重指数平滑法と呼ばれる。

$$y_{t+1} = u_t + m'(u_t - v_{t-1})$$

$$u_t = \alpha x_t + (1-\alpha)u_{t-1}$$ t 時点の u の値（$t+1$ 時点の x の予想値）

$$v_{t-1} = \beta u_{t-1} + (1-\beta)v_{t-2}$$ $t-1$ 時点の v の値（t 時点の u の予想値）

ここに m, α, β はパラメータである。

この式を分かり易く表現すると以下となる。

x の補正予測値＝$t+1$ 時点の x の予測値＋m'（t 時点の u の値－t 時点の u の予測値）

＝$t+1$ 時点の x の予測値＋$t+1$ 時点の予測補正項

実際の計算では、参考文献 1 に従い、 $m'=1$,
$\alpha=\beta$ としており、

$$y_{t+1}=a_t+b_t$$
$$a_t=2u_t-v_t$$
$$b_t=\frac{\alpha}{1-\alpha}(u_t-v_t)$$

以下の初期値をおいている。

$$u_1=v_1=x_1$$
$$a_1=x_1,\ b_1=[(x_2-x_1)+(x_4-x_3)]/2$$

このため予測値は、 $t=5$ から求める。

ブラウン法による最適なパラメータでの予測を図 2.5 に示す。ここでも明らかなように増加・減少のあるデータに対してブラウン法はあまり有効とは言えない。

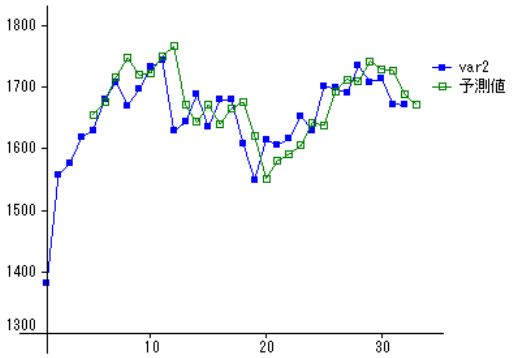


図 2.5 ブラウン法による予測（α=0.42）

12.2.4 最近隣法

最近隣法は現在とその 1 期前のデータに似た過去のデータを探して、次期のデータの予測値を決めるものである。

最近隣法は以下の形で予測を行う。現在とその 1 期前のデータを x_t, x_{t-1} とし、過去のデータ x_{t-m}, x_{t-m-1} との距離 d_m を以下のように考える。

$$d_m=\sqrt{(x_t-x_{t-m})^2+(x_{t-1}-x_{t-m-1})^2}$$

距離の最小値 $d_{\min}$ を求め、距離がその 1.62 倍未満のデータを集める。

$$S=\{d_m \mid d_m<1.62\times d_{\min}\}$$

この 1.62 は黄金分割比と呼ばれ、実用上多く使われる [1)]。その集めた距離の逆数を利用して重み w_m（$d_m\in S$）計算する。但し、距離が 0 の場合はある小さな値（このソフトの場合は 0.0001）として

いる。

$$w_m = \frac{1/d_m}{\sum_{d_k \in S} 1/d_k}$$

この重みを使って予測値 y_{t+1} を以下のように求める。

$$y_{t+1} = \sum_{d_m \in S} w_m x_{m+1}$$

実際に最近隣法を用いた予測は図 2.6 のようになる。

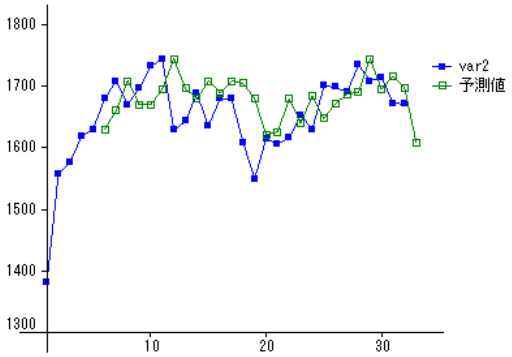


図 2.6 最近隣法による予測

この方法はデータの上がり下がりの変動が大きいほど有効で、上昇下降傾向があるデータには向かない。また過去の似た状況を探すことから、一般に過去のデータが多いほど予測の精度は上がる。

12.2.5 自己相関モデル（ARIMA）

このモデルには 3 つのパラメータ p, d, q があり、これらのパラメータを用いて、ARIMA(p, d, q) と表される。以後各パラメータについて説明し、最後に全体を見渡す。

最初にパラメータ d について述べる。これはデータの差分の回数である。差分は傾向変動などを取り除く 1 つの手段である。$x_t^{(1)}$ を 1 回の差分、$x_t^{(2)}$ を 2 回の差分とするとそれぞれ元のデータを用いて以下のように表される。

$$x_t^{(1)} = x_t - x_{t-1}$$
$$x_t^{(2)} = x_t^{(1)} - x_{t-1}^{(1)} = x_t - 2x_{t-1} + x_{t-2}$$

d 回の差分データに対して ARMA(p, q) モデルを適用する手法が ARIMA(p, d, q) モデルである。但し、d 回の差分データでは利用できるデータが、d+1 期から t 期までとなる。

MA モデル

次にパラメータ q について考える。このパラメータは MA(q) と呼ばれるモデルのパラメータであ

る。このモデルは$t \geq t_0$に対して以下の仮定が基礎になっている。

$$x_t = b_1 u_{t-1} + b_2 u_{t-2} + \cdots + b_q u_{t-q} + b_0 + u_t$$

ここに$u_t, u_{t-1}, \cdots, u_{t-q}$は各時点のホワイトノイズである。特に$b_0 = 0$の場合が教科書などに載っている。

1 期先の予測値y_{t+1}を実測値x_{t+1}からホワイトノイズu_{t+1}を引いたものと定義すると以下のような関係が得られる。

$$\begin{aligned} y_{t+1} &= x_{t+1} - u_{t+1} \\ &= b_1 u_t + b_2 u_{t-1} + \cdots + b_q u_{t-q+1} + b_0 - u_{t+1} + u_{t+1} \\ &= b_1 (x_t - y_t) + b_2 (x_{t-1} - y_{t-1}) + \cdots + b_q (x_{t-q+1} - y_{t-q+1}) + b_0 \end{aligned}$$

計算手順はまず$t < t_0$の間のノイズ$x_t - y_t$の初期値を決める。我々はこれを $N(0,1)$ の正規乱数としている。次にこれらの初期値を用いて$t = t_0$の場合に上式から重回帰分析を用いて予測値y_{t_0+1}求める。但し、計算が可能なためには初項の時期をずらしたデータの組がq個必要であり、少なくとも$t_0 > 2q$でなければならない。我々はこれを$t_0 = 2q + 2$にしている。ここで得た予測値y_{t_0+1}を使って、上式を用いて再度重回帰分析を行うことによって新しい予測値y_{t_0+2}を得る。これを繰り返して行くことで、最終的な予測値y_{t+1}を得る。

この処理では長期予測は不可能である。長期予測のためには実測値の代わりに予測値を用いるしかないが、そうすると説明変数が 0 になって行き、前の予測値が続くようになる。

MA(1) と MA(2) による予測グラフを図 2.7a と図 2.7b に示す。

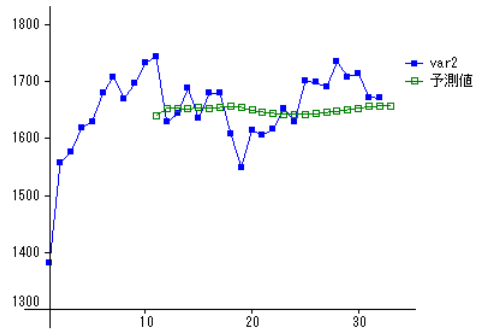


図 2.7a　MA(1) モデルによる予測

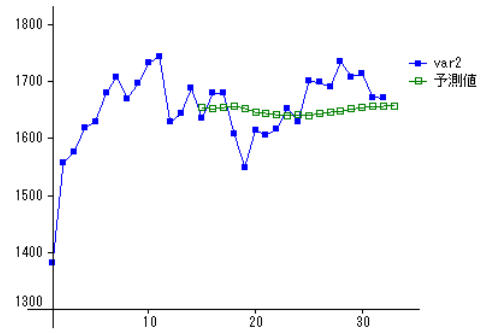


図 2.7b　MA(2) モデルによる予測

AR モデル

パラメータ p は AR(p) と呼ばれるモデルのパラメータである。このモデルは以下の仮定が基礎になっている。

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} + a_0 + u_t$$

ここにu_tはt時点のホワイトノイズである。特に$a_0 = 0$の場合が教科書などによく載っている。

1 期先の予測値 y_{t+1} を実測値 x_{t+1} からホワイトノイズ u_{t+1} を引いたものと定義すると $t \geq t_0$ に対して以下のような関係が得られる。

$$y_{t+1} = a_1 x_t + a_2 x_{t-1} + \cdots + a_p x_{t-p+1} + a_0$$

計算は重回帰分析を用いるが、手順は過去の予測値を使う必要がないので MA モデルと比べると単純である。但し、計算が可能なためには初項の時期をずらしたデータの組が p 個必要であり、少なくとも $t_0 > 2p$ でなければならない。我々はこれを $t_0 = 2p + 2$ にしている。

この処理でも長期予測は不可能である。長期予測のためには実測値の代わりに予測値を用いるしかないが、a_i が殆ど変わらない状況では例えば $p = 1$，$|a_1| < 1$ の場合、

$$y_n = a_1 y_{n-1} + a_0, \quad \lim_{n\to\infty} y_n = a_0/(1-a_1)$$

となり、前の予測値に近い値が続くようになる。AR(1)と AR(2) による予測グラフをそれぞれ図 2.8a と図 2.8b に示す。

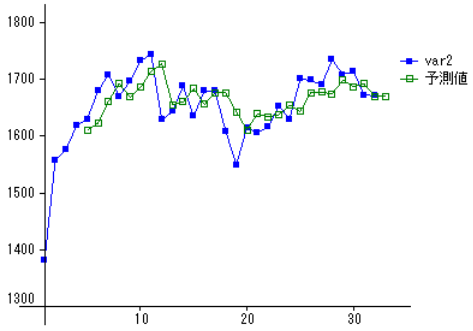


図 2.8a AR(1) モデルによる予測

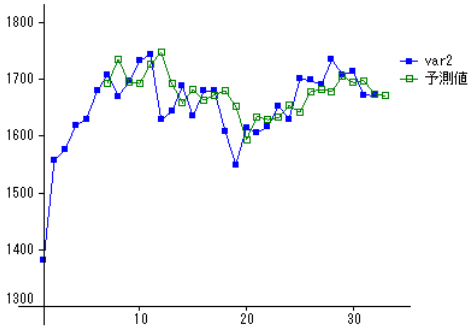


図 2.8b AR(2) モデルによる予測

ARIMA モデル

ここではこれまで学んできたモデルを複合した場合を考える。今 d 回の差分データを $x_t^{(d)}$ とすると、ARIMA(p, d, q) モデルは $t \geq t_0$ で以下のように表される。

$$x_t^{(d)} = \sum_{i=1}^{p} a_i x_{t-i}^{(d)} + \sum_{i=1}^{q} b_i u_{t-i}^{(d)} + c + u_t$$

これを用いて予測値 $y_{t+1}^{(d)}$ は以下のようになる。

$$y_{t+1}^{(d)} = \sum_{i=1}^{p} a_i x_{t-i+1}^{(d)} + \sum_{i=1}^{q} b_i (x_{t-i+1}^{(d)} - y_{t-i+1}^{(d)}) + c$$

計算手順は、まず $t < t_0$ 以前のノイズ $x_t^{(d)} - y_t^{(d)}$ を標準正規乱数で初期化する。後は MA モデルの場合と同様に、$t = t_0$ の場合の予測値 $y_{t_0+1}^{(d)}$ を重回帰分析で求めて、これを利用してさらに次の予測値を求める方法をとる。但し計算が可能なためには、上式に必要なデータが $r = \max(p, q)$ 個、それを時期をずらして $p + q$ 期分必要であることから、少なくとも $t_0 > r + p + q + d$ でなければならない。

我々は少し大きくとって、以下としている。

$$t_0 = r + p + q + d + 2$$

計算が可能であることで上のような条件を付けたが、計算の正確さを考えると十分でない。MA モデルでは計算の初期値を乱数で与えているので、t_0 の近くの推定値は良い近似ではない。我々は値が安定するまで待つ必要がある。そのため、誤差の計算や表示に利用するのは実際には経験的に以下にしている。

$$t_0 = 2p + d + 2 \qquad q = 0 \text{ の場合}$$

$$t_0 = (r + p + q + d + 2) + (2q + 5) \qquad q > 0 \text{ の場合}$$

これで $t_0 + 1$ 期からの予測値 $y_{t+1}^{(d)}$ が求められたが、これは差分を d 回取ったデータの予測値である。我々はこれを元のデータに戻す必要がある。データ間に

$$x_{t+1}^{(d-1)} = x_t^{(d-1)} + x_{t+1}^{(d)}$$

の関係があることから、これを以下のように拡張する。

$$y_{t+1}^{(d-1)} = x_t^{(d-1)} + y_{t+1}^{(d)}$$

即ち、以下のように求められる。

$$y_{t+1} = y_{t+1}^{(0)} = x_t^{(0)} + y_{t+1}^{(1)} = x_t^{(0)} + x_t^{(1)} + y_{t+1}^{(1)} = \cdots = \sum_{i=0}^{d} x_t^{(i)} + y_{t+1}^{(d)}$$

ARIMA(1,0,1) , ARIMA(1,1,1) による予測グラフを図 2.9a と図 2.9b に示す。

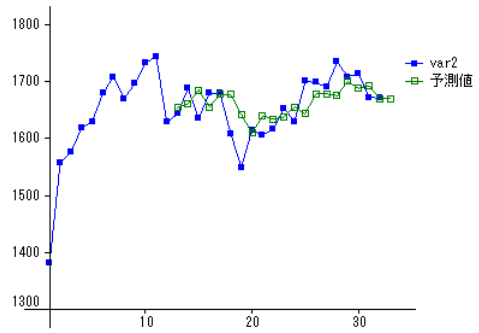


図 2.9a ARIMA(1,0,1) モデルによる予測

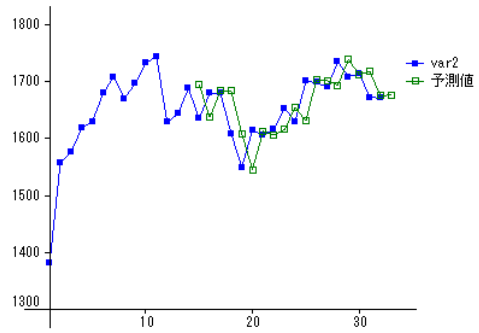


図 2.9b ARIMA(1,1,1) モデルによる予測

差分を入れると 1 期前の実測値に差分の予測値を足すことになり、やはり 1 期前の状態に引きずられるようである。

12.3 変動の分解モデル

具体的なイメージを持ってもらうために、今後しばらく図 3.1 のデータを元にして話を進める。

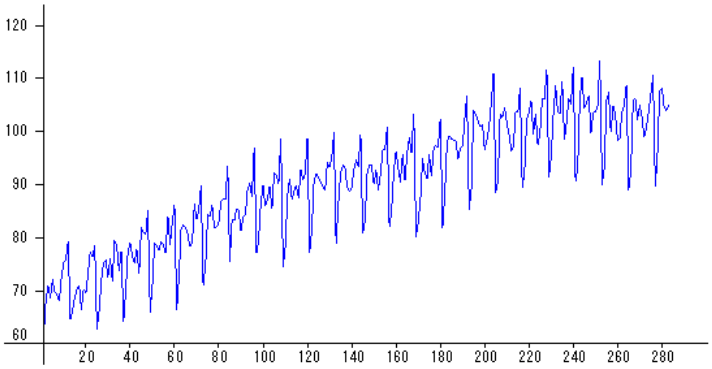


図 3.1 時系列データ decomp_food

データは様々な要因で変動するが、我々は大きくこれを、傾向変動 T 、季節変動 S 、循環変動 C 、残差変動 R に分ける。ここに傾向変動は長期にわたる継続的な変化で、季節変動は周期が一定の変化、循環変動は周期が一定でないものの周期性が認められる変化、残差変動は観測誤差などのゆらぎである。一般に変数 X はこれらの変動の関数として以下のように表される。

$$X = f(T, S, C, R)$$

この一般の関係の中で、実際の分析のためには様々な仮定を置くことが多い。我々のプログラムでは周期が変化する循環変動については考えず、それぞれの変動の合計で表される以下の加法モデルを採用している

$$X = T + S + R$$

但し、傾向変動には通常、移動平均や回帰近似が利用されるが（これを近似モデルと呼ぶ）、我々は傾向変動を２つに分け、近似モデル T_1 と 2.1 節で述べた予測モデル T_2 の和と考える。これによって予測モデルだけの処理も変動の分解モデルと合わせた処理も可能になる。また季節変動について、振幅の変化も考え、季節変動を振幅変動 A と振幅一定の季節変動 S'（以後これを周期変動と呼ぶ）の積に分解する。ここで振幅変動には回帰近似を用い、周期変動の意味を分りやすくするため、大きさの平均を１に近くなるようにとる。これらを合わせて、我々のプログラムでは以下のようなモデルを扱う。

$$X = T_1 + T_2 + A \times S' + R$$

以後予測モデル T_2 を除いて、それぞれの変動の分解について詳細に説明する。

12.3.1 傾向変動の分解

傾向変動の抽出は主に移動平均法による方法と最小２乗法の手法を応用した方法（回帰分析はこれに含まれる）がある。n 期の移動平均法では時点 t のデータの値を以下のようにして、データの平滑

化を図る。

$$d_t = \frac{1}{2m+1}\sum_{i=-m}^{m} x_{t+i} \qquad n = 2m+1 \text{の場合}$$

$$d_t = \frac{1}{2m+2}\left\{\sum_{i=-m}^{m} x_{t+i} + \frac{1}{2}(x_{t-m-1} + x_{t+m+1})\right\} \qquad n = 2m+2 \text{の場合}$$

これは中心法と呼ばれる方法であるが、移動平均を予測に用いる場合には、以下のような方法が使われる。我々はこの方法を用いる。

$$d_t = \frac{1}{n}\sum_{i=-n}^{-1} x_{t+i}$$

また、時間のずれに対して重み係数を掛ける場合もある。データに周期性がある場合、この方法では傾向変動に周期成分が残るが、移動平均を行ったデータに再度移動平均を行うとさらになめらかな傾向が得られる。但し、移動平均では時系列データの前後、または前が使えなくなるので、ある程度データ数も必要である。我々のプログラムでは複数回の移動平均は考えていない。

予め大雑把なデータの変化を近似的につかんでおくことは重要である。最小２乗法の手法を応用した近似手法の中で線形回帰分析を利用するものは計算が容易である。よく使われる線形回帰の方法には以下のようなものがある。

１次近似	$d_t = at + b$	$d_t = at + b$
対数近似	$d_t = a\log t + b$	$d_t = a\log t + b$
べき乗近似	$d_t = bt^a$	$d_t = bt^a$
指数近似	$d_t = be^{at}$	$d_t = be^{at}$
多項式近似	$d_t = a_p t^p + a_{p-1}t^{p-1} + \cdots + a_1 t + a_0$	

ここにべき乗近似と指数近似については両辺の対数をとって線形回帰分析を行う。また、多項式近似は重回帰分析を用いてパラメータの推定を行う。例として２次式による近似結果を図 3.2 に示す。

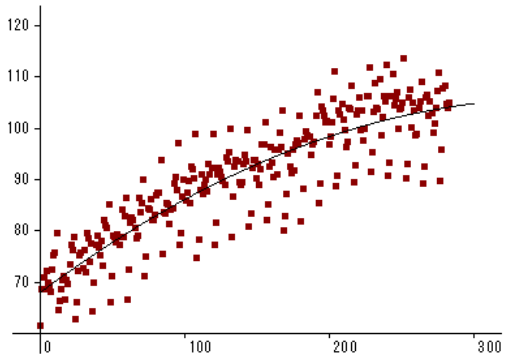


図 3.2　２次曲線の当てはめ

このデータについては以下の２次曲線が最良である。

$$y = -0.00029t^2 + 0.211t + 68.002$$

これら以外の近似には非線形最小２乗法など他の方法を利用する。

この傾向変動の結果を元データから分離するには、我々のモデルでは引き算を用いる。

$$y_t = x_t - d_t$$

この２次曲線を傾向変動として取り除くと図 12.3.3 の結果となる。この段階での実測値と予測値の相関係数の 2 乗（決定係数）R^2 は 0.7904 である。

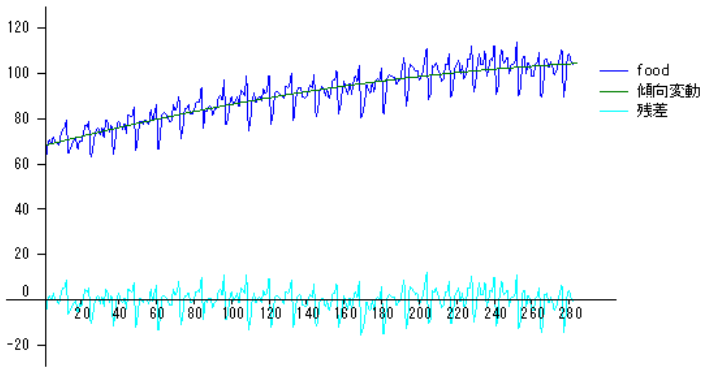


図 12.3.3 傾向変動の分離

もう１つ傾向変動の分解に利用できる方法として、局所回帰分析が考えられる。これは、ウェイトをかけた回帰分析である。予測したい点を要求点として、その近傍に大きなウェイトをかけ、それから離れるに従ってウェイトを小さくする。これにより、関数形を定めることなく、非線形の予測を行うことができる。ウェイトの範囲はバンド幅と呼ばれる値によって決めることができるが、バンド幅が 100 以上の場合はほぼ完全に線形回帰分析となる。通常利用されるのは、バンド幅が 0 から 1 の範囲が多い。

予測モデルの分解については、12.2 節で述べたので省略する。

12.3.2 振幅変動の分解

振幅変動の推定は以下の振幅変動データに対して近似曲線を考えることによって与えることにする。

振幅変動データ＝傾向変動の残差の絶対値

÷傾向変動の残差の絶対値の平均値

これによって振幅変動の値はほぼ 1 に近い値となり、周期変動を平均的な振幅を持つ季節変動と意味付けることができるようになる。図 3.4 に近似直線を求める図を示す。振幅変動を分離した残差は傾

向変動残差÷振幅変動推定値で与えられる。

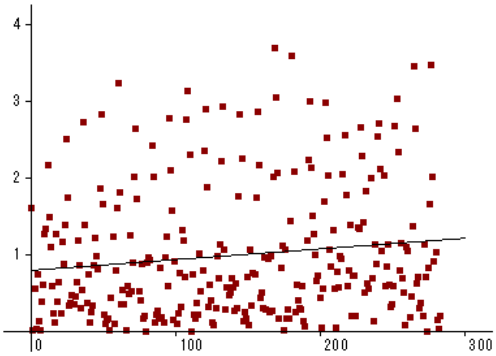


図 3.4 振幅変動の推定

12.3.3 周期変動の分解

周期変動のスペクトル抽出は傾向変動と振幅変動を除去したデータ y_t にどのような周波数成分が含まれるかを知る重要な処理である。最初に時間的なラグの影響を見るために自己相関係数を求め、ラグの値によってそれをプロットするコレログラムを作成する。

自己相関係数 r_k （$k=1,2,\cdots,L<N-1$）は以下の式により求められる。

$$r_k=\frac{s_k^2}{s_0^2},\qquad \text{ここに}\quad s_k^2=\frac{1}{N-k}\sum_{t=k+1}^{N}(x_t-\bar{x}_{k+1}^{N})(x_{t-k}-\bar{x}_1^{N-k}),\quad \bar{x}_a^b=\sum_{t=a}^{b}x_t$$

図 3.5 に最大周期を 70 にしたコレログラムを示す。これによると変動の周期は 12 であることが分かる。

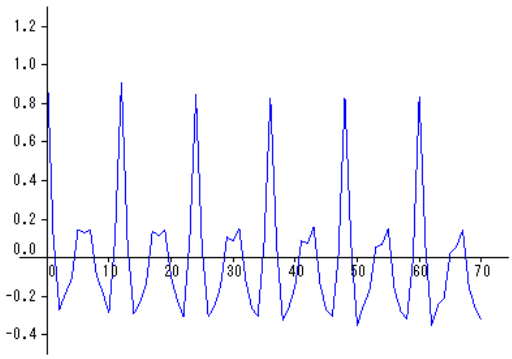


図 3.5 コレログラム

次にこのコレログラムに対してその周波数成分を見ると周期性がより明確になる。このような問題には関数のフーリエ（Fourier）展開という手法が用いられるが、ここでは参考のために期間 $2L$ を周

期に持つ関数 $f(x)$ のフーリエ展開の公式を与えておく。

$$f(x)=\frac{a_0}{2L}+\frac{1}{L}\sum_{k=1}^{\infty}\left(a_k\cos k\pi x/L+b_k\sin k\pi x/L\right)$$

$$a_k=\int_{-L}^{L}f(x)\cos k\pi x/L\,dx,\qquad b_k=\int_{-L}^{L}f(x)\sin k\pi x/L\,dx$$

この式は関数を周波数 $f_k=k/2L$ （$k=1,2,3,\cdots$）の正弦波成分の合計で表したもので、各成分の強さは係数 a_k と b_k で与えられる。

我々の時系列データでは関数が離散的であるため、離散フーリエ変換という手法を利用する。n を時系列データ x_t の周期として、離散フーリエ展開の公式を以下に与える。

$$x_t=\frac{1}{n}\sum_{k=0}^{n-1}(a_k\cos 2\pi kt/n+b_k\sin 2\pi kt/n) \tag{1}$$

$$a_k=\sum_{t=1}^{n}x_t\cos 2\pi kt/n,\qquad b_k=\sum_{t=1}^{n}x_t\sin 2\pi kt/n$$

この公式を自己相関係数 r_i に対して適用する。自己相関係数は $r_i=r_{-i}$ であるため、$-m\le i<m$ の範囲で偶関数である。その際には周期を $2m$ として、以下の形で与えられる。

$$r_i=\frac{1}{2m}\sum_{k=-m}^{m-1}(a_k\cos 2\pi k/2m+b_k\sin 2\pi kt/2m)=\frac{1}{m}\sum_{k=0}^{m-1}a_k\cos\pi kt/m$$

$$a_k=\sum_{i=-m}^{m-1}r_i\cos 2\pi ki/2m=2\sum_{i=0}^{m-1}r_i\cos\pi ki/m$$

この量 a_k を周波数 $f_k=k/2m$ の生スペクトルと呼び、これをラグごとに表したグラフをピリオドグラムという。実用上は生スペクトルより、平滑化という処理を行ったピリオドグラムがよく用いられる [2)]。

実際のデータに対する平滑化したピリオドグラムを図 3.6 に示す。

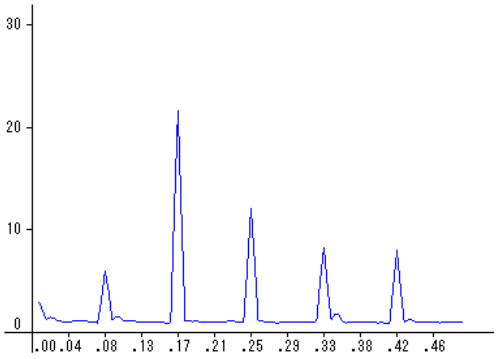


図 3.6 ピリオドグラム

これを詳細に見るとまず、周波数 0.167（周期 6：これらは別に表示されるデータから読み取れる）に大きなピークがあり、同様に周波数 0.25（周期 4）、周波数 0.33（周期 3）、周波数 0.08（周期 12）

などにもピークがある。これらの全体的な周期は、ここに現れた周期の重ね合わせ（最小公倍数、但し時系列の長さの半分より小さいこと）と考えると周期 12 である。

この変動の分離には一般の離散フーリエ変換の式（1）を利用するが、上で考えた周期を n として残差 y_t に適用し、周期変動 u_t を得る。

$$u_t = \frac{1}{n}\sum_{k=0}^{n-1}(a_k \cos 2\pi kt/n + b_k \sin 2\pi kt/n)$$

$$a_k = \sum_{t=1}^{n} y_t \cos 2\pi kt/n, \qquad b_k = \sum_{t=1}^{n} y_t \sin 2\pi kt/n$$

時系列のデータには周期性があると言っても、各周期間には揺らぎが見られる。しかし上の計算では時系列中どの 1 周期を考えればよいのか分からない。そこで実際の計算には特定の 1 周期を選ぶのではなく、各周期中の同一時点の残差の平均 $\bar{y}_t$ を用いて計算を行った。

このようにして季節変動を除去した結果が図 3.7 である。ここでは除去した季節変動と残差のみ示してある。この段階での実測値と予測値の R^2 は 0.9647 である。

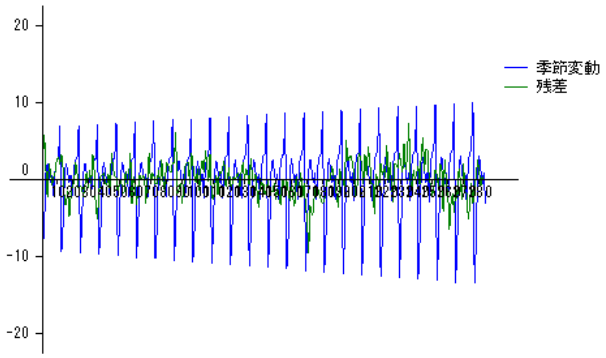


図 3.7 季節変動の分解

もう少し詳細に残差の周波数をながめて（タイムラグ 200 まで）図 3.8 でピリオドグラムを描いてみる。

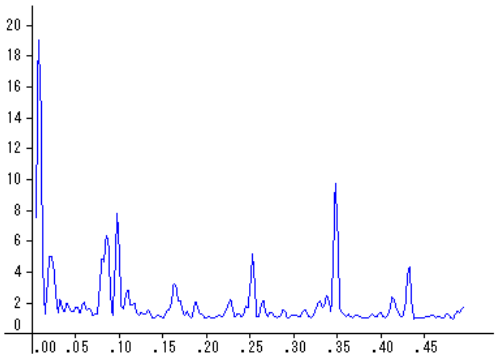


図 3.8 残差のピリオドグラム

これを見ると、0 の近くにピークがあり、これは周期 130 近傍のピークであることが分かる。残差の標準偏差を最小にするように選んでやると、周期は 129 となる。そこでこの周期変動を差し引いて、最終的に図 3.9 の分解になる。最終的な実測値と予測値の R^2 は 0.9838 となる。振幅変動を分離しない場合の R^2 は 0.9830 であり、この場合振幅変動の分解の効果はわずかである。

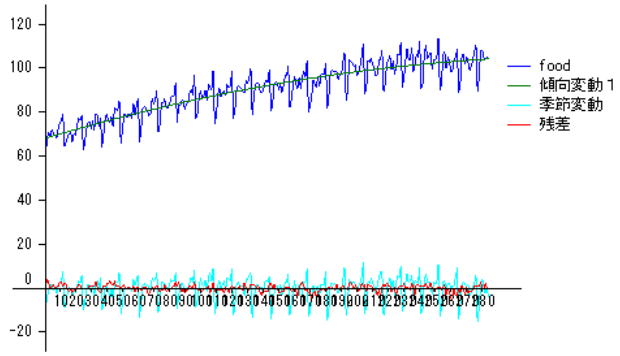


図 3.9 時系列データの分解

実はこの残差にはまだ周期性が残っており、これに対して周期性の分離を行い、さらに残差を小さくできる。実際、例えば 91,90,41 と周期性を取り除いていくと実測値と予測値の R^2 は 0.9941 と大きくできる。これを見ると予測精度が上がっているように思われるが、すでに周期成分 129 を入れているのでこのデータの数 283 個から見れば、わずか 2 周期分を用いて予測を行っていることになる。3 周期目はそれ以前と少しずれることを考えると、いくら残差が小さくできたからといって予測が正しくなる保証はない。ある程度のところで止めておくべきであろう。

さて分解がうまくいき、これ以上分解が難しくなる場合もある。そのとき残差の自己相関係数は 0 に近い値となり、ピリオドグラムは平坦に近くなる。このような波をホワイトノイズと呼ぶ。ホワイトノイズの検定には、Ljung-Box 検定が用いられる。それには、利用するデータ数を t、ラグ i の母相関係数と標本相関係数をそれぞれ ρ_i, r_i として、以下の関係が利用される。

帰無仮説： $\rho_1 = \rho_2 = \cdots = \rho_m = 0$

$$Q = t(t+2)\left\{\frac{r_1^2}{t-1} + \frac{r_2^2}{t-2} + \cdots + \frac{r_m^2}{t-m}\right\} \sim \chi_m^2$$

12.3.4 変動の分解モデルによる予測

時系列データの変動の分解は、データにある程度の周期性があること、その数が最低でも 2 周期分以上あることが条件で可能となる。また傾向変動２（予測手法）を使うと長期予測は難しい。これまで見てきたデータについて 100 期先までの長期予測をしてみよう。見易くするために $t = 200$ からのデータを図 3.10 に表示する。

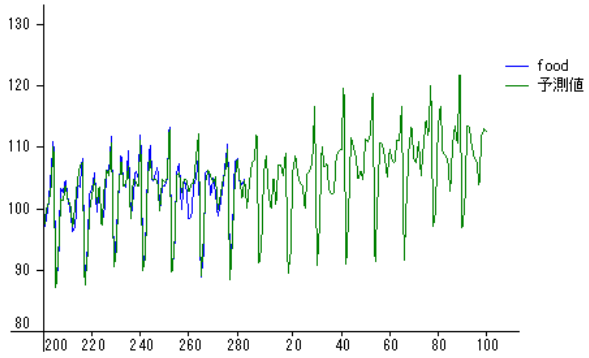


図 3.10 時系列データと長期予測

12.4 プログラムの動作

ここでは具体的に実行画面を見ながらプログラムの動作について説明する。時系列分析のメニュー画面を図 12.4.1 に示す。それぞれのボタンの出力結果については２章の図で示しているので、ここではメニューの使い方に焦点を絞って説明する。

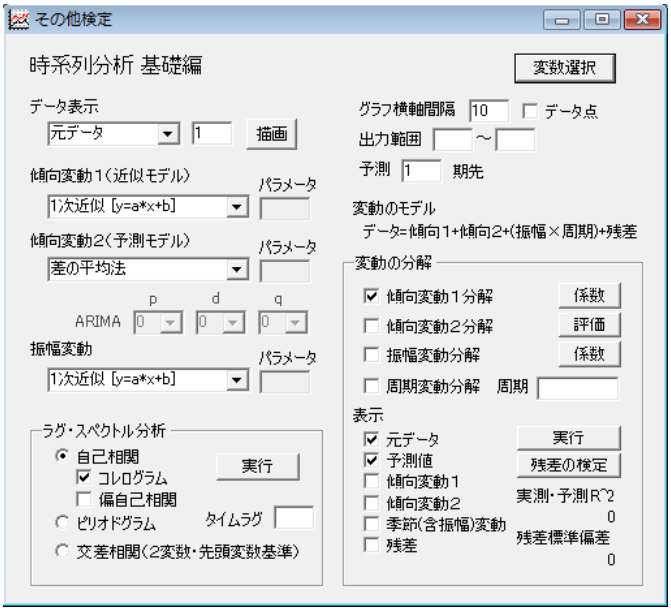


図 12.4.1 時系列分析メニュー

最初に変数選択ボタンで分析対象の変数を選択するが、単独で選択しても、時間を指定する変数と同時に選択してもよい。変数を２つ選択する場合、分析対象の変数を目的変数として先に選択する。入力されたデータを見るためには「データ表示」コンボボックスの形式を選んだ後、「描画」ボタン

をクリックする。データの表示形式には元データ、対数変換、差分、差分期間比がある。このプログラムでは自動的にこれらでデータを変換して分析を実行することはできないが、結果をデータに貼り付けて実行することは可能である。グラフの横軸目盛間隔は右上の「グラフ横軸間隔」テキストボックスで、時間の出力範囲は「出力範囲」テキストボックスで指定できる。グラフのデータポイントの有無はデータ点チェックボックスで選択できる。

変動の分解モデルでの実際の分解は、変動の分解グループボックス内で必要な項目をチェックし、「実行」ボタンをクリックすることで実行できる。特に周期変動の分解では、周期テキストボックスに分解する周期を入力する。周期はカンマ区切りで複数入力できる。「残差の検定」ボタンをクリックすると、変動の分解残差について Ljung-Box 検定が実行される。「係数」や「評価」のコマンドボタンはそれぞれの分解で最適なパラメータを確認するために用いられる。

メニュー左側に並んだコンボボックスでは、傾向変動や振幅変動の分解のメニューが示される。傾向変動１（近似モデル）のコンボボックスには、移動平均、１次近似、対数近似、べき乗近似、指数近似、多項式近似、非線形近似へ、の項目が含まれている。移動平均の期間や多項式近似の次数は、横のテキストボックスで指定する。非線形最小２乗法へを選択すると、すでに設定済みかどうかのメッセージの後、未設定の場合は非線形最小２乗法の分析メニューが表示される。ここで得た結果は傾向変動１の値となる。傾向変動２（予測モデル）のコンボボックスには、差の平均法、指数平滑法、ブラウン法、最近隣法、ARIMA の項目が含まれている。これらの分析のパラメータは横や下にあるテキストボックスで指定する。振幅変動のコンボボックスには１次近似、対数近似、べき乗近似、指数近似、多項式近似の項目が含まれている。多項式近似の次数は横のテキストボックスで指定する。

周期変動の周期は左下のラグ・スペクトルグループボックスで調べる。必要なラジオボタンやチェックボックスを選び、実行ボタンでそれぞれのグラフが表示される。タイムラグテキストボックスでは詳細な検討のためのコレログラムのタイムラグやピリオドグラムの周期の最大を与える。特に指定がなければ、データで利用できる最大値が使われる。

１３．共分散構造分析

共分散構造分析はこれまでの多変量解析の手法を包含する優れた分析手法であり、第２世代の多変量解析と呼ばれることもある。利用者は観測される変数や内部に潜在する直接観測されない変数間の関係を記述するネットワーク型の統計モデルを作成し、そのモデルと観測値とで各変数間の直接的な影響力を推測する。統計モデルはこれまでの多変量解析に比べて複雑な構造を記述可能で、その中に重回帰分析や因子分析などの構造を複数含めることができる。

我々は、社会システム分析教育用ソフトウェア College Analysis の機能拡張のため、新たに共分散構造分析のプログラムを追加することにした。College Analysis には、集計や検定を扱う基本統計や多変量解析のプログラムが含まれているが、共分散構造分析の重要性を考えるとこの分析手法は避けて通ることができないものと思われる。しかしこの分析のプログラムは分量が多く、グラフィックでの構造図入力や複雑なアルゴリズムなど取り組むべき課題も多い。

共分散構造分析は変数間の関係を構造方程式と呼ばれる線形の式で与え、変数間の影響の強さを表すパラメータの値は観測変数の共分散行列から推定する。その際一般にパラメータ数は共分散行列の独立な成分数と異なるため、パラメータの値は厳密には決まらない。パラメータの推定にはある評価関数を用いて、これを最小化するような方法を考える。この評価関数の選び方によって、推定値の導出にはいくつかの方法がある。その中で最もよく利用されるのが最小２乗法や最尤法である。

我々のプログラムの最大の問題はこの最小化のアルゴリズムにある。最小２乗法では評価関数はパラメータについて高次の多項式となり、最尤法では非線形の長大な数式となる。これらの数式の最小化問題は非常に繊細で、これまでの Newton-Raphson 法では限界があるし、計算の手順によっては時間が膨大にかかる場合もある。今回のプログラムではこのアルゴリズムに Levenberg-Marquart 法を応用したものを採用し、計算の方法もできる限り時間的な無駄を省くように考え、簡単なモデルであれば何とか辛抱できる時間で計算できるところまで来た。しかし、Amos などのプログラムでは 1985 年以降発展してきたマルコフ連鎖モンテカルロ法などが採用されており、短時間で比較的安定な解を求めることができるようになってきている。我々も今後このようなアルゴリズムを使ったプログラムに変更して行く必要があるが、現段階では２つのアルゴリズムの違いを実感しておくのも今後のための教訓となる。

この論文では非常に簡単なモデルから、多少複雑な（まだ実用モデルの段階ではないが）モデルまで我々のプログラムと Amos の結果とを比較してみた。その中で我々のプログラムだけでなく、Amos の利用上の注意点も少しだけ見えてきた。これらの問題についても例を見ながら考えて行く。

13.1 モデルの構造と方程式

ここでは図 1.1 の構造モデルを例として共分散構造分析の理論の説明をする。

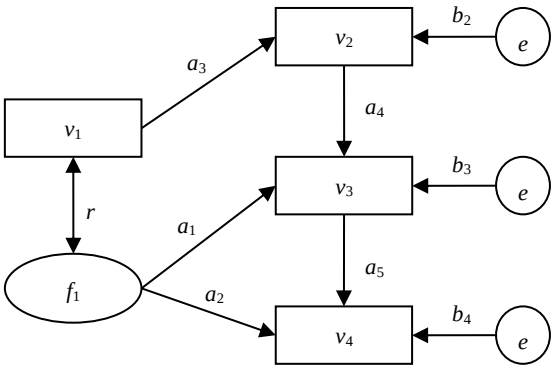


図 1.1 構造モデル

四角や楕円や円で表される量はモデルに含まれる変数で、形によりその意味するところが異なり、それぞれラベルが付けられている。矢印は因果関係を表すパラメータで、これにもラベルが付けられている。また、双方向の矢印は相関を表すパラメータである。

このモデルをよく利用される影響行列の形で表現すると表 1.1 のようになる。左側の変数が始点、上側の変数が終点である。

表 1.1 構造モデルの影響行列

	f_1	v_1	v_2	v_3	v_4	e_2	e_3	e_4
f_1		r		a_1	a_2			
v_1	r		a_3					
v_2				a_4				
v_3					a_5			
v_4								
e_2			b_2					
e_3				b_3				
e_4					b_4			

変数は通常、いくつかの視点から以下のように分けられる。

観測変数と潜在変数

観測変数とは実測値の分かっている変数であり、図 1.1 の構造モデルでは v_1, v_2, v_3, v_4 などの変数がこれに相当し、構造図では四角形で表現される。潜在変数とは直接には観測されない変数で、因子分

析の因子や誤差などがこれに当り、構造図では楕円や円で表現される。図 1.1 の例では f_1, e_2, e_3, e_4 などの変数である。ここでは f_1 が因子変数、e_2, e_3, e_4 が誤差変数である。特に因子変数は楕円、誤差変数は円（または円なし）で表現される場合がある。

外生変数と内生変数

外生変数は構造モデルで相関を除いてどこからも影響を受けない（片側矢印が入らない）変数で、図 1.1 の構造モデルでは v_1, f_1, e_2, e_3, e_4 がこれに当る。内生変数はそれ以外の変数で v_2, v_3, v_4 などである。

構造変数と誤差変数

構造変数とは後に述べるモデルの構成要素に使われる変数で、図 1.1 の構造モデルでは f_1, v_1, v_2, v_3, v_4 などがこれに当る。誤差変数とはモデルでは説明できないゆらぎの成分を表すもので e_2, e_3, e_4 がこれに当る。

これらの変数の関係は構造方程式と呼ばれる式で表現される。図 1.1 の構造モデルでは以下となる。

$$v_2 = a_3 v_1 + b_2 e_2$$
$$v_3 = a_1 f_1 + a_4 v_2 + b_3 e_3$$
$$v_4 = a_2 f_1 + a_5 v_3 + b_4 e_4$$

この方程式の左辺を構造変数に拡張し、以下のような式を考える

$$f_1 = f_1$$
$$v_1 = v_1$$
$$v_2 = a_3 v_1 + b_2 e_2$$
$$v_3 = a_1 f_1 + a_4 v_2 + b_3 e_3$$
$$v_4 = a_2 f_1 + a_5 v_3 + b_4 e_4$$

構造方程式の左辺には構造変数と呼ばれる変数を取るが、そのうちの内生変数は必ず誤差変数からの影響を受けるようにする。上の構造方程式を行列表示すると以下のような形になる。

$$\begin{pmatrix} f_1 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & a_3 & 0 & 0 & 0 \\ a_1 & 0 & a_4 & 0 & 0 \\ a_2 & 0 & 0 & a_5 & 0 \end{pmatrix} \begin{pmatrix} f_1 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & b_2 & 0 & 0 \\ 0 & 0 & 0 & b_3 & 0 \\ 0 & 0 & 0 & 0 & b_4 \end{pmatrix} \begin{pmatrix} f_1 \\ v_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

今、以下のように定義すると、

$$\mathbf{A}=\begin{pmatrix}0&0&0&0&0\\0&0&0&0&0\\0&a_3&0&0&0\\a_1&0&a_4&0&0\\a_2&0&0&a_5&0\end{pmatrix},\quad \mathbf{B}=\begin{pmatrix}1&0&0&0&0\\0&1&0&0&0\\0&0&b_2&0&0\\0&0&0&b_3&0\\0&0&0&0&b_4\end{pmatrix},\quad \mathbf{t}=\begin{pmatrix}f_1\\v_1\\v_2\\v_3\\v_4\end{pmatrix},\quad \mathbf{h}=\begin{pmatrix}f_1\\v_1\\e_2\\e_3\\e_4\end{pmatrix}$$

構造方程式は(1)式のように表すことができる。

$$\mathbf{t}=\mathbf{At}+\mathbf{Bh} \tag{1}$$

ここに $\mathbf{t}$ は構造変数からなるベクトル、$\mathbf{h}$ は外生変数からなるベクトルである。またパラメータは行列 $\mathbf{A}$ と $\mathbf{B}$ に含まれる。

構造方程式は以下のように変形できる。

$$\mathbf{t}=(\mathbf{I}-\mathbf{A})^{-1}\mathbf{Bh} \tag{2}$$

ここでベクトル $\mathbf{t}$ のうち観測変数に注目し、観測変数で作られたベクトル $\mathbf{v}$ とそれを取り出す行列 $\mathbf{G}$ を以下のように定義する。

$$\mathbf{v}=\mathbf{Gt},\ \text{ここに}\quad \mathbf{v}=\begin{pmatrix}v_1\\v_2\\v_3\\v_4\end{pmatrix},\quad \mathbf{G}=\begin{pmatrix}0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{pmatrix}.$$

この関係を用いると、上式は(3)式のように変形される。

$$\mathbf{v}=\mathbf{G}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{Bh} \tag{3}$$

次に観測変数 $\mathbf{v}$ および外生変数 $\mathbf{h}$ の共分散行列を考える。簡単のため潜在変数は平均が 0、分散が 1 になるように標準化されているものとする。変数 $\mathbf{v}$ の共分散行列を $E(\mathbf{vv}')$、変数 $\mathbf{h}$ の共分散行列を $E(\mathbf{hh}')$ とするとそれらの関係は(4)式のようになる。

$$E(\mathbf{vv}')=\mathbf{G}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}E(\mathbf{hh}')\mathbf{B}'(\mathbf{I}-\mathbf{A})'^{-1}\mathbf{G}' \tag{4}$$

実際の計算では $E(\mathbf{vv}')$ を標本から得られた不偏共分散行列（共分散行列の不偏推定量）で置き換え、$E(\mathbf{hh}')$ についても観測変数部分は不偏共分散行列、潜在変数部分は分散を 1、共分散には必要に応じて共分散を表すパラメータを設定する。図 1 の構造モデルの場合は、潜在変数間または外生の観測変数と潜在変数間で、f_1 と v_1 の間だけに共分散 r を仮定しているので、以下の形となる。

$$E(\mathbf{vv}')\Rightarrow\mathbf{U}=\begin{pmatrix}u_{11}&u_{12}&u_{13}&u_{14}\\u_{21}&u_{22}&u_{23}&u_{24}\\u_{31}&u_{23}&u_{33}&u_{34}\\u_{41}&u_{24}&u_{34}&u_{44}\end{pmatrix},\quad E(\mathbf{hh}')\Rightarrow\mathbf{H}=\begin{pmatrix}1&r&0&0&0\\r&u_{11}&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{pmatrix}$$

ここで$\mathbf{U}$は不偏共分散行列であるが、標準化したデータの場合には相関行列となる。これを用いて(4)式を書き換えると以下のようになる。

$$\mathbf{G}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\mathbf{H}\mathbf{B}'(\mathbf{I}-\mathbf{A})'^{-1}\mathbf{G}'=\mathbf{U} \qquad (5)$$

これは観測値とパラメータを結びつける方程式である。この方程式を丁度方程式と呼び、一意的な解が存在する場合、その解を丁度解と呼ぶ。しかし丁度解が存在する場合はまれで、一般には解が不定になっていたり、不能になっていたりする。解が不定になっている場合をパラメータは識別不能という。不能になっている場合は最適近似解を求める。最適近似解を求める方法はいくつかあるが、ここでは主に利用される２つの方法について紹介する。

13.2 パラメータの推定

パラメータの推定は方程式の近似解を求めるための評価関数を作り、それを最小化する方法が採られるが、この節ではよく利用される２つの評価関数について説明する。

最小２乗法

方程式(5)の左辺と右辺の差の２乗和を最小化するために以下の評価関数を考える。

$$f_{MS}(\boldsymbol{\theta})=\sum_{i=1}^{n}\sum_{j=i}^{n}\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})_{ij}-u_{ij}\right)^2$$

ここに$\boldsymbol{\theta}$はパラメータを総称したものであり、nは観測変数の数、$\boldsymbol{\Sigma}(\boldsymbol{\theta})$は以下のように(5)式の左辺を表す。

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})=\mathbf{G}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\mathbf{H}\mathbf{B}'(\mathbf{I}-\mathbf{A})'^{-1}\mathbf{G}'$$

丁度解の場合$f(\boldsymbol{\theta})$の値は０である。

最尤法

我々はまず観測値を与える確率変数$\mathbf{x}_\lambda$（$\lambda=1,2,\cdots,N$）がそれぞれ独立にn変量正規分布に従うと考える。共分散行列を$\boldsymbol{\Sigma}(\boldsymbol{\theta})$とすると、$\mathbf{x}_\lambda$の確率密度関数は以下で与えられる。

$$f(\mathbf{x}_\lambda,\boldsymbol{\theta})=(2\pi)^{-p/2}\,|\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{x}_\lambda-\boldsymbol{\mu})'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{x}_\lambda-\boldsymbol{\mu})\right]$$

N回の独立な観測に関する確率密度関数は以下で与えられる。

$$f(\mathbf{x},\boldsymbol{\theta})=\prod_{\lambda=1}^{N}f(\mathbf{x}_\lambda,\boldsymbol{\theta})$$

最尤法ではこの確率密度関数に実測値$\hat{\mathbf{x}}_\lambda$を代入した尤度関数$f(\boldsymbol{\theta})$を最大化するようにパラメータを決定する。実際には計算の簡単化のため、尤度関数を対数変換した対数尤度関数の符号を変えたものを最小化する。符号を変えた対数尤度関数は以下で与えられる。

$$
\begin{aligned}
-\log f(\boldsymbol{\theta}) &= -\sum_{\lambda=1}^{N} \log f(\hat{\mathbf{x}}_\lambda, \boldsymbol{\theta}) \\
&= \frac{1}{2}\sum_{\lambda=1}^{N} (\hat{\mathbf{x}}_\lambda - \bar{\mathbf{x}})' \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} (\hat{\mathbf{x}}_\lambda - \bar{\mathbf{x}}) - \frac{N}{2} \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}| + const. \\
&= \frac{N}{2}\left(\mathrm{tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{S}\right) - \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}|\right) + const.
\end{aligned}
$$

但し、

$$
\mathbf{S} = \frac{1}{N}\sum_{\lambda=1}^{N} (\hat{\mathbf{x}}_\lambda - \bar{\mathbf{x}})(\hat{\mathbf{x}}_\lambda - \bar{\mathbf{x}})'
$$

通常最尤法の評価関数としては、上の対数尤度関数に定数を加えた以下の式が用いられることが多い。

$$
f_{ML}(\theta) = \mathrm{tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{S}\right) - \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{S}| - n
$$

これらの評価関数の最小化法には様々な方法が用いられるが、現在我々は最小 2 乗法では、Levenberg-Marquart 法、最尤法では最初のパラメータ設定に最小 2 乗法を用い、求められた値を初期値として Levenberg-Marquart 法を応用した Newton-Raphson 法を用いている。

13.3 モデルの評価

ここではモデルの良し悪しを評価するいくつかの指標とその性質についてまとめておく。

解の検定

帰無仮説 H_0：構成されたモデルは正しい。

対立仮説 H_1：構成されたモデルは正しくない。

$$
\chi^2 = (N-1) f_{ML} \sim \chi^2_{df}, \qquad df = \frac{1}{2} n(n+1) - p
$$

ここに N はデータ数、n は観測変数の数、p は自由パラメータ数（外生観測変数数＋パス係数数＋誤差変数数＋共分散（相関）数）であり、df は χ^2 分布の自由度である。この検定はデータ数を増やして精度を上げるほど対立仮説である「モデルは正しくない」という結果が出やすくなるという矛盾を含んでいる。

適合度指標

GFI（Goodness of Fit Index）

これは実測値による共分散行列とパラメータで表された共分散行列の類似の程度を見る指標で以下のように与えられる。

$$\mathrm{GFI} = 1 - \frac{\mathrm{tr}\left(\left(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{S} - \mathbf{I}\right)^2\right)}{\mathrm{tr}\left(\left(\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{S}\right)^2\right)} \quad \text{ここに } \mathrm{tr}(\mathbf{A}^2) = \mathrm{tr}(\mathbf{A}\mathbf{A}')$$

この指標の値は 0.9 以上が良いとされるが、モデルの自由度が大きくなると値を大きくすることが難しくなる。

AGFI（Adjusted Goodness of Fit Index）

これは GFI の自由度の問題を改善した指標で、相関を加えて自由度を見かけ上小さくしても値が改善されるとは限らない指標である。

$$\mathrm{AGFI} = 1 - \frac{n(n+1)}{2df}(1 - \mathrm{GFI})$$

一般に AGFI≦GFI の関係がある。

情報量基準

AIC（Akaike's Information Criterion）

これは一般の統計モデルの評価指標として有名であり、以下で定義される。

$$\mathrm{AIC} = \chi^2 - 2df$$

この値が小さいほど良いモデルとされる。この指標には、標本数が多い場合、自由度が小さい（パラメータ数が多い）モデルが良いモデルと判断される傾向がある。

CAIC（Consistent Akaike's Information Criterion）

これは AIC の標本数の影響を抑えた指標である。

$$\mathrm{CAIC} = \chi^2 - (\log(N) + 1)df$$

パラメータの検定

最尤法の推定値 $\hat{\boldsymbol{\theta}}$ を用いると、以下のようになることが知られている。

$$z_i = \frac{\hat{\theta}_i}{\sigma_{\hat{\theta}_i}} \sim \mathrm{N}(0,1) \quad \text{ここに、} \sigma_{\hat{\theta}_i} \equiv \frac{N-1}{2}\frac{\partial^2}{\partial \theta_i^{\,2}} f_{ML}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

これを用いてパラメータの値を 0 と比較する検定を行うことができる。

13.4 プログラムの動作

ここでは２章で述べた例を用いてプログラムの動作を説明する。プログラムを起動すると図 4.1a のような初期メニューが表示される。これは授業用にできるだけ簡易化したメニューである。この中で拡張メニューボタンをクリックすると図 4.1b のような拡張メニューが表示される。

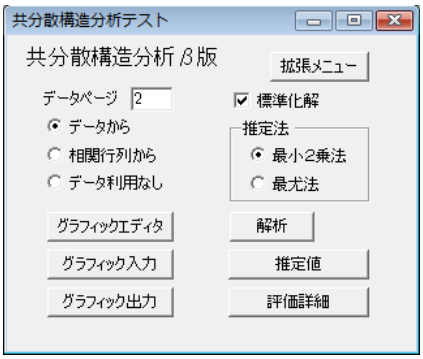


図 4.1a　初期メニュー画面

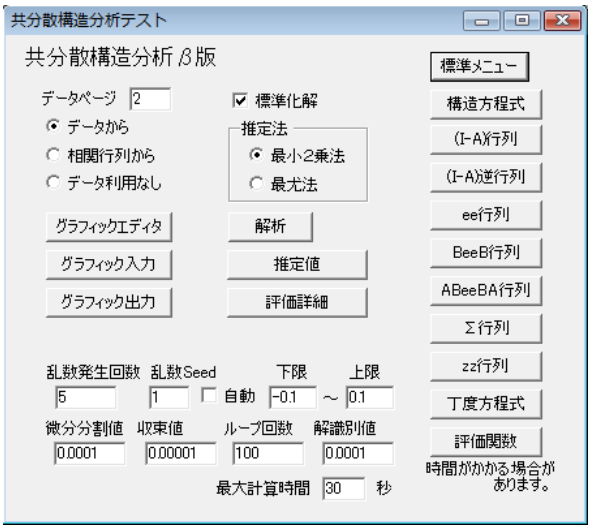


図 4.1b　拡張メニュー画面

拡張メニューには細かな設定や、数式表示のためのボタンが含まれている。以後すべての機能が揃った拡張メニューをもとに説明していく。これらのメニューの中の「グラフィックエディタ」、「グラフィック入力」、「グラフィック出力」ボタンについては他の分析との関係でまだ十分に検討しておらず、ここでは触れない。

共分散構造分析のデータは基本的にデータ構造を記述したページと観測変数のデータ値を表すページに分かれる。前者を図 4.2 に示す。後者については通常の統計データの画面である。

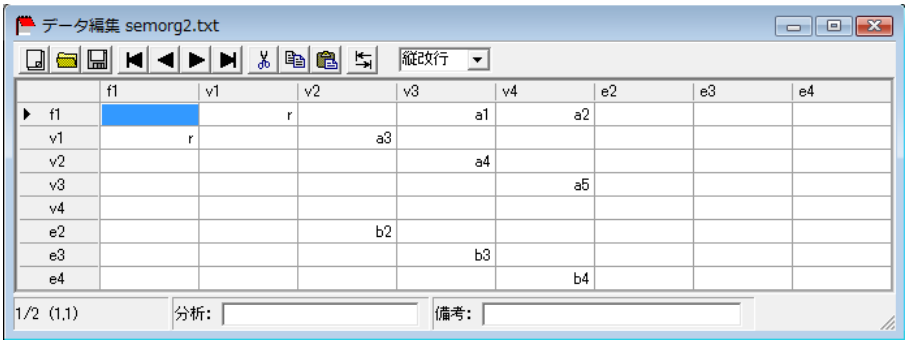


図 4.2　構造データ

分析は、メインメニュー左上の「データページ」テキストボックスに観測値のページ番号を記入し、図 4.2 の構造データを表示して実行する。１つの観測データに複数のモデルを考える場合は、データを１ページ目にして、２ページ目以降を構造データにするのがよい。

最初に共分散構造分析の基礎となる数式について、表示結果を説明する。図 4.1b のメニュー画面の「構造方程式」ボタンをクリックすると構造方程式が図 4.3 のように表示される。

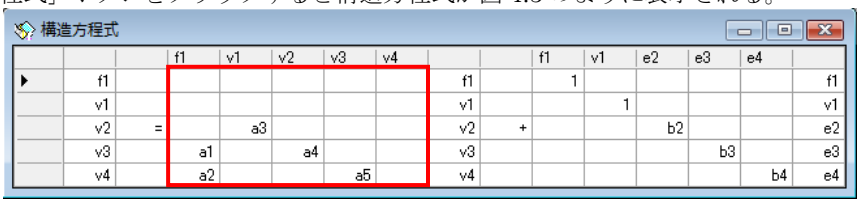


図 4.3　構造方程式

ここに２節（1）式中の行列 **A** は図の四角形で囲まれた部分である。メニューの「(I-A) 行列」ボタンクリックすると図 4.4 のように **I**−**A** 行列の結果が表示される。

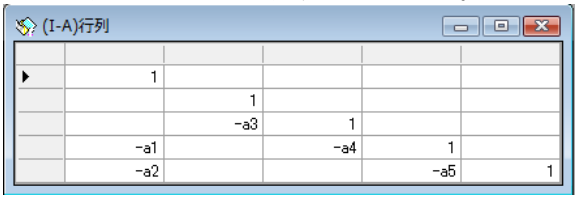


図 4.4（I-A）行列

メニューの「(I-A) 逆行列」ボタンクリックすると図 4.5 の **I**−**A** 逆行列が表示される。

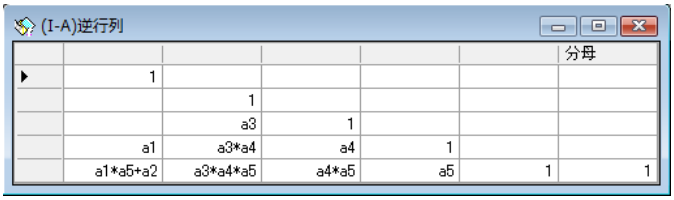


図 4.5（I-A）逆行列

分母の列の最下行では **I**−**A** 行列の行列式を表す。メニューの「ee 行列」ボタンをクリックすると図

4.6 のように行列 **H** が表示される。

ee 行列

	f1	v1	e2	e3	e4
f1	1	r			
v1	r	1			
e2			1		
e3				1	
e4					1

図 4.6　ee 行列

モデルで相関を仮定した部分はここにそのパラメータが残る。後はすべて無相関と仮定される。

メニューの「BeeB 行列」ボタンクリックすると図 4.7 のように行列 $\mathbf{BHB'}$ が表示される。

beeb 行列

	1	r			
	r	1			
			b2*b2		
				b3*b3	
					b4*b4

図 4.7 BHB' 行列

「ABeeBA 行列」ボタンクリックすると図 4.8 のように行列　$\mathbf{(I-A)^{-1}BHB'(I-A)'^{-1}}$ が表示される。

(I-A)Bee'B'(I-A)' 行列

						分母
	1	r	r*a3	a1+r*a3*a4	a1*a5+a2+r...	
	r	1	a3	r*a1+a3*a4	r*a1*a5+r*a...	
	a3*r	a3	a3*a3+b2*b2	a3*r*a1+a3...	a3*r*a1*a5...	
	a1+a3*a4*r	a1*r+a3*a4	a1*r*a3+a3...	a1*a1+a3*a...	a1*a1*a5+a...	
	a1*a5+a2+a...	a1*a5*r+a2...	a1*a5*r*a3...	a1*a5*a1+a...	a1*a5*a1*a...	1

図 4.8 (I-A)$^{-1}$BHB'(I-A)'$^{-1}$ 行列

「Σ行列」ボタンクリックすると図 4.9 のように丁度方程式左辺の $\mathbf{\Sigma(\theta)}$ が表示される。

Σ行列

					分母
	1	a3	r*a1+a3*a4	r*a1*a5+r*a...	
	a3	a3*a3+b2*b2	a3*r*a1+a3...	a3*r*a1*a5...	
	a1*r+a3*a4	a1*r*a3+a3...	a1*a1+a3*a...	a1*a1*a5+a...	
	a1*a5*r+a2...	a1*a5*r*a3...	a1*a5*a1+a...	a1*a5*a1*a...	1

図 4.9　Σ行列

「zz 行列」ボタンクリックすると図 4.10 のように行列 **U** が表示される。これは観測変数の共分散行列（標準化解の場合は相関行列）である。

zz 行列

	v1	v2	v3	v4
v1	1	0.834553	0.895450	0.887599
v2	0.834553	1	0.879707	0.826780
v3	0.895450	0.879707	1	0.953547
v4	0.887599	0.826780	0.953547	1

図 4.10 観測変数の相関行列

メニューの「丁度方程式」ボタンクリックすると図 4.11 のように丁度方程式が表示される。

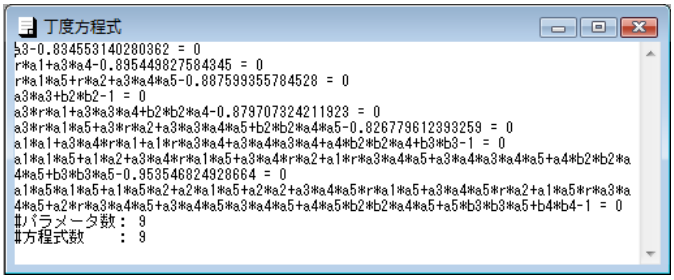


図 4.11　丁度方程式

メニューの「評価関数」ボタンをクリックすると図 4.12 のように評価関数が表示される。

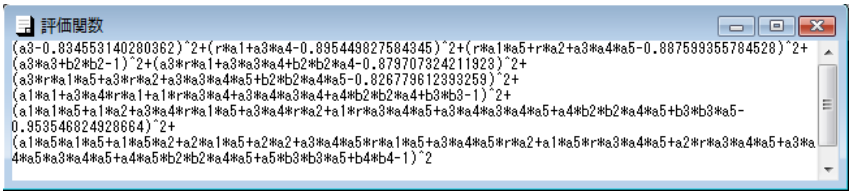
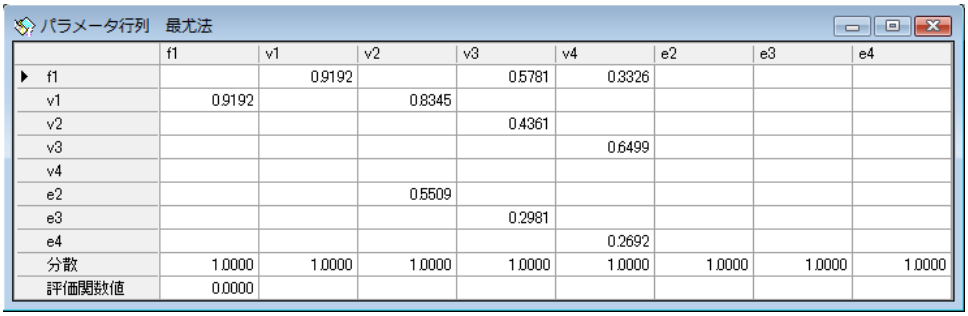


図 4.12　評価関数

これは最小 2 乗法における評価関数で、これを最小化するようにパラメータは選ばれる。最尤法の場合、表示が膨大になるのでかなり時間がかかる場合がある。

推定値については、「推定法」のグループの「最尤法」ラジオボタンを選択して、最初に「解析」ボタンクリックし、それから「推定値」をクリックすると図 4.13 のように表示される。



	f1	v1	v2	v3	v4	e2	e3	e4
f1		0.9192		0.5781	0.3326			
v1	0.9192		0.8345					
v2				0.4361				
v3					0.6499			
v4								
e2			0.5509					
e3				0.2981				
e4					0.2692			
分散	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
評価関数値	0.0000							

図 4.13　最尤法の推定値

これは最尤法の推定値であるが、丁度方程式の解でもある。グラフィックエディタを用いて構造図を作成した場合は、構造図中にも推定値が表示されるようにしたい。さらに「評価詳細」ボタンをクリックすると、異なった形式の推定値と評価値が図 4.14a と図 4.14b のように表される。

パラメータ推定値

	変数		変数	推定値	標準誤差	検定値	両側確率
r	v1	<->	f1	0.9192	0.0620	14.8289	0.0000
a1	v3	<-	f1	0.5781	0.1353	4.2719	0.0000
a2	v4	<-	f1	0.3326	0.2424	1.3724	0.1699
a3	v2	<-	v1	0.8345	0.1023	8.1566	0.0000
a4	v3	<-	v2	0.4361	0.1265	3.4490	0.0006
a5	v4	<-	v3	0.6499	0.2319	2.8030	0.0051
b2	v2	<-	e2	0.5509	0.0723	7.6156	0.0000
b3	v3	<-	e3	0.2981	0.0849	3.5097	0.0004
b4	v4	<-	e4	0.2692	0.0514	5.2398	0.0000
評価関数値	0.0000						

図 4.14a 推定値の詳細表示

モデルの評価

```
推定値の評価
推定値の検定（データ数が増えれば「正しくない」と結論され易いことに注意）
自由度の値が不適切で計算できません。

適合度指標
  自由度    0
  GFI（≧0.9 が良いが、自由度が小さくなると改善されることに注意
    1.0000
  AGFI（AGFI≦GFI 自由度を小さくしても必ずしも改善されない）
    自由度の値が不適切で計算できません。

情報量基準　χ2-2*df
  AIC（モデル比較の代表的指標、値が小さいほど良いモデル
      標本数が多くなると自由度が小さいほど「良く」なる特徴を持つ）
    0.0002
  CAIC（AICより標本の影響を抑えた指標）
    0.0002

自由度当りのモデルの分布と最尤モデルとの乖離を表す指標
  RMSEA（≦0.05:当てはまりが良い、≧0.1:当てはまりが悪い）
    自由度の値が不適切で計算できません。
```

図 4.14b　モデルの評価（表示の後半部分）

最小 2 乗法の場合、推定値の検定部分や評価指標の適合度指標以外の部分は表示されない。

13.5 Amos との比較

我々はプログラムの評価のために、我々の結果と Amos の結果とを以下の構造図の場合について比較した[1)]。まだ我々の計算のアルゴリズムが不十分なため、ごく小さなモデルについてのみの比較に限られている。なお名称は参考文献に名前がある場合はその名前を使用し、名前がない場合は我々が与えた。また結果の符号については、潜在変数の符号の任意性に起因すると思われる場合は、結果が同一のものと判断した。

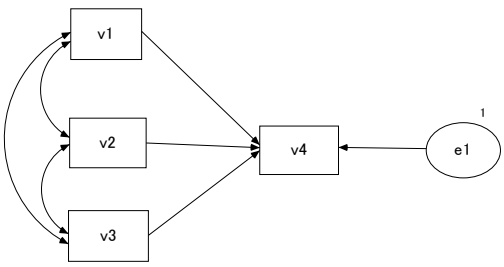


図 5.1 回帰分析モデル

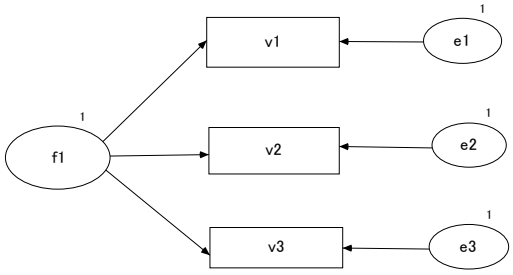


図 5.2 因子分析モデル

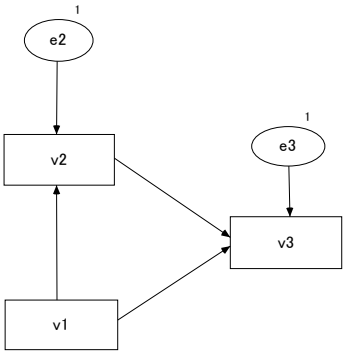


図 5.3 回帰分析の複合モデル 1

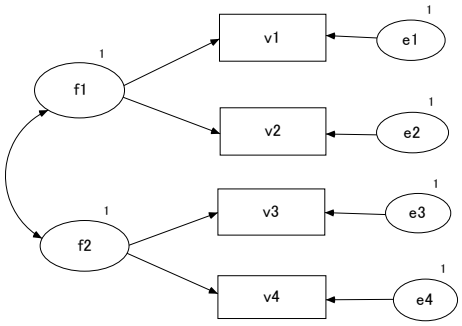


図 5.4 因子分析の複合モデル 1

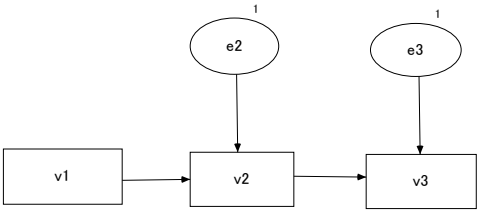


図 5.5 連結モデル

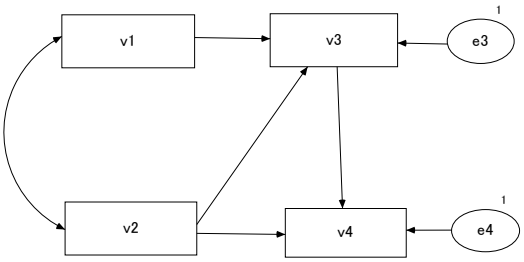


図 5.6 逐次モデル

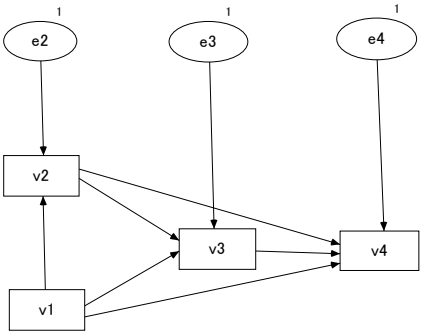


図 5.7 回帰分析の複合モデル 2

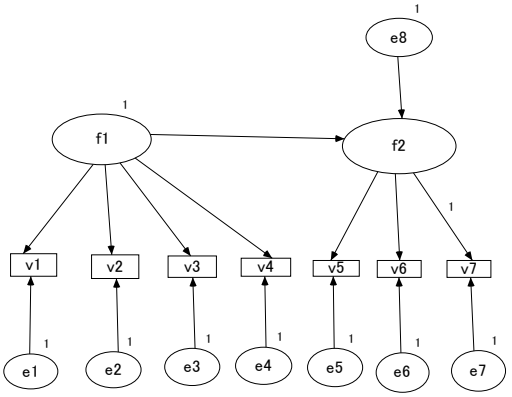


図 5.8 因子分析の複合モデル 2

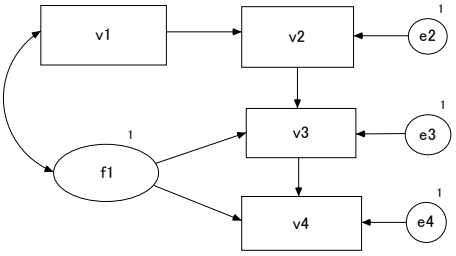


図 5.9 回帰分析と因子分析の複合モデル

いずれの場合もモデルがデータをよく表す場合は Amos の結果と我々のプログラムの結果は一致

する。しかしモデルがデータに適合しない場合（我々は乱数を用いてデータを作成して試した)、興味深い結果が出たので紹介する。

表 5.1 は図 5.4 の場合の最尤法による両者の比較である。

表 5.1 図 5.4 の結果の比較

	変数		変数	Amos	CAnalysis	CAnalysis 別解
r	f2	<->	f1	-0.253	0.2532	45.8391
a1	v1	<-	f1	1.000	1.0000	0.0069
a2	v2	<-	f1	-0.244	-0.2444	0.0060
a3	v3	<-	f2	1.000	-0.9998	-1.0218
a4	v4	<-	f2	0.133	-0.1333	-0.1361
b1	v1	<-	e1	0.000	0.0000	1.0000
b2	v2	<-	e2	0.970	0.9697	1.0000
b3	v3	<-	e3	0.000	0.0197	0.0001
b4	v4	<-	e4	0.991	-0.9911	-0.9911
評価関数値				0.087	0.0867	0.0664

我々の結果はパラメータの初期値の与え方によって何種類かの異なる結果が得られ、その中で Amos の結果と一致する解の他に、例えば CAnalysis 別解のような解が得られた。別解では相関係数が 1 以上の値になるが、評価関数値は Amos の値より小さくなる。Amos ではこのような非現実な解は排除しているように見える。さらに我々のプログラムでパラメータを一部固定してみると評価関数値が Amos の値より小さい現実的な推定値を求めることもできた。我々はこれまで評価関数が極値となる推定値を求めようとしてきたが、パラメータの値が 1 の近傍になる場合には境界を持つ最小化問題となっているように思われる。Amos であってもこのような場合には注意する必要がある。

次に表 5.2 は図 5.8 の結果の比較である。

表 5.2 図 5.8 の結果の比較

	変数		変数	AMOS	CAnalysis
a0	f2	<-	f1	0.173	0.8608
a1	v1	<-	f1	-0.157	0.0983
a2	v2	<-	f1	0.692	0.1063
a3	v3	<-	f1	0.139	0.4869
a4	v4	<-	f1	0.270	-0.0344

a5	v5	<-	f2	-0.235	-0.6597
a6	v6	<-	f2	0.106	0.0874
a7	v7	<-	f2	1.000	0.5206
c1	v1	<-	e1	0.988	0.9952
c2	v2	<-	e2	0.722	0.9944
c3	v3	<-	e3	0.990	-0.8736
c4	v4	<-	e4	0.963	0.9994
c5	v5	<-	e5	0.972	0.8233
c6	v6	<-	e6	0.994	-0.9972
c7	v7	<-	e7	0.000	0.8940
c8	f2	<-	e8	0.986	0.0001
評価関数値				0.190	0.1451

ここでは現実的な値の範囲で Amos より良い解が得られている。この場合にも Amos の推定値の中に境界値 1 が含まれている。またこのような場合でも Amos での GFI の値が 0.951 と高いことにも注意を要する。

さらに図 9 については Amos と College Analysis で同じ解が得られ、いずれも標準化解のパラメータの推定値が非現実な値となる場合も見られた。これを見ると Amos でも完全に非現実なパラメータを除外しているわけではなさそうである。

実際の分析ではパラメータの推定値が現実的な値となるようなモデルを考えるため、ここで述べたようなことは起こらないが、分析に不慣れな利用者は十分注意する必要がある。特に非標準化解の場合はそれに気が付かない可能性もあるので、結果の検討が必要である。

13.6 今後の課題と展望

我々は共分散構造分析についてプログラムの開発を進め、中間段階にまで到達した。殆ど知識のない状態から始めたので、計算の手順の失敗やアルゴリズムの問題から計算時間の短縮にかなり回り道をした。しかしこれらの問題を考える過程で知識を得ることもできた。特に計算時間については実際にプログラムを作成しなければ分らない部分も多い。対話的に処理を行う場合、著者らは人がストレスなく待てる計算時間の上限を 10 秒程度に考えているが、これまでに College Analysis の中で開発してきたプログラムでは特に気になることはなかった。しかし共分散構造分析のプログラムでは、今のままのアルゴリズムでは、4 章で試したモデル程度が限界である。この意味でも Amos で採用されているマルコフ連鎖モンテカルロ法は優れている。我々のプログラムを実用的なものにするためには

どうしても取り入れなくてはならない。

構造図について我々のプログラムでは行列形式で入力するが、紙に書かれたものを入力する場合はかなり効率良く行える。しかし、頭の中でモデルを考える場合、この行列形式の入力は有効とは言えない。このため我々は新しくグラフィック入力用のエディタを開発しなければならない。これは共分散構造分析のメニューから呼び出し、構造図を作成して、結果を表形式のエディタに戻すものにする。また他の分析でも使用するため、汎用的なものにすることも必要である。現在、その大部分は開発が終わっており、その実行画面は図 6.1 のようになる。このエディタの機能やデータ構造については他の分析との関係もあるので、別の機会に詳しく説明する。

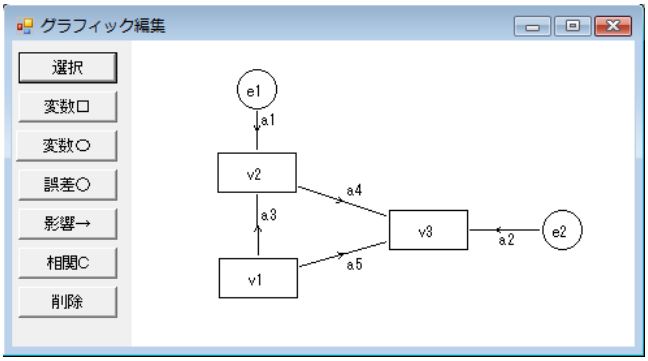


図 6.1 グラフィックエディタ画面

１４．パス解析

パス解析は観測変数間に線形の関係が仮定されるとき、因果関係の方向性を議論するために利用される手法で、共分散構造分析の特別な場合に相当する。ここではプログラムを実際に動かし、動きを見ながら、理論についても解説する。

メニュー［分析－多変量解析－パス解析］を選択すると、図１のようなパス解析実行メニューが表示される。

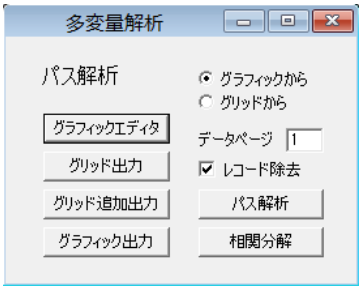


図１　パス解析実行メニュー

「グラフィックエディタ」ボタンで、グラフィックエディタを起動し、例えば図２のような構造図を描く。

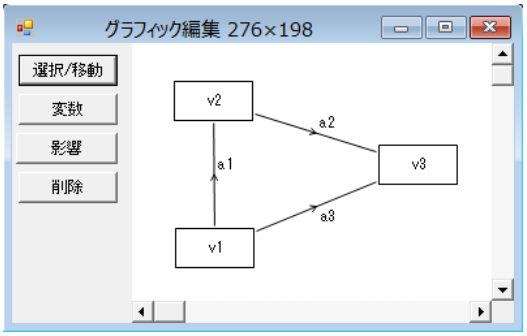


図２　構造図

共分散構造分析の構造図では誤差変数についても描画するが、パス解析では誤差変数の入り方は明らかであるため描画しない。図は単純に観測変数とそれらの間の影響だけで描かれる。但し、影響はすべての変数を結ぶものとし、影響のループは含まないものとする。

これらの変数名のデータは、グリッドエディタで、実行メニューの「データページ」テキストボックスに指定されたページに含まれるものとする。プログラムはデータページの変数の中で、変数名に合うデータを利用する。

変数間の構造データは、ラジオボタンにより、グリッドエディタとグラフィックエディタのどちらかを選ぶことができる。通常は、グラフィックエディタからの入力にしておき、良い構造が出来上が

ったら、「グリッド出力」か「グリッド追加出力」によって、構造データをグリッドエディタに移し、保存する。

実行メニューで「パス解析」ボタンをクリックすると図 3 のように、構造間の影響の強さが表示される。

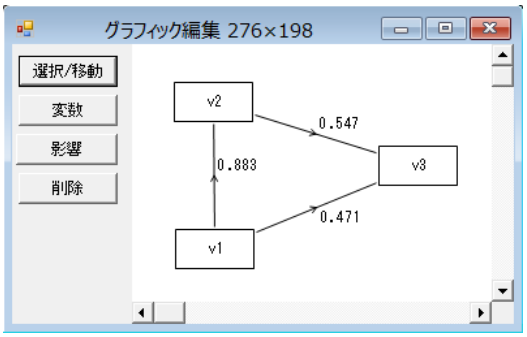


図 3　パス解析結果

これを見て我々は影響の強さ、影響の方向の良し悪しを判定する。これらの影響の強さの数値は以下のような標準化した重回帰式から求められる。

v2=a1*v1+e2

v3=a3*v1+a2*v2+e3

ここで、e2 と e3 は誤差項であり、自分自身を除いて他の変数との相関はないものと考える。

これらの式から、各変数の相関について以下のような関係が分かる。

cov(v1, v2)=cov(v1, a1*v1+e2)=a1*cov(v1, v1)+cov(v1, e2)=a1

cov(v1, v3)=cov(v1, a3*v1+a2*v2+e3)=a3*cov(v1, v1)+a2*cov(v1, v2)=a3+a1*a2

cov(v2, v3)=cov(v2, a2*v2+a3*v1+e3)=a2*cov(v2, v2)+a3*cov(v2, v1)+cov(e3, v2)=a2+a1*a3

第 1 式について a1 を直接相関、第 2 式について、a3 を直接相関、a1*a2 を間接相関、第 3 式について、a2 を直接相関、a1*a3 を擬似相関と呼ぶ。直接相関は変数間を直接的に結ぶ関係、間接相関は変数間の影響を及ぼす方向通りにたどって行って 2 回以上でたどりつく関係、擬似相関は他の変数（ここでは v1）が両者に影響を及ぼしているような関係である。

左辺は相関係数であるので、これらの式は相関係数を、直接相関、間接相関、擬似相関に分解することに相当する。この関係は、実行メニューの「相関分解」をクリックすることで示される。結果を図 4 に示す。

	パス係数	相関係数	直接効果	間接効果	擬似相関
v1->v2	0.8833	0.8833	0.8833	0	0
v1->v3	0.4712	0.9544	0.4712	0.4832	0
v2->v3	0.547	0.9632	0.547	0	0.4162

図 4　相関分解

直接効果、間接効果、擬似相関の合計が相関係数になっていることが分かる。

次に、もう少しだけ複雑なモデルを使って、これらの計算法を考えてみる。図 5 にモデルを示すが、ここではウィンドウの表示は省略する。

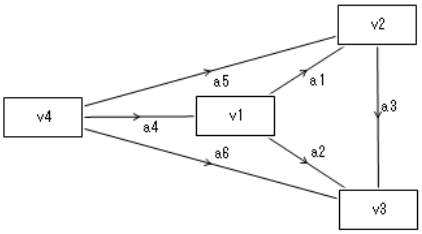


図 5　パスの例 2

ここではこの例を用いて v2, v3 への v4 の擬似相関を見てみよう。重回帰分析の計算を用いると、v2 と v3 の相関係数は以下のように与えられる。

cov(v2, v3)=cov(v2, a3*v2+a2*v1+a6*v4)

=a3+a1*a2+a5*a6+a5*a2*a4+a6*a1*a4

最初の項は直接相関、次の項は v1 からの擬似相関、下線の項は v4 からの擬似相関とみると、ある変数から影響をたどって行った道筋で、同一の変数を通る道筋を除いたものの総和となっている。この場合、v4 から v1 が v4 からの同一の道筋と考えると、そこを通る経路は v1 からの影響に置き換えると考える。これは v4 から v1 への影響が単純に係数の掛け算ではなく、

a1*a2*(a4*a4+cov(e1, e1))=a1*a2

のように回帰分析の際の誤差項の分散も含まれることから納得できる。

参考文献

1) 多変量解析法入門, 永田靖, 棟近雅彦, サイエンス社, 2001.

１５．多次元尺度構成法

多次元尺度構成法（MDS: Multi Dimensional Scaling）は個体間に与えられた、類似度または非類似度（距離）を元に各個体の位置（嗜好性等抽象的な位置関係も含む）を求める手法である。個体間の非類似度がユークリッド空間上の距離として与えられる場合を計量 MDS、非類似度が順序のみ意味を持つ場合を非計量 MDS と呼ぶ。我々はこれらの手法を順番に説明する。

15.1 計量 MDS

個体 i と個体 j（$1 \le i, j \le n$）の距離を d_{ij} とし、距離が以下の関係を満たすとき計量 MDS の手法が利用できる。

$$d_{ij} \ge 0$$
$$d_{ij} = d_{ji}$$
$$d_{ij} + d_{jk} \ge d_{ik}$$

今、 p 次元のユークリッド空間中の個体 i の位置を $x_{i\alpha}$（$1 \le \alpha \le p < n$）とする。個体 i と個体 j との距離 d_{ij} は以下のように求められる。

$$d_{ij} = \left(\sum_{\alpha=1}^{p} (x_{i\alpha} - x_{j\alpha})^2 \right)^{1/2}$$

この距離は原点の取り方に依存しないので、原点を個体の重心に設定するものとする。そのとき、

$$\bar{x}_\alpha = \frac{1}{n} \sum_{i=1}^{n} x_{i\alpha} = 0 \qquad (1)$$

である。原点から個体 i, j へのベクトルの内積を z_{ij} とすると、これは余弦定理により、以下のように与えられる。

$$z_{ij} = \sum_{\alpha=1}^{p} x_{i\alpha} x_{j\alpha} = \frac{1}{2} \left(d_{i0}^2 + d_{j0}^2 - d_{ij}^2 \right) \qquad (2)$$

ここに、 d_{i0} は原点から個体 i までの距離である。(1) の関係式を使うと以下となるが、

$$\sum_{i=1}^{n} z_{ij} = \sum_{j=1}^{n} z_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{n} z_{ij} = 0$$

これに(2) 式を代入して、 d_{i0} についての関係式を求め、 z_{ij} を以下のように書き換えることができる。

$$z_{ij} = \frac{1}{2} \left(\sum_{k=1}^{n} \frac{d_{kj}^2}{n} + \sum_{k=1}^{n} \frac{d_{ik}^2}{n} - \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{d_{kl}^2}{n^2} - d_{ij}^2 \right) \qquad (3)$$

我々は求められた距離行列から、(3)式によってこの内積で作られた行列 $\mathbf{Z}$ を求め、(2)の最初の等号関係を用いて、後に示す方法で位置 $x_{i\alpha}$ を求める。

行列 $\mathbf{Z}$ は p 個の固有値 $\lambda_\alpha\ (\geq 0)$ を対角成分に並べた対角行列 $\mathbf{\Lambda}(p\times p)$ とその固有値に対する固有ベクトル $\mathbf{y}_\alpha$ を横に並べた行列 $\mathbf{Y}=(\mathbf{y}_1\cdots\mathbf{y}_p)$ によって以下のように分解できることが知られている（エッカート・ヤングの定理)。

$$\mathbf{Z}=\mathbf{Y}\mathbf{\Lambda}^t\mathbf{Y}$$

今、固有値の平方根を対角成分に並べた対角行列を $\mathbf{\Lambda}^{1/2}$ として、$\mathbf{X}=\mathbf{Y}\mathbf{\Lambda}^{1/2}$ とおくと、上の関係式は以下のようになる。

$$\mathbf{Z}=\mathbf{X}^t\mathbf{X}$$

これを(2)式と比較すると、以下の関係を得る。

$$x_{i\alpha}=\sqrt{\lambda_\alpha}\,y_{i\alpha}$$

15.2 非計量 MDS

非計量 MDS では、非類似度 s_{ij} を用いるが、これをディスパリティと呼ばれる量 $\hat{d}_{ij}$ に変換して利用する。これらは以下の関係を満たすようにする。

$$s_{ij}>s_{kl}\Rightarrow\hat{d}_{ij}\geq\hat{d}_{kl}$$

$$s_{ij}=s_{kl}\Rightarrow\hat{d}_{ij}=\hat{d}_{kl}$$

ディスパリティの生成は参考文献 2) に示された以下の手順で行う。ある手法で（我々のプログラムでは非類似度 s_{jk} を用いた計量 MDS の手法)、位置が求まっているとする。その位置から距離 d_{jk} を求める。非類似度 s_{jk} を小さい順に並べ、それに $s_1, s_2, \cdots, s_l$ と番号を付ける。s_i に対応する非類似度 s_{jk} に対応する距離 d_{jk} についても同様に番号付けを行っておく。但し、s_i に同順位のものがある場合、それに対応する d_i について、平均をとっておくものとする。

この準備を行った後、以下の手順を実行する。

1）$\hat{d}_1=d_1$ とする。

2）$(k-1)$ 番目までの $\{\hat{d}_i\}$ を作ったとする。

3）$d_k\geq\hat{d}_{k-1}$ のとき、$\hat{d}_k=d_k$ と定める。2）に行き、$\hat{d}_{k+1}$ に移る。

4）$d_k<\hat{d}_{k-1}$ のとき $\hat{d}_k$ とその前の値を以下のように決定、変更する。2）に行き、$\hat{d}_{k+1}$ に移る。

$i=1,2,\cdots,k-2$ と順に変えて、以下を満たす最小の i を見つける。

$$\hat{d}.=\frac{1}{i+1}\left(d_k+\sum_{j=1}^{i}\hat{d}_{k-j}\right)\geq\hat{d}_{k-i-1}$$

見つかれば、$\hat{d}_k=\hat{d}_{k-1}=\cdots=\hat{d}_{k-i}=\hat{d}.$ とする。

見つからなければ、$\hat{d}_k = \hat{d}_{k-1} = \cdots = \hat{d}_1 = \frac{1}{k}\sum_{j=0}^{k-1}\hat{d}_{k-j}$ とする。

5）$\hat{d}_n$ を定めたとき、プロセスを終了する。

ディスパリティ $\hat{d}_{ij}$ の導入は、矛盾を含む非類似度 s_{ij} を、矛盾なく求められる距離 d_{ij} を使って、順序関係を変えずにできるだけ実現可能な値に近づける操作と考えられる。

ディスパリティが求められたら、その値にできるだけ近づけるように再度 d_{ij} を構成しなおす。その基準をストレスと呼び、以下のように定義する。

$$S = \sqrt{\sum_{i=1}^{n}\sum_{j=i+1}^{n}(d_{ij}-\hat{d}_{ij})^2 \Big/ \sum_{i=1}^{m}\sum_{j=i+1}^{m}d_{ij}^2} \qquad \text{ここに、} d_{ij} = \left(\sum_{\alpha=1}^{p}\left|x_{i\alpha}-x_{j\alpha}\right|^t\right)^{1/t}$$

一般にこの距離をミンコフスキー距離、t の値をミンコフスキー定数と呼ぶ。特ミンコフスキー定数が 2 の場合がユークリッド距離である。我々のプログラムでは、S の最適化の方法は最急降下法を用い、$x_{i\alpha}$ の初期値には、元の計量 MDS から求めた値を使っている。ストレスの定義は参考文献 2) で別の定義を示しているが、ここでは参考文献 1）の定義に従っている。

次元数 p を増やして行く際のストレスの変化を表す折れ線グラフ（ストレスプロット）を描き、どの次元から適合度が良くなるか調べる。また、s_{ij} の値を横軸に取り、縦軸にその値に対応する $\hat{d}_{ij}$ 及び d_{ij} の値を 2 種類のマーカーでポイントする。これをシェパードダイアグラムと呼ぶ。$\hat{d}_{ij}$ の値は同じ値を取るものがあるので、$\hat{d}_{ij}$ の上下に d_{ij} が散らばる傾向があるが、これらの点が $\hat{d}_{ij}$ に近く、s_{ij} の大きさによる逆転が起こらないほど適合度は高い。推測されたデータの点 $x_{i\alpha}$ の α の 2 つの次元（$\alpha = 1, 2$ の場合が多いが）について平面上に点を描いて、位置を確かめることも多い。

15.3 プログラムの動作

メニュー［分析－多変量解析他－多次元尺度構成法］を選択すると図 1 のような多次元尺度構成法実行メニューが表示される。

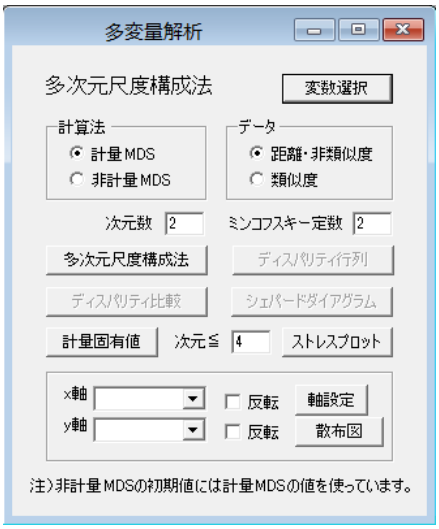
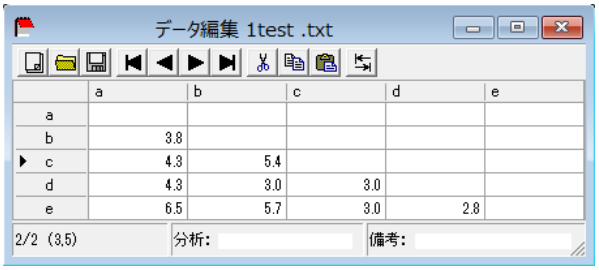


図 1 多次元尺度構成法実行メニュー

データには、図 2 のように、類似度が低いほど大きな値を取る非類似度データ（または距離データ）か、類似度が高いほど大きな値を取る類似度データを用いる。これらの選択は「データ」グループボックスで指定する。



	a	b	c	d	e
a					
b	3.8				
c	4.3	5.4			
d	4.3	3.0	3.0		
e	6.5	5.7	3.0	2.8	

図 2 非類似度データ

非類似度データの場合、対角成分は空欄か 0 にする。類似度データの場合、対角成分は空欄か、最も大きな値を取るものとする。類似度データの場合はこの最大の値から各セルの値を引いたものを非類似度データの値として用いている。データは図 2 のように三角データか、対称データを用いる。非対称データの場合の処理もできるが、我々のプログラムでは、2 つの対応するデータの平均を取ることで対称化して利用している。

計量 MDS か非計量 MDS かは「計算法」グループボックスで指定する。計量 MDS の場合、ミンコフスキー定数は通常 2 で考える。このデータでは次元数を 2 として、「変数選択」した後、「多次元尺度構成法」ボタンをクリックすると図 3 のような位置座標に関する実行結果が表示される。

	1次元	2次元
a	2.844	-1.747
b	2.308	2.019
c	-1.463	-1.846
d	-0.473	0.974
e	-3.216	0.600

図 3 計量 MDS の実行結果

計算途中の非類似度行列の固有値と固有ベクトルは、「計量固有値」ボタンをクリックすることで、図 4 のように表示される。

	1次元	2次元
固有値	26.124	11.840
a	0.556	-0.508
b	0.452	0.587
c	-0.286	-0.536
d	-0.093	0.283
e	-0.629	0.174

図 4 非類似度行列の固有値と固有ベクトル

「次元≦」を 4 で、「ストレスプロット」ボタンをクリックすると、図 5 のようなグラフが表示される。

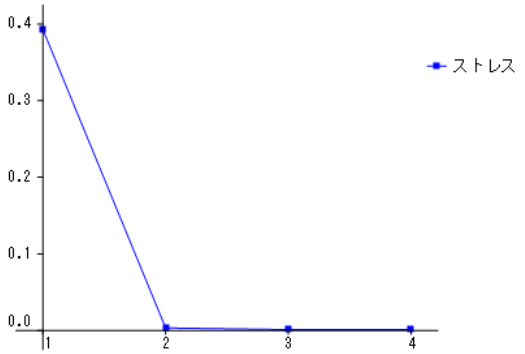


図 5 ストレスプロット

ストレス値の評価は、0.2：良くない、0.1：悪くはない、0.05：良い、0.025：非常に良い、というように言われている。この例の場合だと、2 次元の段階で評価が良くなっているので、2 次元の結果を受け入れる。

2 次元の実行結果の位置を図として表示するために、「軸設定」ボタンで軸を選択し（この場合は自動的に 2 つの次元)、「散布図」ボタンをクリックすると、図 6 のような結果が表示される。

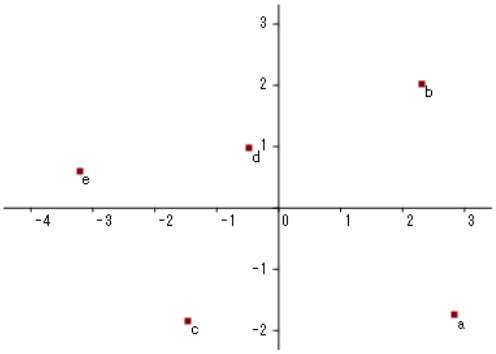


図 6 位置関係結果

次に、参考文献１）にある例題を用いて非計量 MDS の操作を説明する。図 7 に距離行列を示す。これは類似度データである。

	クラウン	セドリック	サニー	マークⅡ	カローラ	スカイライン	マーチ	ヴィッツ	RAV4	パジェロ
クラウン	10									
セドリック	9	10								
▶ サニー	6	7	10							
マークⅡ	7	9	8	10						
カローラ	5	6	8	8	10					
スカイライン	2	3	6	3	6	10				
マーチ	2	3	5	4	7	6	10			
ヴィッツ	1	2	4	3	5	5	9	10		
RAV4	1	1	2	1	3	3	7	8	10	
パジェロ	2	3	3	4	5	2	5	5	4	10

図 7 非計量、類似度データ

実行メニューの「計算法」で「非計量 MDS」を選び、「データ」グループボックスで「類似度」を選ぶ。「次元数」を 2 にして、すべての変数を選択し、「多次元尺度構成法」ボタンをクリックすると、図 8 のような結果が表示される。

	1次元	2次元
▶ クラウン	4.494	-0.955
セドリック	3.983	-0.576
サニー	2.131	1.772
マークⅡ	3.335	-0.779
カローラ	0.806	0.610
スカイライン	-1.225	4.126
マーチ	-2.914	0.601
ヴィッツ	-4.009	0.105
RAV4	-4.889	-1.177
パジェロ	-1.713	-3.727

図 8 非計量 MDS の実行結果

実行メニューの「ディスパリティ行列」ボタンをクリックすると、図 6 の類似度データに対応するディスパリティを図 9 のように表示する。但し、類似度データは、非類似度＝類似度最大値－類似度、によって、非類似度に変更されている。

ディスパリティ行列

	クラウン	セドリック	サニー	マークII	カローラ	スカイライン	マーチ	ヴィッツ	RAV4	パジェロ
クラウン	0.0000	0.8391	3.8135	2.6355	4.6990	7.5870	7.5870	8.7703	8.7703	7.5870
セドリック	0.8391	0.0000	2.6355	0.8391	3.8135	6.7177	6.7177	7.5870	8.7703	6.7177
サニー	3.8135	2.6355	0.0000	2.2557	2.2557	3.8135	4.6990	5.6701	7.5870	6.7177
マークII	2.6355	0.8391	2.2557	0.0000	2.2557	6.7177	5.6701	6.7177	8.7703	5.6701
カローラ	4.6990	3.8135	2.2557	2.2557	0.0000	3.8135	2.6355	4.6990	6.7177	4.6990
スカイライン	7.5870	6.7177	3.8135	6.7177	3.8135	0.0000	3.8135	4.6990	6.7177	7.5870
マーチ	7.5870	6.7177	4.6990	5.6701	2.6355	3.8135	0.0000	0.8391	2.6355	4.6990
ヴィッツ	8.7703	7.5870	5.6701	6.7177	4.6990	4.6990	0.8391	0.0000	2.2557	4.6990
RAV4	8.7703	8.7703	7.5870	8.7703	6.7177	6.7177	2.6355	2.2557	0.0000	5.6701
パジェロ	7.5870	6.7177	6.7177	5.6701	4.6990	7.5870	4.6990	4.6990	5.6701	0.0000

図 9 ディスパリティ行列

「ディスパリティ比較」ボタンをクリックすると、図 10 のように非類似度、ディスパリティ、距離を非類似度の昇順に並べた表が表示される。

ディスパリティ比較

	S	D.P.	Dist
s2,1	1.0000	0.8391	0.6361
s4,2	1.0000	0.8391	0.6796
s8,7	1.0000	0.8391	1.2015
s4,3	2.0000	2.2557	2.8210
s5,3	2.0000	2.2557	1.7614
s5,4	2.0000	2.2557	2.8851
s9,8	2.0000	2.2557	1.5551
s4,1	3.0000	2.6355	1.1728
s3,2	3.0000	2.6355	2.9910
s7,5	3.0000	2.6355	3.7209
s9,7	3.0000	2.6355	2.6574
s3,1	4.0000	3.8135	3.6088
s5,2	4.0000	3.8135	3.3913

図 10 非類似度、ディスパリティ、距離比較表

この関係を図で表したものがシェパードダイアグラムである。「シェパードダイアグラム」ボタンをクリックすると図 11 のようなグラフが表示される。距離の点の散らばり方で、適合の良し悪しをみることができる。

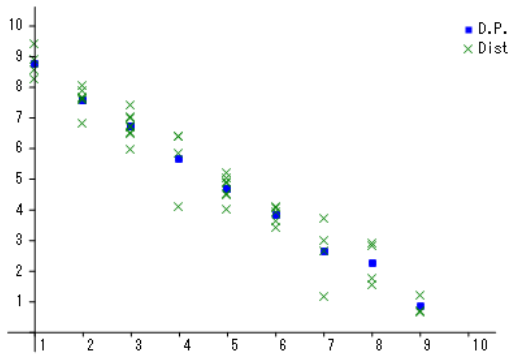


図 11 シェパードダイアグラム

軸を設定して「散布図」ボタンをクリックすると、図 12 のような位置表示のグラフが表示される。

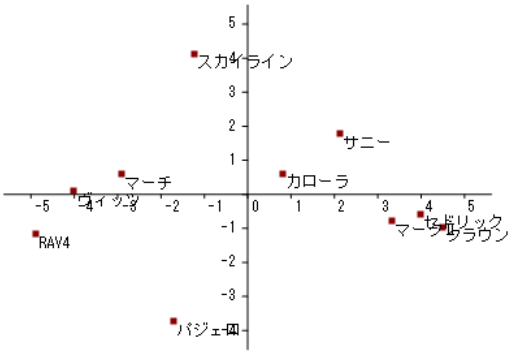


図 12 位置関係結果

次元を増やして行った際の、ストレスの変化は「ストレスプロット」ボタンをクリックすることで図 13 のように得られる。

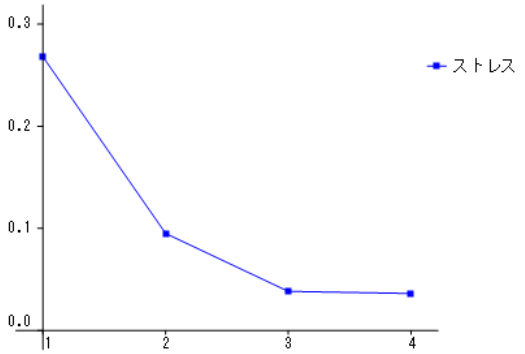


図 13 ストレスプロット

参考文献

1) 多変量解析法入門, 永田靖, 棟近雅彦, サイエンス社, 2001.

2) 関連性データの解析法 多次元尺度構成法とクラスター分析法, 齋藤堯幸, 宿久洋, 共立出版, 2006.

１６．局所重回帰分析

これまでの重回帰分析や非線形最小２乗法の予測手法は、パラメータを含んだ関数形を仮定し、最小２乗法によってパラメータの値を定め、予測関数を確定するものであった。しかし、局所重回帰分析は要求点を与えることによって、その近傍の点による重回帰分析の結果から直接予測値を求める方法で、関数形を必要としない興味深い予測手法である。

16.1　局所重回帰分析の理論

変数i（$i=1,\cdots,p$）、時刻t（$t=T,\cdots,0$）の時系列データ$x_{i,t}$があるとき、その中から時刻tを含めてr期分のそれ以前のデータを取り出す。それらのデータを説明変数とし、時刻$t+a$（$a\geq 1$）のある変数dのデータ$x_{d,t+a}$を目的変数として予測する重回帰分析をパネル重回帰分析という。これはa期先の予測である。

予測値を$X_{d,t+a}$とすると予測式は以下のように与えられる。

$$X_{d,t+a}=\sum_{i=1}^{p}\sum_{j=0}^{r-1}b_{i,j}x_{i,t-j}+b_0 \tag{1}$$

係数$b_{i,j},b_0$は以下の量Lを最小化することによって求める。

$$L=\sum_{t=a+r-1}^{T}\left(x_{d,t}-\sum_{i=1}^{p}\sum_{j=0}^{r-1}b_{i,j+1}x_{i,t-a-j}-b_0\right)^2 \tag{2}$$

今、目的変数と説明変数をそれぞれ以下のように定義し、

$y_\lambda=x_{d,\lambda+a+r-2}$（$\lambda=1,\cdots,T-a-r+2$）

$z_{\alpha,\lambda}=z_{i+pj,\lambda}=x_{i+pj,\lambda+r-2-j}$　　（$i=1,\cdots,p,\ j=0,\cdots,r-1,\alpha=1,\cdots,pr$）

係数をb_αにして(2)式を書き変えると、以下のような式になる。

$$L=\sum_{\lambda=1}^{T-a-r+2}\left(y_\lambda-\sum_{\alpha=1}^{pr}b_\alpha z_{\alpha,\lambda}-b_0\right)^2 \tag{3}$$

これから、偏回帰係数 $\mathbf{b}={}^t\begin{pmatrix}b_0 & b_1 & b_2 & \cdots & b_s\end{pmatrix}$，$s=pr$ は以下のように求めることができる。

$$\mathbf{b}=\left({}^t\mathbf{\Omega\Omega}\right)^{-1}{}^t\mathbf{\Omega y} \tag{4}$$

ここに、

$$\mathbf{y}={}^t\begin{pmatrix}y_1 & y_2 & \cdots & y_N\end{pmatrix},\quad N=T-a-r+2$$

$$\mathbf{\Omega}=\begin{pmatrix}1 & z_{11} & z_{21} & \cdots & z_{s1}\\ 1 & z_{12} & z_{22} & \cdots & z_{s2}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & z_{1N} & z_{2N} & \cdots & z_{sN}\end{pmatrix}$$

時系列分析ではデータが時間の経過とともに明らかになっていくので、現在のすべてのデータから求めたパラメータを使って、過去の各時間の予測を行うことはその時点のデータの影響を強く受け過ぎるという難点がある。そこで、過去の予測を行う際には、その時点までのデータから計算されたパラメータを用いることとし、これによって実測値と予測値の相関を求めることにする。これは一種の交差検証になっている。プログラムにはこの交差検証を付け加えている。

パネル重回帰分析には、他の分析で予測した結果を組み込むことができる。そこで時系列分析の結果をデータとして組み込むことを考えてみた。時系列分析は、傾向変動と周期変動を分解するモデルを考える。データの不規則な大きな変動も考える必要があるので、傾向変動には自然に傾向を求めることができる局所回帰分析を採用した。そのためバンド幅によって局所的な回帰式に影響を与える範囲を限定することができる。また周期変動については、分解する周期（周波数）を複数指定できるようにしている。

標準的な重回帰分析は、目的変数 $y_\lambda\ (\lambda=1,2,\cdots,N)$ と説明変数 $x_{i\lambda}\ (i=1,2,\cdots,p)$ の線形結合 $Y_\lambda=\sum_{i=1}^{p}b_i x_{i\lambda}+b_0$ との差の 2 乗の和 L を最小にするようにパラメータ $b_i\ (i=0,1,2,\cdots,p)$ を決定する。ここに L は以下で与えられる。

$$L=\sum_{\lambda=1}^{N}\left(y_\lambda-Y_\lambda\right)^2=\sum_{\lambda=1}^{N}\left(y_\lambda-\sum_{i=1}^{p}b_i x_{i\lambda}-b_0\right)^2$$

これに対して局所重回帰分析は、各観測値に対してウェイト w_λ をかけて以下の L' を最小化する。

$$L'=\sum_{\lambda=1}^{N}w_\lambda\left(y_\lambda-Y_\lambda\right)^2=\sum_{\lambda=1}^{N}w_\lambda\left(y_\lambda-\sum_{i=1}^{p}b_i x_{i\lambda}-b_0\right)^2$$

この解は、$\mathbf{b}={}^t\begin{pmatrix}b_0 & b_1 & b_2 & \cdots & b_p\end{pmatrix}$ として、以下のように求めることができる。

$$\mathbf{b}=\left({}^t\mathbf{\Omega\Pi\Omega}\right)^{-1}{}^t\mathbf{\Omega\Pi y} \tag{1}$$

ここに、

$$\mathbf{y}={}^t\begin{pmatrix}y_1 & y_2 & \cdots & y_N\end{pmatrix},$$

$$\mathbf{\Omega}=\begin{pmatrix}1 & x_{11} & x_{21} & \cdots & x_{p1}\\ 1 & x_{12} & x_{22} & \cdots & x_{p2}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_{1N} & x_{2N} & \cdots & x_{pN}\end{pmatrix},\quad \mathbf{\Pi}=\begin{pmatrix}w_1 & 0 & \cdots & 0\\ 0 & w_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & w_N\end{pmatrix}$$

要求点 x_i^r の予測値 Y^r は、以下のように与えられる。

$$Y^r=\sum_{i=1}^{p}b_i x_i^r+b_0 \tag{2}$$

ウェイト w_λ は以下のように求める。まず、説明変数についての要求点 x_i^r とバンド幅（調整パラメータ）$p\,(>0)$ を定める。要求点は局所重回帰分析のウェイトの中心を表す点である。次に標準化された観測点 $\tilde{x}_{i\lambda}=\dfrac{x_{i\lambda}-\overline{x}_i}{\sigma_i}$ と標準化された要求点 $\tilde{x}_i^r=\dfrac{x_i^r-\overline{x}_i}{\sigma_i}$ との間のユークリッド距離

$$\Gamma_\lambda=\sqrt{\sum_{i=1}^{p}\left(\tilde{x}_{i\lambda}-\tilde{x}_i^r\right)^2}$$

を求める。但し、標準化の際の標準偏差は不偏分散からのものとする。

この距離 Γ_λ について、その平均を $\overline{\Gamma}$、不偏分散からの標準偏差を σ_Γ とし、これらを用いて、ウェイト w_λ を以下のように定義する。

$$w_\lambda=\exp\left[-\left(\Gamma_\lambda/p\sigma_\Gamma\right)^2\right] \tag{3}$$

これによって要求点の近傍の点にウェイトをかけて最小 2 乗法の解を求めることになる。

標準化偏回帰係数については、標準化されたデータ $\tilde{y}_\lambda,\ \tilde{x}_{i\lambda}$ を用いて、以下のように求めることもできる。

$$\tilde{\mathbf{b}}=\left({}^t\tilde{\mathbf{\Omega}}\mathbf{\Pi}\tilde{\mathbf{\Omega}}\right)^{-1}{}^t\tilde{\mathbf{\Omega}}\mathbf{\Pi}\tilde{\mathbf{y}} \tag{4}$$

ここに、

$$\tilde{\mathbf{y}}={}^t\left(\tilde{y}_1\quad \tilde{y}_2\quad \cdots\quad \tilde{y}_N\right),\qquad \tilde{y}_\lambda=\frac{y_\lambda-\overline{y}}{\sigma_y}\quad \text{（不偏分散を用いた標準化）}$$

$$\tilde{\mathbf{\Omega}}=\begin{pmatrix}1 & \tilde{x}_{11} & \tilde{x}_{21} & \cdots & \tilde{x}_{p1}\\ 1 & \tilde{x}_{12} & \tilde{x}_{22} & \cdots & \tilde{x}_{p2}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & \tilde{x}_{1N} & \tilde{x}_{2N} & \cdots & \tilde{x}_{pN}\end{pmatrix},\quad \mathbf{\Pi}=\begin{pmatrix}w_1 & 0 & \cdots & 0\\ 0 & w_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & w_N\end{pmatrix}$$

別の書式で書くと以下となる。

$$\tilde{b}_i = \frac{\sigma_i}{\sigma_y} b_i , \qquad \tilde{b}_0 = \frac{1}{\sigma_y}\left(b_0 + \sum_{i=1}^{p} b_i \overline{x}_i - \overline{y} \right) \tag{5}$$

この関係は、以下のように求めることができる。

$$\frac{Y_\lambda - \overline{y}}{\sigma_y} = \frac{1}{\sigma_y}\left(\sum_{i=1}^{p} b_i x_{i\lambda} + b_0 - \overline{y} \right)$$

$$= \sum_{i=1}^{p} \frac{\sigma_i}{\sigma_y} b_i \frac{x_{i\lambda} - \overline{x}_i}{\sigma_i} + \frac{1}{\sigma_y}\left(b_0 + \sum_{i=1}^{p} b_i \overline{x}_i - \overline{y} \right)$$

通常の重回帰分析では $\overline{y} = \overline{Y} = \sum_{i=1}^{p} b_i \overline{x}_i + b_0$ であるから、標準化された定数項は 0 になるが、局所重回帰分析では一般に $\overline{y} \neq \overline{Y}$ であるので、標準化された定数項は 0 にならない。

偏回帰係数と標準化偏回帰係数の関係は、(5)式とは逆に以下のように書くこともできる。我々のプログラムではこの関係を利用している。

$$b_i = \frac{\sigma_y}{\sigma_i} \tilde{b}_i , \qquad b_0 = \sigma_y \tilde{b}_0 - \sum_{i=1}^{p} \frac{\sigma_y}{\sigma_i} \tilde{b}_i \overline{x}_i + \overline{y} \tag{6}$$

局所重回帰分析はバンド幅（調整パラメータ） p が無限大になるとウェイトがすべて 1 になり、通常の重回帰分析に近づく。

局所重回帰分析は要求点の近傍で成り立つ近似手法であるので、通常の RMSE や重相関係数の指標は使えず、その信頼性を求める指標は 1 個抜き交差検証法（HOOCV：Leave-One-Out Cross-Validation）を用いて与える。即ち、データ中の 1 点を抜き、その説明変数の座標 $x_{i\lambda}$ を要求点とし、残りの点で局所重回帰分析を行い、要求点の予測値 Y_λ を求める。元々この点には実測値 y_λ があるので予測の誤差が求められる。

局所重回帰分析の精度の指標はこの実測値と予測値を利用し、通常の重回帰分析の RMSE や重相関係数の定義を用いて以下のように与える。もちろんこの指標はバンド幅に影響される。

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{\lambda=1}^{N}\left(y_\lambda - Y_\lambda\right)^2} , \qquad \text{重相関係数} = \frac{\sum_{\lambda=1}^{N}\left(y_\lambda - \overline{y}\right)\left(Y_\lambda - \overline{Y}\right)}{\sqrt{\sum_{\mu=1}^{N}\left(y_\mu - \overline{y}\right)^2 \sum_{\nu=1}^{N}\left(Y_\nu - \overline{Y}\right)^2}} \tag{7}$$

局所重回帰分析は、バンド幅や 1 個抜く点によって必ずしも予測値が求められるとは限らない。そのため、RMSE や重相関係数の値は求められた点だけを用いて計算することもある。

16.2　プログラムの利用法

メニュー［分析－多変量解析等－重回帰分析－局所重回帰分析］をクリックすると図 1 に示すような局所重回帰分析のメニューが表示される。

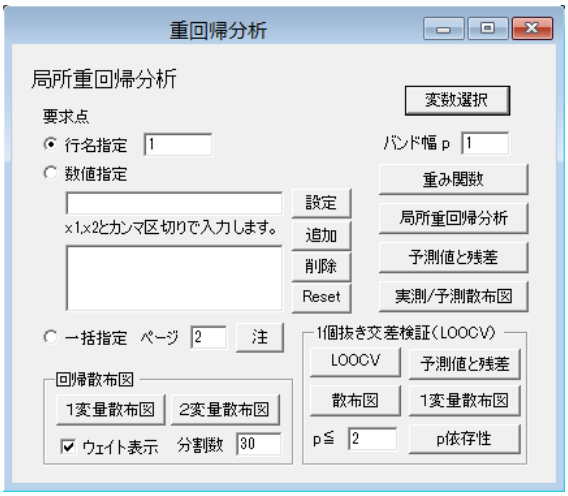


図 1　実行メニュー

通常の重回帰分析と同様に「変数選択」で、目的変数、説明変数の順番に変数を選ぶ。要求点は、「行名指定」でデータから選択するか、「数値指定」で外部から入力する。行名指定は、データの行名の部分の表示で指定する。レコード名が見当たらない場合は、実行の際にメッセージが表示される。数値指定の場合は、テキストボックスに説明変数の値をカンマ区切りで入力する。複数の要求点を調べることが必要であるので、プログラムには入力した値を保存しておく機能が付いている。テキストボックスに書いた要求点のデータは、「追加」ボタンで下のリストボックスに追加保存される。リストボックスのデータは選択して、「設定」ボタンでテキストボックスに呼び戻すことができる。また、選択して「削除」ボタンで 1 つだけリストから削除でき、「Reset」ボタンですべて削除することができる。変数選択の場合と同じ要領で活用できる。

バンド幅を適当な値（ここでは 1）に設定し、適当な行名を指定して「局所重回帰分析」ボタンをクリックすると、図 2 のような分析結果が得られる。

偏回帰係数と要求点

	偏回帰係数	標準化係数	要求点	標準化点
▶ 説明1	0.3812	0.1982	31	-0.9408
説明2	0.4113	0.2146	19	-1.2370
切片/要求予測値	51.6478	-0.5032	71.2791	-0.9552

図 2　偏回帰係数の出力結果

重回帰式による推測結果と各観測点のウェイト値は「予測値と残差」ボタンで図 3 のように表示さ

れる。

予測値と残差

	実測値	予測値	残差	ウェイト
1	66	71.2791	-5.2791	1.0000
2	89	82.2936	6.7064	0.4037
3	73	75.1613	-2.1613	0.6670
4	80	76.4653	3.5347	0.5477
5	75	74.8603	0.1397	0.9383
6	79	67.4070	11.5930	0.8994
7	81	80.4278	0.5722	0.0679
8	66	72.2221	-6.2221	0.9224
9	70	75.4421	-5.4421	0.7975
10	78	77.7895	0.2105	0.6696

図 3　実測値と予測値

実測値と予測値の関係は「実測/予測散布図」をクリックすると、図 4 のように表示される。

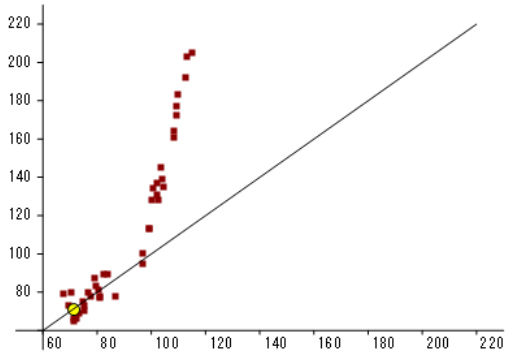


図 4　実測/予測値散布図 1

図中の黄色い点は要求点で、直線は実測と予測が同じであるとする直線である。要求点近傍の点の予測がうまく行っている状況が見える。

偏回帰係数は、要求点とバンド幅に大きく影響を受ける。要求点を変更したときの結果を図 5 に示す。今度は別の点の予測がうまく行っている。

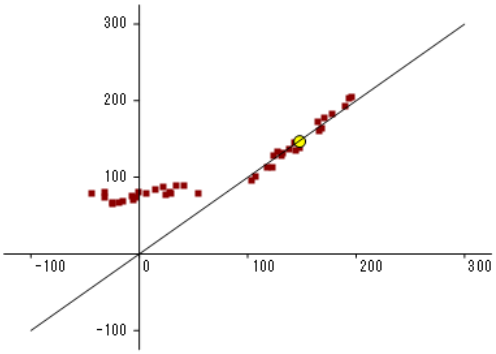


図 5　実測/予測値散布図 2

実際の x,y 軸の上で回帰直線を引いてみる。変数を目的変数と説明変数を 1 つにして、「1 変量回帰

散布図」を描くと図 6 のようになる。2 つの図は要求点を変えて描いている。

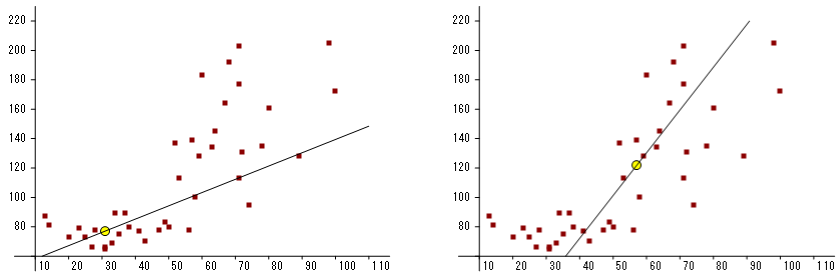


図 6　1 変量回帰散布図（p=1）

これは、データの散布図であり、図中の直線は回帰直線である。要求点によって回帰直線が変化しているのが分かる。

また、実際の x,y,z 軸上で回帰平面を描いてみる。変数を目的変数と説明変数を 2 つにして、「2 変量回帰散布図」を描くと図 7 のようになる。2 つの図は要求点を変えて描いている。

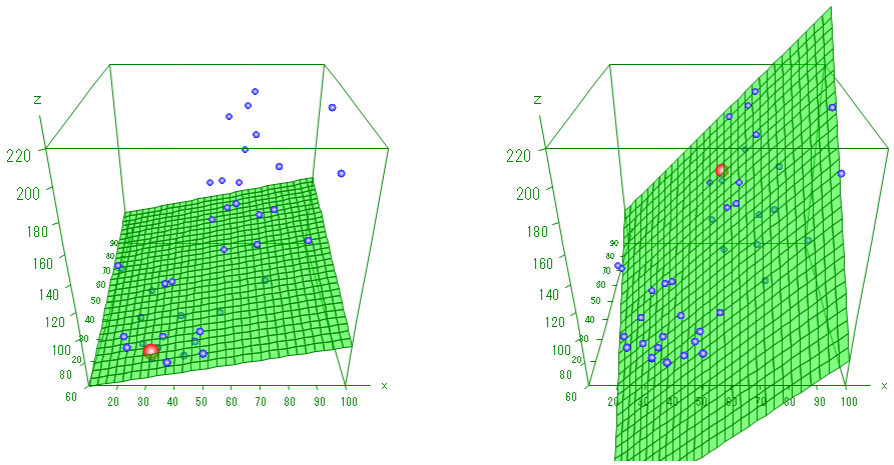


図 7　2 変量回帰散布図（p=1）

次にバンド幅を p=0.5 と p=5 にし、説明変数の数を 1 つにして、1 変数回帰散布図を描く。結果を図 8 に示す。

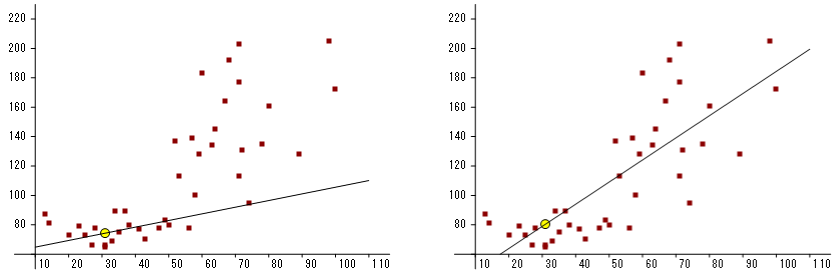


図 8　図 6 左の要求点で p=0.5（左）と p=5（右）の 1 変量回帰散布図

バンド幅の値により、局所性が大きく変更を受けていることが分かる。右側の図は通常の回帰直線に近い。

分析メニューで「重み関数」ボタンをクリックすると 2 変数グラフ描画メニューが表示される。その中の「グラフ描画」ボタンをそのままクリックすると、図 9 左のような実際の重み関数のグラフ（この場合は 2 変量）が表示される。1 変量の場合は図 9 右のようなグラフになる。

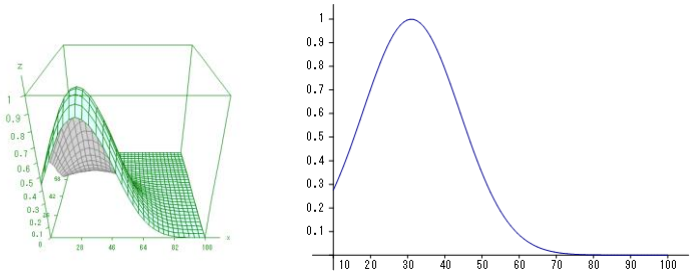


図 9　重み関数グラフ（左は 2 変量、右は 1 変量）

これまでは要求点を 1 点だけ指定したが、現実の分析では多くの要求点を一度に与えて予測値を求めることも考えられる。実行メニューで、要求点の「一括指定」ラジオボタンを選択すると、別のページに与えられた複数の要求点のデータから一括で予測値を求めることもできる。要求点のページはラジオボタン右側の「ページ」テキストボックスに与える。デフォルトは 2 頁目になっているので必要なら変更する。要求点の頁の例を図 10 に示す。

図 10　要求点の一括指定

ここで注意することは、変数名を必ず正確に（全角半角や大文字小文字の区別を付けて）指定することである。分析では変数選択の数や順番が要求点の指定通りとは限らないので、プログラムでは変数名を探して順番等を合わせるようにしている。

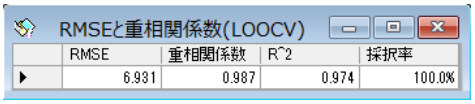
一括指定した要求点を用いた場合は、重回帰式の偏相関係数などは重要でないので、結果は要求点と予測値を表形式で与える。要求点指定に空欄がある場合は、予測値の欄が空欄になる。予測値の出力例を図 11 に与える。


要求点と予測値

	予測値	説明1	説明2
1	71.279	31	19
2	84.700	34	43
3	75.704	25	34
4	78.734	50	14
5	74.550	35	24
11	82.866	13	55
12	71.279	31	19
13	80.777	41	33

図 11　要求点一括指定の出力

局所重回帰分析の予測精度を与えるために、1 個抜き交差検証（LOOCV）を用いた RMSE と重相関係数を与える。「LOOCV」ボタンをクリックすると図 12 のような結果が表示される。

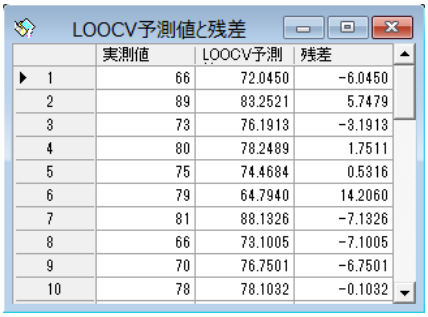

RMSEと重相関係数(LOOCV)

RMSE	重相関係数	R^2	採択率
6.931	0.987	0.974	100.0%

図 12　1 個抜き交差検証による RMSE と重相関係数

ここで採択率は、1 個抜いたデータで計算ができない場合があるので、計算できるデータ点の割合を示したものである。

この求めた予測値と実測値の具体的な値は 1 個抜き交差検証中の「予測値と残差」ボタンをクリックすることで図 13 のように与えられる。予測値が求められなかった部分は空白になっている。


LOOCV予測値と残差

	実測値	LOOCV予測	残差
1	66	72.0450	-6.0450
2	89	83.2521	5.7479
3	73	76.1913	-3.1913
4	80	78.2489	1.7511
5	75	74.4684	0.5316
6	79	64.7940	14.2060
7	81	88.1326	-7.1326
8	66	73.1005	-7.1005
9	70	76.7501	-6.7501
10	78	78.1032	-0.1032

図 13　1 個抜き交差検証による実測値と予測値

この関係は 1 個抜き交差検証中の「散布図」ボタンで、実測/予測散布図として図 14 のように与えられる。

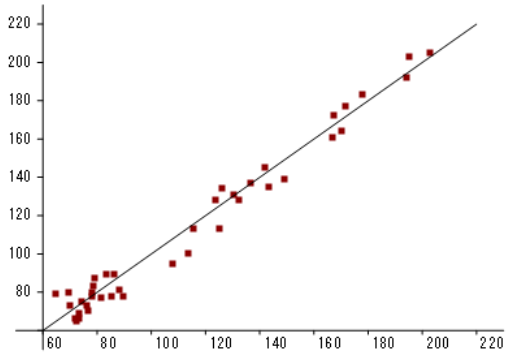


図 14　1 個抜き交差検証による実測/予測散布図

説明変数による予測値と実測値の関係は、1 変量の場合「1 変量散布図」をクリックして図 15 のように与えられる。この図の場合、特別に説明変数を 1 個だけにした。

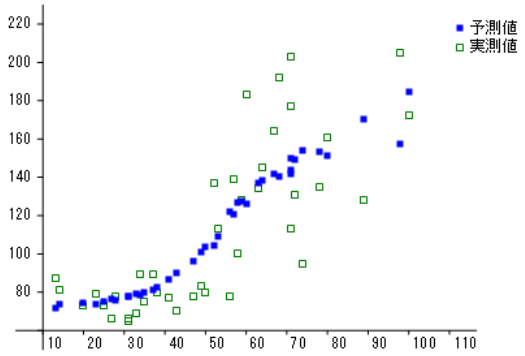


図 15　1 個抜き交差検証による 1 変量散布図

バンド幅によって、RMSE や重相関係数の値は変化する。「p 依存性」ボタンをクリックすると、RMSE のバンド幅 p の値による変化が図 16 のように示される。

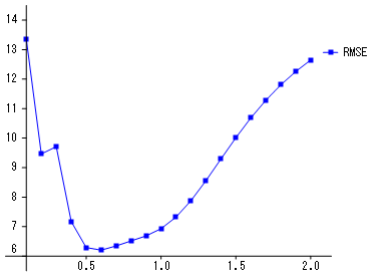


図 16　バンド幅の値による RMSE の変化

ここで、p=0.3 のところで値が急に大きくなっているが、この部分は 1 個抜き交差検証ですべての点が利用できなかった部分である。

参考文献

W.S.Cleveland and S.J.Delvin, Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, Journal of the American Statistical Association, Vol.83, No.403, 596-610 (1988).

１７．数量化Ⅳ類

17.1　数量化Ⅳ類の理論

林の数量化Ⅳ類はデータ間の親近性を仮定し、その中に内在するパターンをデータの空間配置として表現する手法である。r 次元ユークリッド空間中に m 個のデータがあり、データ i とデータ j との親近性（類似度）を e_{ij} とする。親近性には正負の符号の制限はないが、親近性が高いほど大きな値を取るものとする。また一般に対称性 $e_{ij}=e_{ji}$ を仮定しない。同一のデータ同士の親近性 e_{ii} は、後の議論から定義する必要はないが、取り敢えず 0 としておく。r 次元の空間中のデータ i の位置座標を $x_{i\alpha}$（$\alpha=1,2,\cdots,r$）とし、これをベクトルで表し $\mathbf{x}_\alpha=(x_{1\alpha},x_{2\alpha},\cdots,x_{m\alpha})'$ とする。

データ i とデータ j の距離 $d_{ij}\ (\geq 0)$ を位置座標 $x_{i\alpha}$ と $x_{j\alpha}$ を使って、以下のように定義する。

$$d_{ij}^{(r)2}=\sum_{\alpha=1}^{r}(x_{i\alpha}-x_{j\alpha})^2 \tag{1}$$

今、親近性の高いデータ同士は近い距離に位置するように配置したいが、これを実現するために、以下の量 Q を最大化することを考える。

$$Q=-\sum_{i=1}^{m}\sum_{j=1}^{m}e_{ij}d_{ij}^{(r)2}=-\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{\alpha=1}^{r}e_{ij}(x_{i\alpha}-x_{j\alpha})^2 \tag{2}$$

ここで、

$$g_{ij}=h_{ij}-\delta_{ij}\sum_{k=1}^{m}h_{ik},\qquad h_{ij}=e_{ij}+e_{ji} \tag{3}$$

と定義とすると、Q は以下のように書ける。

$$Q=\sum_{\alpha=1}^{r}\sum_{i=1}^{m}\sum_{j=1}^{m}g_{ij}x_{i\alpha}x_{j\alpha}=\sum_{\alpha=1}^{r}\mathbf{x}'_\alpha\mathbf{G}\mathbf{x}_\alpha \tag{4}$$

ここで、$\mathbf{x}_\alpha$ の値によって Q の値はいくらでも大きくできるため、以下の条件を付けることにする。

$$\sum_{i=1}^{m}x_{i\alpha}^2=1 \tag{5}$$

制約条件を付けたラグランジュの未定定数法を用いて、Q の式を以下のように変更する。

$$L=\sum_{\alpha=1}^{r}\mathbf{x}'_\alpha\mathbf{G}\mathbf{x}_\alpha-\sum_{\alpha=1}^{r}\lambda_\alpha(\mathbf{x}'_\alpha\mathbf{x}_\alpha-1) \tag{6}$$

これを $\mathbf{x}_\alpha$ で微分して以下の固有値方程式を得る。

$$\mathbf{G}\mathbf{x}_\alpha=\lambda_\alpha\mathbf{x}_\alpha \tag{7}$$

固有値方程式を成分で書き換えると以下のようになる。

$$\sum_{k=1}^{m}g_{ik}x_{k\alpha}=\lambda_\alpha x_{i\alpha} \tag{8}$$

これより以下となる。

$$\lambda_\alpha \sum_{i=1}^{m} x_{i\alpha} = \sum_{i=1}^{m}\sum_{k=1}^{m} g_{ik} x_{k\alpha} = 0 \tag{9}$$

ここで定義式によって成り立つ以下の関係を使った。

$$\sum_{j=1}^{m} g_{ij} = 0 \tag{10}$$

(9)式より $\lambda_\alpha \neq 0$ の場合、以下となる

$$\sum_{i=1}^{m} x_{i\alpha} = 0 \tag{11}$$

また、(10)式が成り立つことから方程式の 1 つの解として

$$\lambda_\alpha = 0, \quad x_{i\alpha} = 1/\sqrt{m} \tag{12}$$

を持つことも分かる。この場合(9)式の関係から、(11)式は成り立たなくてもよい。

最後に、方程式(8)を用いると、(4)の定義と(5)の制約より以下となる。

$$Q = \sum_{\alpha=1}^{r}\sum_{i=1}^{m} \lambda_\alpha x_{i\alpha}^2 = \sum_{\alpha=1}^{r} \lambda_\alpha \tag{13}$$

親近性 e_{ij} の線形変換に対する固有値と固有ベクトルの変化を調べてみる。

$$e'_{ij} = a e_{ij} + b \tag{14}$$

の変換に対して、

$$h'_{ij} = a h_{ij} + 2b$$

$$g'_{ij} = a g_{ij} - 2b(m\delta_{ij} - 1), \qquad \sum_{j=1}^{m} g'_{ij} = 0 \tag{15}$$

これにより、固有方程式は以下となる。

$$\sum_{k=1}^{m} (a g_{ik} + 2b) y_{k\alpha} = (\lambda'_\alpha + 2mb) y_{i\alpha} \tag{16}$$

これは $y_{k\alpha} = x_{k\alpha}$ とすると以下の関係を得る。

$$\lambda'_\alpha = a\lambda_\alpha - 2mb \qquad \text{for } \lambda_\alpha \neq 0, \sum_{i=1}^{m} x_{i\alpha} = 0$$

$$\lambda'_\alpha = 0 \qquad \text{for } \lambda_\alpha = 0, \; x_{i\alpha} = const. \tag{17}$$

即ち、固有値も線形の変換を受ける。これより、0 でない固有値の分布の間隔比

$$\gamma(\alpha) = \frac{\lambda_{\max} - \lambda_\alpha}{\lambda_{\max} - \lambda_{\min}} \tag{18}$$

は変換(14)に対して不変である。これにより、データに固有の親近性の特徴を調べることができると

考えられる。最後に、数量化の適合度の１つの指標として、距離 $-e_{ij}$ と(1)で与えられる r 次元の距離 $d_{ij}^{(r)}$ との順位相関係数を考えることもある。しかし、これは次元数を増やせば必ず適合度が上がるとは限らず、注意が必要である。

17.2　プログラムの利用法

数量化Ⅳ類のデータは、数間の親近性（類似度）または距離（非類似度）を表すデータである。その例を図１に示す。

図１　距離を表すデータ

メニュー［分析－多変量解析－数量化理論－数量化Ⅳ類］を選択すると図２のような数量化Ⅳ類分析メニューが表示される。

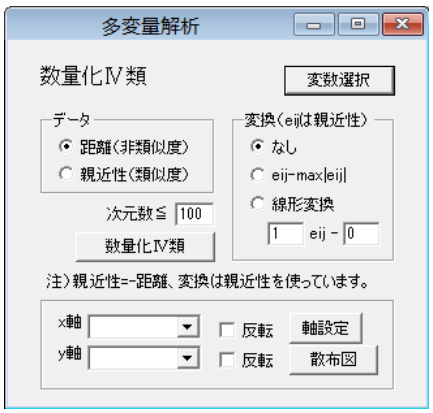


図２　数量化Ⅳ類分析メニュー

変数選択ですべての変数を選択し、データによって「距離」か「親近性」を選択する。距離の場合はデータの符号を変えて親近性にして分析を進める。変数の変換が必要な場合は変換ラジオボタンで指定する。特に「eij-max|eij|」は固有値をすべて正にするための設定であり、「線形変換」は他の多次元尺度構成法と合わせるための設定である。「次元数」大きな値を設定しておけば、変数数-1 の値になる。もちろん見やすくするため小さな値に設定することもできる。

「数量化Ⅳ類」ボタンをクリックすると、図３のような実行結果が示される。

座標

	1次元	2次元	3次元	4次元	5次元
固有値	39.788	36.280	33.808	33.202	32.522
みかん	0.111	−0.160	0.564	−0.070	−0.687
りんご	0.242	−0.275	0.130	0.734	0.378
いちご	0.218	0.539	−0.585	0.185	−0.346
ぶどう	0.158	0.495	0.372	−0.418	0.501
なし	0.179	−0.603	−0.426	−0.494	0.111
メロン	−0.908	0.004	−0.056	0.062	0.043
間隔比	0.000	0.483	0.823	0.907	1.000
ed順位相関	0.570	0.953	0.770	0.511	

図 3　分析結果

固有値、固有ベクトルが表示され、その下に固有値の間隔比と親近性と予測距離との順位相関が表示される。

「軸設定」をして、「散布図」ボタンをクリックすると、パラメータ（固有ベクトル）の値が散布図として図 4 のように表示される。軸の向きは「反転」チェックボックスによって変更できる。

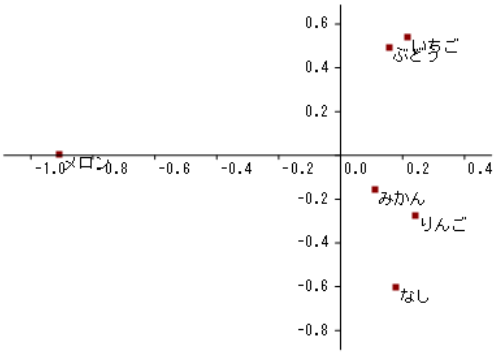


図 4　パラメータ散布図

参考文献

[1] 齋藤堯幸・宿久洋，関連性データの解析法，共立出版，2006

１８．パネル重回帰分析

18.1　パネル重回帰分析の理論

変数i（$i=1,\cdots,p$）、時刻t（$t=T,\cdots,0$）の時系列データ$x_{i,t}$があるとき、その中から時刻tを含めてr期分のそれ以前のデータを取り出す。それらのデータを説明変数とし、時刻$t+a$（$a\geq 1$）のある変数dのデータ$x_{d,t+a}$を目的変数として予測する重回帰分析をパネル重回帰分析という。これはa期先の予測である。

予測値を$X_{d,t+a}$とすると予測式は以下のように与えられる。

$$X_{d,t+a}=\sum_{i=1}^{p}\sum_{j=0}^{r-1}b_{i,j}x_{i,t-j}+b_0 \tag{1}$$

係数$b_{i,j}, b_0$は以下の量Lを最小化することによって求める。

$$L=\sum_{t=a+r-1}^{T}\left(x_{d,t}-\sum_{i=1}^{p}\sum_{j=0}^{r-1}b_{i,j+1}x_{i,t-a-j}-b_0\right)^2 \tag{2}$$

今、目的変数と説明変数をそれぞれ以下のように定義し、

$y_\lambda=x_{d,\lambda+a+r-2}$（$\lambda=1,\cdots,T-a-r+2$）

$z_{\alpha,\lambda}=z_{i+pj,\lambda}=x_{i+pj,\lambda+r-2-j}$　　（$i=1,\cdots,p,\ j=0,\cdots,r-1,\alpha=1,\cdots,pr$）

係数をb_αにして(2)式を書き変えると、以下のような式になる。

$$L=\sum_{\lambda=1}^{T-a-r+2}\left(y_\lambda-\sum_{\alpha=1}^{pr}b_\alpha z_{\alpha,\lambda}-b_0\right)^2 \tag{3}$$

これから、偏回帰係数 $\mathbf{b}={}^t\begin{pmatrix}b_0 & b_1 & b_2 & \cdots & b_s\end{pmatrix}$, $s=pr$ は以下のように求めることができる。

$$\mathbf{b}=\left({}^t\mathbf{\Omega\Omega}\right)^{-1}{}^t\mathbf{\Omega y} \tag{4}$$

ここに、

$$\mathbf{y}={}^t\begin{pmatrix}y_1 & y_2 & \cdots & y_N\end{pmatrix},\quad N=T-a-r+2$$

$$\mathbf{\Omega}=\begin{pmatrix}1 & z_{11} & z_{21} & \cdots & z_{s1}\\ 1 & z_{12} & z_{22} & \cdots & z_{s2}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & z_{1N} & z_{2N} & \cdots & z_{sN}\end{pmatrix}$$

時系列分析ではデータが時間の経過とともに明らかになっていくので、現在のすべてのデータから求めたパラメータを使って、過去の各時間の予測を行うことはその時点のデータの影響を強く受け過ぎるという難点がある。そこで、過去の予測を行う際には、その時点までのデータから計算されたパ

ラメータを用いることとし、これによって実測値と予測値の相関を求めることにする。これは一種の交差検証になっている。プログラムにはこの交差検証を付け加えている。

パネル重回帰分析には、他の分析で予測した結果を組み込むことができる。そこで時系列分析の結果をデータとして組み込むことを考えてみた。時系列分析は、傾向変動と周期変動を分解するモデルを考える。データの不規則な大きな変動も考える必要があるので、傾向変動には自然に傾向を求めることができる局所回帰分析を採用した。そのためバンド幅によって局所的な回帰式に影響を与える範囲を限定することができる。また周期変動については、分解する周期（周波数）を複数指定できるようにしている。

18.2　プログラムの利用法

パネル重回帰分析のデータは複数変数の時系列データである。その例を図 1 に示す。

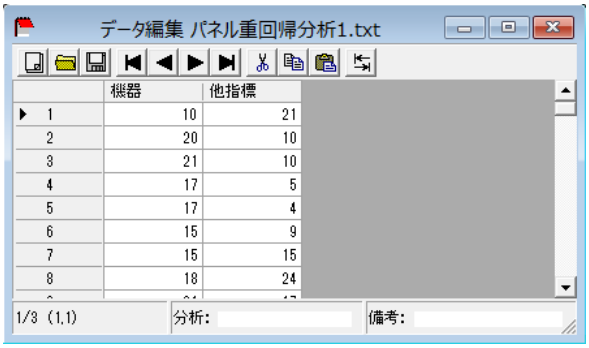


図 1　パネル重回帰分析のデータ

メニュー［分析－多変量解析他－予測手法－パネル重回帰分析］を選択すると図 2 のようなパネル重回帰分析実行メニューが表示される。

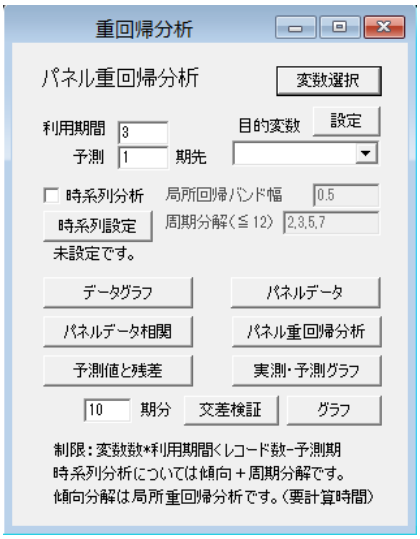


図 2　分析実行メニュー

使用するデータをすべて「変数選択」ボタンで選ぶが、変数間の時間的な影響を調べるツールとして使うことも考えているため、通常の重回帰分析のように目的変数を最初に選択することはしない。目的変数は、変数選択した候補をコンボボックスに読み込んだ後で、その中から「設定ボタン」で選択する。選択肢の中には単独の変数の他に「すべて」というものがあり、選択したすべての変数を目的変数にして、素早く結果を求めるときに利用する。ボタンによってはこれが使えないものもある。

この分析では、何期分のデータを利用するか、何期先の予測をするかを設定することができる。それに応じて、「パネルデータ」ボタンでは時系列データを通常の重回帰分析の形式に変形して出力する。出力結果をそのまま重回帰分析のデータとして利用することもできる。変数「機器」を目的変数とし、3 期分のデータを利用し、1 期先の予測をする場合の出力データを図 3 に示す。

	機器	機器_1	他指標_1	機器_2	他指標_2	機器_3	他指標_3
4	17	21	10	20	10	10	21
5	17	17	5	21	10	20	10
6	15	17	4	17	5	21	10
7	15	15	9	17	4	17	5
8	18	15	15	15	9	17	4
9	21	18	24	15	15	15	9
10	26	21	17	18	24	15	15
11	23	26	4	21	17	18	24

図 3　計算用データ

この中で「機器」は目的変数で、左に月単位で与えられているデータとする。また、例えば「機器_2」は変数「機器」の 2 期前のデータを表している。

図 3 の計算用データの各変数間の相関係数は、「パネルデータ相関」ボタンをクリックすることで図 4 のように与えられる。

パネルデータ相関係数

	機器	機器_1	他指標_1	機器_2	他指標_2	機器_3	他指標_3
▶ 機器	1.000	0.824	−0.052	0.833	0.105	0.834	−0.146
機器_1	0.824	1.000	−0.095	0.817	−0.054	0.825	0.102
他指標_1	−0.052	−0.095	1.000	−0.093	−0.056	−0.019	−0.137
機器_2	0.833	0.817	−0.093	1.000	−0.099	0.807	−0.060
他指標_2	0.105	−0.054	−0.056	−0.099	1.000	−0.094	−0.060
機器_3	0.834	0.825	−0.019	0.807	−0.094	1.000	−0.113
他指標_3	−0.146	0.102	−0.137	−0.060	−0.060	−0.113	1.000

図 4　パネルデータ相関出力結果

このデータを使った重回帰分析の詳細は、「パネル重回帰分析」ボタンで図 5 のように与えられる。

パネル重回帰分析

	偏回帰係数	標準化係数	t検定値	自由度	確率値
▶ 機器_1	0.3258	0.3200	3.4654	90	0.0008
他指標_1	0.0243	0.0112	0.2531	90	0.8008
機器_2	0.3579	0.3433	4.1283	90	0.0001
他指標_2	0.3891	0.1784	4.0746	90	0.0001
機器_3	0.3166	0.2978	3.3981	90	0.0010
他指標_3	−0.2432	−0.1121	−2.3768	90	0.0196
切片	−1.2488	0.0000	−0.3919	90	0.6960
重相関・寄与率	0.912	0.833			

図 5　目的変数を「機器」とした場合のパネル重回帰分析結果

目的変数を「すべて」に設定すると、「パネル重回帰分析」ボタンで図 6 のような結果になる。

パネル重回帰分析（すべて）

	機器:偏回帰:	標準化	確率値	他指標:偏回	標準化	確率値
▶ 機器_1	0.3258	0.3200	0.0008	−0.1127	−0.2414	0.2784
他指標_1	0.0243	0.0112	0.8008	−0.0719	−0.0718	0.4983
機器_2	0.3579	0.3433	0.0001	0.0622	0.1300	0.5159
他指標_2	0.3891	0.1784	0.0001	−0.1324	−0.1324	0.2105
機器_3	0.3166	0.2978	0.0010	0.0234	0.0480	0.8198
他指標_3	−0.2432	−0.1121	0.0196	0.0167	0.0168	0.8826
切片	−1.2488	0.0000	0.6960	16.8657	0.0000	0.0000
重相関・寄与率	0.912	0.833		0.195	0.038	

図 6　目的変数をすべてとした場合のパネル重回帰分析結果

これは各変数を目的変数にして、偏回帰係数、標準化偏回帰係数、確率値、重相関係数、寄与率を出力している。どの変数の何期前のデータが重要であるか、標準化係数や確率値を見ることで知ることができる。

目的変数を「機器」とした場合の実測値、予測値、残差は、「予測値と残差」ボタンをクリックすることで図 7 のように求められる。ここで一番下の予測値は、1 期先（設定で変更可能）の予測値で、実測値はまだない。

予測値と残差

	機器	予測値	残差
94	57.000	56.017	0.983
95	62.000	64.020	-2.020
96	78.000	59.132	18.868
97	61.000	66.715	-5.715
98	70.000	73.932	-3.932
99	69.000	70.957	-1.957
100	66.000	65.528	0.472
1期先		70.471	

図 7　目的変数を「機器」とした場合の予測値と残差結果

また、目的変数を「すべて」とした場合の実測値、予測値、残差は、同様にして図 8 のように求められる。

パネル重回帰分析

	機器	予測値	残差	他指標	予測値	残差
94	57.000	56.017	0.983	4.000	13.589	-9.589
95	62.000	64.020	-2.020	7.000	13.965	-6.965
96	78.000	59.132	18.868	23.000	14.173	8.827
97	61.000	66.715	-5.715	21.000	10.749	10.251
98	70.000	73.932	-3.932	12.000	11.852	0.148
99	69.000	70.957	-1.957	15.000	11.333	3.667
100	66.000	65.528	0.472	18.000	12.551	5.449
		70.471			12.274	

図 8　目的変数をすべてとした場合の予測値と残差結果

実測値と予測値について、結果をグラフで表示するためには、「実測・予測グラフ」ボタンをクリックする。実行結果は図 9 に示す。

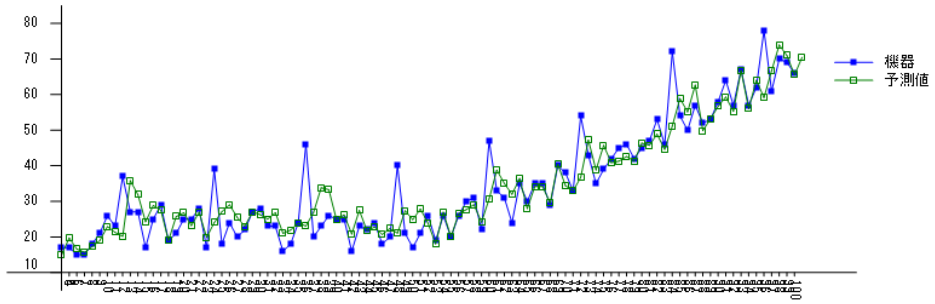


図 9　実測値と予測値グラフ

ここにデータの名前（年月）は縦表示にしてある。

我々がこれまで求めてきた各時点の予測値は、全体の結果を使って求めた係数から計算して得られた値である。それゆえ、この係数には各時点の実測値の結果が含まれている。そのためこれらのデータは厳密には予測値ではない。これを補正するためには、予測値は各時点のそれより過去のデータから求めるべきであろう。この考え方は交差検証の考え方に通じる。「期分」のテキストボックスに予測したい期間の数値を入れ、「交差検証」ボタンをクリックすると、過去のデータからだけで作られた予測値と残差が図 10 のように表示される。但し、表示期間を 50 期分にしている。

	機器	予測値	残差
94	57.000	54.447	2.553
95	62.000	63.934	-1.934
96	78.000	57.643	20.357
97	61.000	68.532	-7.532
98	70.000	74.934	-4.934
99	69.000	71.341	-2.341
100	66.000	65.481	0.519
R･R^2	0.907	0.824	

図 10　目的変数を「機器」とした場合の 50 期分の交差検証結果

目的変数をすべてにして同様の結果を得ることもできる。「グラフ」ボタンをクリックすると、図 10 の結果をグラフ化することができる。結果を図 11 に示す。

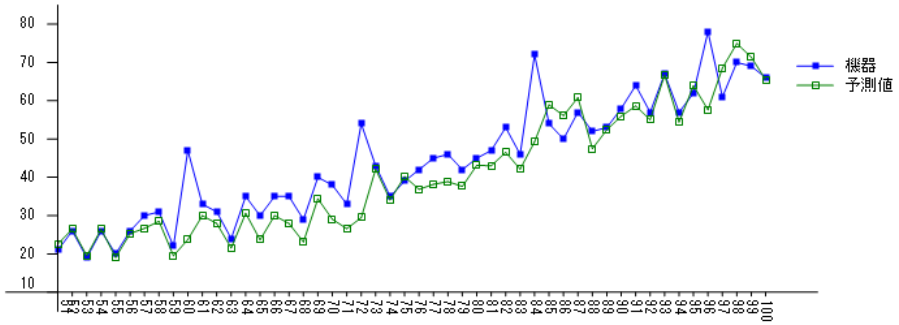


図 11　交差検証での実測値と予測値

純粋なパネル重回帰分析の結果は以上であるが、我々はさらに予測精度を上げるために、傾向変動や周期変動の分解を考える従来の時系列分析の予測値をパネルデータに加え、2 つの分析の良い部分を組み合わせることにした。ここで、傾向変動には局所回帰分析を用いている。図 2 の分析実行メニューの時系列分析チェックボックスにチェックを入れると、「局所回帰バンド幅」と「周期分解（≦12)」のテキストボックスが利用できるようになる。バンド幅の値はデフォルトでほぼ良い結果が得られるが、例えば 12 ヶ月周期が明らかな場合には、周期分解に 12 を含める。周期分解のためのデータ数は最低でも最大周期の 2 倍必要なので、周期は適当に小さくという意味で「(≦12)」の指摘を加えてある。しかし、この範囲に縛られる必要はない。ここでは 12 を加えている。

時系列分析を加えた場合、データの数によっては計算時間がかかる場合があるので、最初に「時系列設定」のボタンをクリックする。「計算が終わりました。」の表示が出たら、以後はすぐに表示される。「パネルデータ」ボタンをクリックすると、図 12 のように最後の列に時系列分析の予測値が追加される。但し、計算が可能な途中からの挿入となる。プログラムはこの部分を利用して計算をする。

パネルデータ

	機器	機器_1	他指標_1	機器_2	他指標_2	機器_3	他指標_3	機器_ts
22	28	25	1	25	22	21	17	0.000000
23	17	28	5	25	1	25	22	0.000000
24	39	17	4	28	5	25	1	0.000000
25	18	39	15	17	4	28	5	31.967064
26	24	18	9	39	15	17	4	24.869150
27	20	24	13	18	9	39	15	25.120598
28	22	20	23	24	13	18	9	18.966236
29	27	22	25	20	23	24	13	18.279529
30	28	27	16	22	25	20	23	21.529595
31	23	28	15	27	16	22	25	26.493811

図 12　時系列分析を加えた計算用データ

重回帰分析では、変数の数が増えると寄与率の値は増加するので、前以上の結果は期待できるが、増加の程度は、元のデータの性質による。例えば周期性が強いデータならば、時系列分析の変数の効果が強く効いてくる。

これ以降の分析は時系列分析を含めない場合と同様であるので、図 13 と図 14 に交差検証の結果のみを示しておく。データがそろってきた最後の方の数値はよく合っている。

交差検証

	機器	予測値	残差
94	57.000	58.364	-1.364
95	62.000	58.160	3.840
96	78.000	75.271	2.729
97	61.000	64.673	-3.673
98	70.000	71.280	-1.280
99	69.000	72.271	-3.271
100	66.000	67.021	-1.021
R･R^2	0.947	0.896	

図 13　時系列分析を加えた交差検証結果

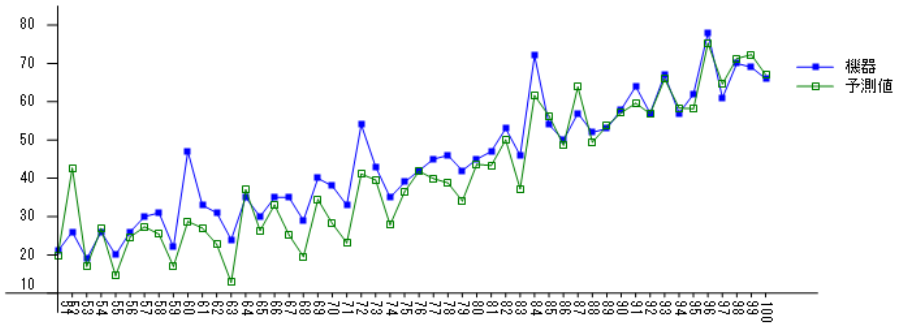


図 14　時系列分析を加えた交差検証での実測値と予測値

１９．メタ分析

19.1　メタ分析の手法

メタ分析は、多くの研究資料から同一の調査内容を選び出し、それらを再度集計して結果をより強固なものにしようとする分析手法である。１つの研究資料からは、効果量と呼ばれる統計量とその分散及び、データ数を取り出す。代表的な効果量には標準化された平均値差、オッズ比、相関係数などがある。しかし、研究資料ごとにこれらが同じである保証はないので、必要があれば、これらを統一的な効果量に変換する。その後、各研究資料にデータ数でウェイトをかけて、研究で与えられた結果が保証されるかどうか検討する。この一連の手法をメタ分析という。

我々はこの一連の過程を計算するプログラムの開発を考えた。ここでは、参考文献[1] に従い、効果量の入力、効果量の変換、統計的分析に分けて、理論的にどのような式が使われているのかをまとめて紹介する。

19.1.1　効果量とその入力

我々がプログラムの中で扱う効果量は以下で述べる通りである。種々の資料には効果量（または検定確率）とデータ数は記載されているが、効果量の分散が記載されていないことが多い。また、参考文献[1] では、後の統計的分析のために分散は記載されているが、データ数が記載されていない。これらの状況に対処するために、我々は結果表示に必要なデータは何か、またそれを得るためにはどのようなデータが必要かを検討した。結論は、比較的良い近似として、結果表示に必要なデータは、効果量と全データ数または、効果量と分散であった。ここでは、効果量と、全データ数または分散のどちらかが分かっているものとして、他方を求める近似式を与えておく。但しこの結果には $y=1/x(1-x)$ のグラフの性質を利用している。

１）標準化平均値差 d （ヘッジスの g とも呼ばれる）

対応のない場合

効果量　$d=\dfrac{\bar{x}_1-\bar{x}_2}{u_{pooled}}$,　$u_{pooled}=\sqrt{\dfrac{(n_1-1)u_1^2+(n_2-1)u_2^2}{n_1+n_2-2}}$　（pooled 標準偏差）

分散　$V_d=\dfrac{n_1+n_2}{n_1n_2}+\dfrac{d^2}{2(n_1+n_2)}$

全データ数　$N=n_1+n_2$

$d, N \to V_d$ のとき、$V_d \simeq \dfrac{4\alpha+d^2/2}{N}$

$d, V_d \to N$ のとき、$N \simeq \dfrac{4\alpha+d^2/2}{V_d}$

ここで、分散の $(n_1+n_2)/n_1n_2 = N/n_1(N-n_1)$ の項については、例えば、$N=100$ とすると、図１のようなグラフとなる。

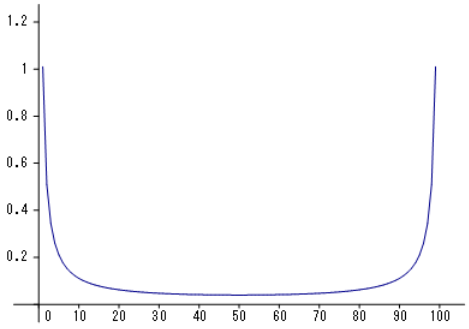


図１　$y=100/x(100-x)$

このグラフは、中央部で $4/N$ に近いほぼ安定な値を取っており、この項による変動は少ないと考えられる。そこで我々は、この関数の $x=N/2$ の値を中心とした正規分布による加重平均を考え、その結果を $(n_1+n_2)/n_1n_2 \simeq 4\alpha/N$ とした。

α の値については、以下のように計算した。

$$\alpha = \frac{1}{4A}\int_{0.1}^{0.9} \frac{1}{x(1-x)} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-0.5)^2}{2\sigma^2}\right] dx$$

$$A = \frac{1}{\sqrt{2\pi}\sigma}\int_{0.1}^{0.9} \exp\left[-\frac{(x-0.5)^2}{2\sigma^2}\right] dx$$

この場合、例えば、$\sigma=0.2$ とすると、$\alpha=1.187$ となる。我々はこの値を利用する。

標準化平均値差の代わりに、資料で t 統計量が使われている場合は、簡単に標準化平均値差に変換することができる。

$$t = \sqrt{\frac{n_1n_2}{n_1+n_2}} \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{(n_1-1)u_1^2+(n_2-1)u_2^2}{n_1+n_2-2}}} = \sqrt{\frac{n_1n_2}{n_1+n_2}}d \simeq \sqrt{\frac{N}{4\alpha}}d$$

$$V_t \simeq \frac{NV_d}{4\alpha} \simeq \frac{4\alpha+d^2/2}{4\alpha} = 1+\frac{d^2}{8\alpha}$$

これを利用すると、以下の変換も可能になる。

$t, N \to V_t$ のとき、$V_t \simeq 1+\dfrac{t^2}{2N}$

$t, V_t \to N$ のとき、$N \simeq \dfrac{t^2}{2(V_t-1)}$

2）バイアス修正平準化平均値差 g

効果量　$g = J \times$ ，　$J = 1 - \dfrac{3}{4(n_1 + n_2 - 2) - 1}$

分散　$V_g = J^2 \times V_d$

全データ数　$N = n_1 + n_2$

$N \to V_g$ のとき、$J = 1 - \dfrac{3}{4(N-2)-1}$ として、$V_g = J^2 V_d \simeq \dfrac{4\alpha J^2 + g^2/2}{N}$

$V_g \to N$ のとき、$J \simeq 1$ であると考え、$N \simeq \dfrac{4\alpha + g^2/2}{V_g}$

3）対数オッズ比

以下の 2 次元分割表を考える。

	効果あり	効果なし	合計
介入群	a	b	$a+b$
統制群	c	d	$c+d$

効果量　$LOR = \ln\left(\dfrac{a/b}{c/d}\right) = \ln\left(\dfrac{a}{b}\right)$

分散　$V_{LOR} = \dfrac{1}{a} + \dfrac{1}{b} + \dfrac{1}{c} +$

全データ数　$N = a+b+c+d = n_1 + n_2$

$N \to V_{LOR}$ のとき、$V_{LOR} \simeq \dfrac{16\alpha^2}{N}$　$V_{LOR} = \dfrac{1}{a} + \dfrac{1}{b} + \dfrac{1}{c} + \dfrac{1}{d} \simeq \dfrac{4\alpha}{n_1} + \dfrac{4\alpha}{n_2} \simeq \dfrac{16\alpha^2}{N}$

$V_{LOR} \to N$ のとき、$N \simeq \dfrac{16\alpha^2}{V_{LOR}}$

効果量の代わりに、資料で χ^2 統計量が使われていた場合は、簡単に効果量に変換することができない。

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

分割表の度数から効果量を計算する必要がある。

4）相関係数

効果量　$r = \dfrac{s_{xy}}{s_x s_y}$　　分散　$V_r = \dfrac{(1-r^2)^2}{n-1}$

データ数　$N = n$

$N \to V_r$ のとき、$V_r = \dfrac{(1-r^2)^2}{N-1}$

$V_g \to N$ のとき、$N = \dfrac{(1-r^2)^2}{V_r} + 1$

19.1.2 効果量の変換

効果量は相互に変換可能である。ここではプログラムで用いられる変換について式を与える。

$d \leftrightarrow g$

効果量：$g = J \times d$　　分散：$V_g = J^2 \times V_d$

ここに、$J = 1 - \dfrac{3}{4(N-2)-1}$　（入力の際に N は設定済みとする）

$LOR \leftrightarrow d$

効果量：$d = LOR \times \dfrac{\sqrt{3}}{\pi}$　　分散：$V_d = V_{LOR} \times \dfrac{3}{\pi^2}$

$r \to d$

効果量：$d = \dfrac{2r}{\sqrt{1-r^2}}$　　分散：$V_d = \dfrac{4V_r}{(1-r^2)^3}$

$d \to r$

効果量：$r = \dfrac{d}{\sqrt{d^2 + a}}$　　分散：$V_r = \dfrac{a^2 V_d}{(d^2+a)^3}$

ここに、$a = \dfrac{(n_1+n_2)^2}{n_1 n_2} \simeq 4\alpha$

19.1.3 統計的分析

1) 固定効果モデル

固定効果モデルでは、研究間の差はなく、研究 i の効果量 d_i は独立に $d_i \sim N(0, V_i)$ に従うと仮定し、以下の集計を考える。

$$d = \frac{\sum_{i=1}^{n} d_i / V_i}{\sum_{i=1}^{n} 1/V_i} \sim N\left(0, \frac{\sum_{i=1}^{n} V_i / V_i^2}{\left(\sum_{i=1}^{n} 1/V_i\right)^2}\right) = N\left(0, 1 \Big/ \sum_{i=1}^{n} 1/V_i\right)$$

ここで、$w_i = 1/V_i$ として、これをウェイトと考え、$w = \sum_{i=1}^{n} w_i$ とすると、以下となる。

$$d_i \sim N(0, 1/w_i), \quad d = \sum_{i=1}^{n} w_i d_i \Big/ w \sim N\left(0, 1/w\right)$$

この性質より、研究を結合した検定は、検定統計量 $z = d/\sqrt{1/w} \sim N(0,1)$ を使って行う。

2) 変量効果モデル

変量効果モデルでは、研究間に差があり、研究 i の効果量 d_i は広く拡がり、$d_i \sim N(0, V_i + \sigma^2)$ に従うと考える。$w_i' = 1/(V_i + \sigma^2)$ とおくと、$d_i \sim N(0, 1/w_i')$ より、$\sqrt{w_i'} d_i \sim N(0,1)$ となり、以下を得る。

$$Q' = \sum_{i=1}^{n} w_i'(d_i - d)^2 = \sum_{i=1}^{n} \frac{(d_i - d)^2}{V_i + \sigma^2} \sim \chi^2_{n-1}$$

一方、

$$Q = \sum_{i=1}^{n} w_i (d_i - d)^2 = \sum_{i=1}^{n} \frac{(d_i - d)^2}{V_i}$$

は元の分散で測った量である。その差は、以下で与えられる。

$$Q - Q' = \sum_{i=1}^{n} \frac{(d_i - d)^2 \sigma^2}{V_i (V_i + \sigma^2)} = \sum_{i=1}^{n} (d_i - d)^2 w_i' w_i \sigma^2 = C\sigma^2$$

ここに、Q' と C には、期待値を使って、

$$E(Q') = n - 1$$

また、$C = E[\sum_{i=1}^{n} (d_i - d)^2 w_i' w_i]$ は、

$$E[(d_i - d)^2] = E[d_i^2] - 2E[d_i d] + E[d^2]$$

$$E[d_i^2] = V_i' = 1/w_i'$$

$$E[d_i d] = E[d_i \sum_{j=1}^{n} d_j w_j' / w'] = V_i' w_i' / w' = 1/w'$$

$$E[d^2] = E[\sum_{i=1}^{n} d_i w_i' / w' \sum_{j=1}^{n} d_j w_j' / w'] = \sum_{i=1}^{n} V_i w_i'^2 / w'^2 = \sum_{i=1}^{n} w_i' / w'^2 = 1/w'$$

より、

$$C = \sum_{i=1}^{n} (1/w_i' - 1/w') w_i' w_i = \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} w_i' w_i / w' \simeq \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} w_i^2 / w$$

これらより、σ^2 が以下のように求められる。

$$\sigma^2 = \frac{Q-(n-1)}{C}$$

以後、ウェイトとして、$w_i' = \dfrac{1}{V_i + \sigma^2}$，$w' = \sum_{i=1}^{n} w_i'$ を用いて、計算を行えばよい。即ち、研究を結合した検定は、検定統計量 $z' = d \Big/ \sqrt{1/w'} \sim N(0,1)$ を使って行う。

19.1.4　研究群間の比較

何らかの指標の違いにより、研究が k 個のグループに分けられるとする。各グループの研究の数を n_i，全体の研究の数を n とするとき、そのグループ間の効果量の差を検定するには、以下の性質を用いる。

$$Q_{Total} \sim \chi^2_{n-1}\ ,\ Q_i \sim \chi^2_{n_i-1},\ n = \sum_{i=1}^{k} n_i \quad \text{より、}$$

$$Q_{Total} - \sum_{i=1}^{k} Q_i \sim \chi^2_{df}\ ,\ df = (n-1) - \sum_{i=1}^{k} (n_i - 1) = k-1$$

この計算には、固定効果モデルではウェイト $w_i = 1/V_i$ を用い、変量効果モデルではウェイト $w_i' = 1/V_i'$ を用いる。

19.2　プログラムの利用法

メニュー［分析－多変量解析等－メタ分析］を選択すると、メタ分析の分析実行メニューが図 2 のように表示される。

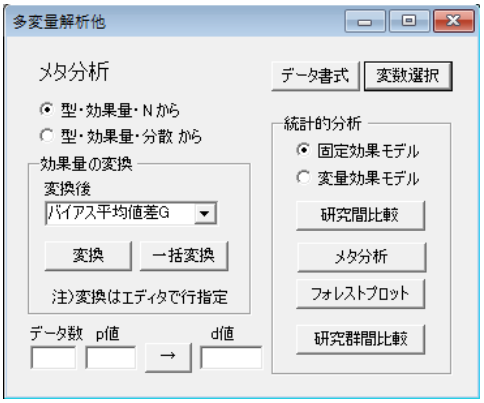


図 2　分析実行メニュー

ここでは、参考文献 1 で与えられた図 3 のデータを元にプログラムの利用法を説明する。

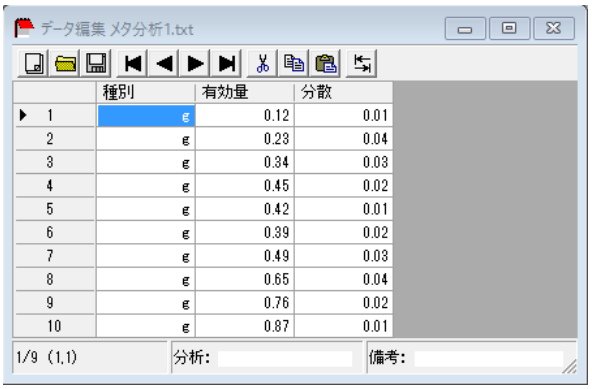


図 3　メタ分析データ

このデータではデータ型、有効量、分散が用いられているので、分析実行メニューのデータ型を、「型・効果量・分散」に変えて、分析を実行する。しかし、一般には分散が与えられる場合は少なく、むしろ、データ型、有効量、データ数が与えられることが多いと思われる。その場合には、分析実行メニューのデータ型を、「型・効果量・N」にして実行する。データ型には、G：バイアス修正平準化平均値差、D：標準化平均値差、LOR：対数オッズ比、R：相関係数が指定できる。なお、指定する文字は大文字でも小文字でも同じである。

また、2 群の差の検定などでは、検定統計量を省略し、検定確率だけを表示している場合もあるので、その際には、標準化平均値差 D の値を簡易的に計算できる機能をメニューの下に設けている。その他の対数オッズ比や相関係数では、殆どの場合、値を記述するので、ここでは標準化平均値差 D に限定している。また、ノンパラメトリック検定の確率から近似的に D を求めても、少し乱暴ではあるが、経験上特に大きな差は出ないように思う。

一般に各研究では効果量が同一とは限らない。異なる効果量の場合は、効果量の変換を行い、同じ効果量に合わせて分析する。そのためにプログラムには効果量の変換機能を付けている。変数選択で 3 つの変数を選択し、「変換後」コンボボックスで変換先の型を選び、「一括返還」ボタンをクリックすると、図 4 のような結果が得られる。ここでは、相関係数として出力している。

図 4　効果量の相関係数への変換

グリッドの一部分のデータについて変換をしたい場合は、種別・効果量・分散の必要な行を連続的

に選択して「変換」ボタンをクリックする。出力結果は省略する。

すべての研究結果を統合して検定を行いたい場合、研究間の効果量の値にばらつきがあるかどうか知らなければならない。それを調べる場合は、「研究間比較」ボタンをクリックする。結果を図 5 に示す。

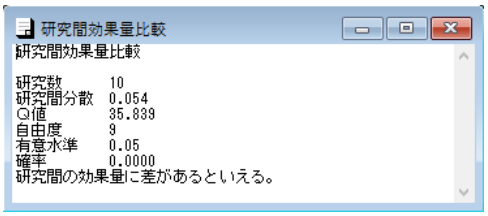


図 5　研究間効果量の差の比較検定

この結果から、研究間の効果量に差が見られたので、分析には「変量効果モデル」を用いる。これは「固定効果モデル」に比べて差が検出しにくい検定である。

変量効果モデルを用いた最終的な分析結果を得るには、「変量効果モデル」ラジオボタンを選択し、「メタ分析」ボタンをクリックする。結果を図 6 に示す。

フォレストデータ

	種別	効果量	N	分散	標準誤差	p値	2.5%下限	2.5%上限
▶ 1	G	0.1200	476	0.0100	0.1000	0.2301	-0.0760	0.3160
2	G	0.2300	119	0.0400	0.2000	0.2501	-0.1620	0.6220
3	G	0.3400	160	0.0300	0.1732	0.0496	0.0005	0.6795
4	G	0.4500	242	0.0200	0.1414	0.0015	0.1728	0.7272
5	G	0.4200	484	0.0100	0.1000	0.0000	0.2240	0.6160
6	G	0.3900	241	0.0200	0.1414	0.0058	0.1128	0.6672
7	G	0.4900	162	0.0300	0.1732	0.0047	0.1505	0.8295
8	G	0.6500	124	0.0400	0.2000	0.0012	0.2580	1.0420
9	G	0.7600	252	0.0200	0.1414	0.0000	0.4828	1.0372
10	G	0.8700	513	0.0100	0.1000	0.0000	0.6740	1.0660
結合		0.4747	2773	0.0076	0.0870	0.0000	0.3041	0.6453

図 6　変量効果モデルを用いた分析結果

各研究の結果がまとめて表示され、一番下の行に結合された結果が表示されている。

さらに、この結果を分かり易く表す図がフォレストプロットである。「フォレストプロット」ボタンをクリックすると、図 7 のような結果が表示される。

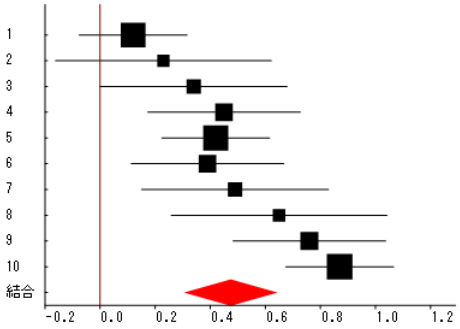


図 7　フォレストプロット

一番下のひし形が、0 をまたいでいないことから、この結果では、有意に差があるといえるということになる。

次に、研究がいくつかの特徴に分かれ、その研究群間に差があるかどうか調べてみたいと考えたとする。その際には、先頭列に分類変数を加えた図 8 のようなデータを用いる。

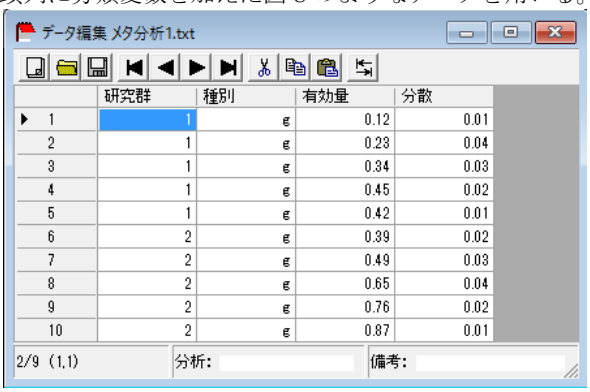


図 8　2 つの研究群による比較データ

すべてのデータを並んだ順に選択し、分析実行メニューの「研究群間比較」ボタンをクリックすると、図 9 のような結果が得られる。

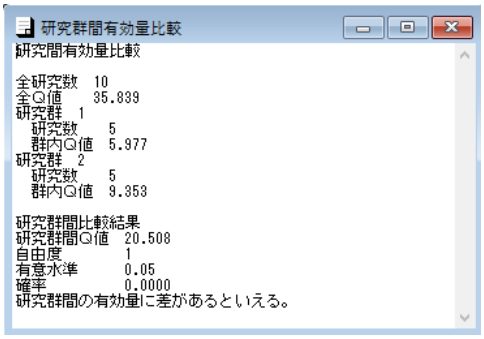


図 9　研究群間の比較結果

参考文献

[1] 山田剛史，井上俊哉編，メタ分析入門 心理・教育研究の系統的レビューのために，東京大学出版会，2012.

２０．２値ロジスティック回帰

20.1　一般化線形モデルの理論

1)　指数分布族

最初に、参考文献[1]に従って理論を整理しておく。

ある単一のパラメータ θ を持つ確率変数 Y が以下の確率密度関数に従うとき、その分布を指数型分布族という。

$$f(y;\theta)=\exp[a(y)b(\theta)+c(\theta)+d(y)]$$

指数型分布族には、ポアソン分布、正規分布、2 項分布等が含まれる。特に $a(y)=y$ のとき分布は正準形であると言われ、$b(\theta)$ は分布の自然パラメータと呼ばれる。

確率変数 $a(\mathrm{Y})$ については

$$\frac{d}{d\theta}\int f(y;\theta)dy=\int[a(y)b'(\theta)+c'(\theta)]f(y;\theta)dy=E[a(y)]b'(\theta)+c'(\theta)=0$$

より、

$$E[a(Y)]=-c'(\theta)/b'(\theta) \tag{1.1}$$

$$\frac{d^2}{d\theta^2}\int f(y;\theta)dy=\int[a(y)b''(\theta)+c''(\theta)]f(y;\theta)dy+\int[a(y)b'(\theta)+c'(\theta)]^2 f(y;\theta)dy$$

$$=E[a(Y)^2]b'(\theta)^2+E[a(Y)][b''(\theta)+2b'(\theta)c'(\theta)]+c''(\theta)+c'(\theta)^2$$

$$=E[a(Y)^2]b'(\theta)^2-\frac{c'(\theta)}{b'(\theta)}[b''(\theta)+2b'(\theta)c'(\theta)]+c''(\theta)+c'(\theta)^2$$

$$=E[a(Y)^2]b'(\theta)^2-E[a(Y)]^2b'(\theta)^2+\frac{1}{b'(\theta)}[-b''(\theta)c'(\theta)+c''(\theta)b'(\theta)]$$

$$=V[a(Y)]b'(\theta)^2+\frac{1}{b'(\theta)}[-b''(\theta)c'(\theta)+c''(\theta)b'(\theta)]=0$$

より、

$$V[a(Y)]=\frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{b'(\theta)^3} \tag{1.2}$$

という性質がある。

対数尤度関数 $l(y;\theta)=\log f(y;\theta)$ の θ に関する導関数の確率変数

$$U(Y;\theta)=a(Y)b'(\theta)+c'(\theta) \tag{1.3}$$

は、スコア統計量とも呼ばれ、その分布の期待値と分散は (1.1), (1.2), (1.3) 式を使うと以下となる。

$$E[U]=0 \tag{1.4}$$

$$V[U]=V[a(Y)]b'(\theta)^2=\frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{b'(\theta)} \tag{1.5}$$

さらに、

$$V[\mathrm{U}]=E[U^2]-E[U]^2=E[U^2]$$

$$E[U']=E[a(Y)]b''(\theta)+c''(\theta)=-\frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{b'(\theta)}=-V[U]$$

の関係より、以下も成り立つ。

$$V[U]=E[U^2]=-E[U'] \tag{1.6}$$

スコア統計量の分散$V[U]$は情報量とも呼ばれる。

2) 正準形の一般化線形モデル

正準形の指数分布族の分布に従う確率変数$Y_i\ (i=1,2,\cdots,N)$が、パラメータθ_iの同じ形の以下の独立な確率密度関数の分布に従うと考える。

$$f(y_i;\theta_i)=\exp[y_i b(\theta_i)+c(\theta_i)+d(y_i)] \tag{2.1}$$

確率密度関数の対数$l(y_i;\theta_i)$は以下で与えられる。

$$l(y_i;\theta_i)=y_i b(\theta_i)+c(\theta_i)+d(y_i) \tag{2.2}$$

確率変数Y_iの平均と分散は前節の議論より、以下のように与えられる。

$$E[Y_i]=-c'(\theta_i)/b'(\theta_i)\equiv\mu_i \tag{2.3}$$

$$V[Y_i]=\frac{b''(\theta_i)c'(\theta_i)-c''(\theta_i)b'(\theta_i)}{[b'(\theta_i)]^3} \tag{2.4}$$

ここで、θ_iはμ_iの関数であるとみることができる。

我々はこのμ_iに対して、ある説明変数${}^t\mathbf{x}_i=(1,x_{i1},\cdots,x_{ip})$（$i=1,\cdots,N$）とパラメータを用いて以下のような仮定をする。

$$\eta_i\equiv g(\mu_i)=\sum_{j=1}^{p}\beta_j x_{ij}+\beta_0={}^t\boldsymbol{\beta}\mathbf{x}_i \tag{2.5}$$

この仮定により、θ_iは$\boldsymbol{\beta}$の関数と見ることができる。またこの関係を与える関数$\eta_i=g(\mu_i)$を連結関数という。

確率変数Y_iの同時確率密度関数（尤度関数）は以下で与えられる。

$$f(\mathbf{y};\boldsymbol{\theta})=\exp\left[\sum_{i=1}^{N}\{y_i b(\theta_i)+c(\theta_i)+d(y_i)\}\right] \tag{2.6}$$

また対数尤度関数$l(\mathbf{y};\boldsymbol{\theta})$は以下のようになる。

$$l(\mathbf{y};\boldsymbol{\theta}) = \sum_{i=1}^{N} l_i = \sum_{i=1}^{N} [y_i b(\theta_i) + c(\theta_i) + d(y_i)] \tag{2.7}$$

この対数尤度関数の β_j による微分をスコアベクトルと呼び、U_j とすると、スコアベクトル U_j は以下のようになる。

$$U_j = \sum_{i=1}^{N} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial l_i}{\partial \theta_i} \tag{2.8}$$

ここで、

$$\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)\left[y_i + c'(\theta_i)/b'(\theta_i)\right]$$
$$= b'(\theta_i)(y_i - E[Y_i]) = b'(\theta_i)(y_i - \mu_i)$$
$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{b'(\theta_i)^2}{-c''(\theta_i)b'(\theta_i) + c'(\theta_i)b''(\theta_i)} = \frac{1}{b'(\theta_i)V[Y_i]}$$
$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \eta_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \eta_i} = x_{ij} \frac{\partial \mu_i}{\partial \eta_i}$$

となることから、以下の表式を得る。

$$U_j = \sum_{i=1}^{N} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{V[Y_i]} \frac{\partial \mu_i}{\partial \eta_i} \tag{2.9}$$

また、尤度が最大となるパラメータでは、以下も成り立つ。

$$E[U_j] = 0 \tag{2.10}$$

さらに U_j の β_k による微分を U_{jk} とすると、U_{jk} は以下のようになる。

$$U_{jk} \equiv \frac{\partial U_j}{\partial \beta_k} = \frac{\partial}{\partial \beta_k} \sum_{i=1}^{N} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta_k} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial}{\partial \theta_i} \left(\frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial l_i}{\partial \theta_i} \right)$$
$$= \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k} \left(\frac{\partial \theta_i}{\partial \mu_i} \right)^2 \frac{\partial}{\partial \theta_i} \left(\frac{\partial l_i}{\partial \theta_i} \right) + \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta_k} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial l_i}{\partial \theta_i} \frac{\partial}{\partial \theta_i} \left(\frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \theta_i}{\partial \mu_i} \right)$$
$$= \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \beta_k} \left(\frac{\partial \theta_i}{\partial \mu_i} \right)^2 [y_i b''(\theta_i) + c''(\theta_i)]$$
$$+ \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ik}}{V[Y_i]} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial}{\partial \theta_i} \left(\frac{\partial \mu_i}{\partial \beta_j} \frac{\partial \theta_i}{\partial \mu_i} \right)$$

また、

$$E[Y_i b''(\theta_i)+c''(\theta_i)]=E[Y_i]b''(\theta_i)+c''(\theta_i)=-c'(\theta_i)b''(\theta_i)/b'(\theta_i)+c''(\theta_i)$$
$$=\frac{-c'(\theta_i)b''(\theta_i)+b'(\theta_i)c''(\theta_i)}{b'(\theta_i)}=-b'(\theta_i)^2V[Y_i]$$

$$E\left[\sum_{i=1}^{N}\frac{(Y_i-\mu_i)x_{ik}}{V[Y_i]}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial}{\partial\theta_i}\left(\frac{\partial\mu_i}{\partial\beta_j}\frac{\partial\theta_i}{\partial\mu_i}\right)\right]=0$$

であることから、(2.9)式を求める際の計算により、U_{jk} の変数の値を確率変数で置き換えて計算すると以下となる。

$$E[U_{jk}]=-\sum_{i=1}^{N}\frac{\partial\mu_i}{\partial\beta_j}\frac{\partial\mu_i}{\partial\beta_k}\left(\frac{\partial\theta_i}{\partial\mu_i}\right)^2 b'(\theta_i)^2V[U_i]=-\sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{V[Y_i]}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \qquad (2.11)$$

また、(2.9)の関係より、以下のようにも書ける。

$$E[U_{jk}]=-E[U_jU_k] \qquad (2.12)$$

ここで、$(\Im)_{jk}=-E[U_{jk}]=E[U_jU_k]$ とすると、行列 $\Im$ は情報行列と呼ばれる。

$$(\Im)_{jk}=E(U_jU_k)=\sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{V[Y_i]}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \qquad (2.13)$$

今、(2.7)で与えられる対数尤度関数が最大となる $\boldsymbol{\beta}$ の値を求めてみよう。これには、

$$\frac{\partial l}{\partial\beta_j}=U_j=0 \qquad (2.14)$$

という方程式を解くことになる。

（$f(x)=0$ を解くには $y_m=f'(x_{m-1})(x_m-x_{m-1})-f(x_{m-1})=0$ を計算することから）

式(2.13)の解はニュートン・ラフソン法によると、

$$U_j^{(m)}=\sum_{k=1}^{p}U_{jk}^{(m-1)}(\beta_k^{(m)}-\beta_k^{(m-1)})+U_j^{(m-1)}=0$$

のように、$\beta_k^{(m)}$ の値を逐次求めて行くことになるが、実際の計算では $U_{jk}^{(\mathrm{m}-1)}$ の代わりに、情報行列 $-(\Im^{(m-1)})_{jk}$ を用いる。この式を書き変えると、以下となる。

$$\boldsymbol{\beta}^{(m)}=\boldsymbol{\beta}^{(m-1)}+(\Im^{(m-1)})^{-1}\mathbf{U}^{(m-1)} \qquad (2.15)$$

(2.10)式と(2.13)式を元にして、大標本においては、スコアベクトルの分布は漸近的に $\mathbf{U}\sim\mathbf{N}(\mathbf{0},\Im)$,

$${}^t\mathbf{U}\Im^{-1}\mathbf{U}\sim\chi^2(p) \qquad (2.16)$$

であることも示される。

最尤推定量 $l(\boldsymbol{\beta})$ の推定値 $\mathbf{b}$ の近傍でのテイラー展開近似は以下となり、

$$l(\boldsymbol{\beta}) = l(\mathbf{b}) + {}^t(\boldsymbol{\beta}-\mathbf{b})\mathbf{U}(\mathbf{b}) - \frac{1}{2}{}^t(\boldsymbol{\beta}-\mathbf{b})\mathfrak{I}(\mathbf{b})(\boldsymbol{\beta}-\mathbf{b})$$

スコアベクトルの推定値 $\mathbf{b}$ の近傍でのテイラー展開近似は以下となる。

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{U}(\mathbf{b}) - \mathfrak{I}(\mathbf{b})(\boldsymbol{\beta}-\mathbf{b}) = -\mathfrak{I}(\mathbf{b})(\boldsymbol{\beta}-\mathbf{b}) \tag{2.17}$$

ここでは $E[\partial U_j / \partial \beta_k] = -\mathfrak{I}_{jk}$ や $\mathbf{U}(\mathbf{b}) = \mathbf{0}$ を使っている。(2.16)と(2.17)より、

$${}^t(\boldsymbol{\beta}-\mathbf{b})\mathfrak{I}(\mathbf{b})(\boldsymbol{\beta}-\mathbf{b}) \sim \chi^2(p) \tag{2.18}$$

も示される。また、同様にして以下も示される。

$$\mathbf{b} = \boldsymbol{\beta} + \mathfrak{I}^{-1}\mathbf{U} \sim N(\beta, \mathfrak{I}^{-1}) \tag{2.19}$$

モデルの最適値からのずれを表す逸脱度 D を以下のように定義する。

$$D = 2[l(\mathbf{b}_{\max};\mathbf{y}) - l(\mathbf{b};\mathbf{y})] \sim \chi^2(N-p)$$

ここに、$l(\mathbf{b}_{\max};\mathbf{y})$ はパラメータ数 N の飽和モデルでの対数尤度、$l(\mathbf{b};\mathbf{y})$ は現在考えている、パラメータ数 p のモデルでの対数尤度である。同じパラメータ数では、この値が小さい連結関数のモデルほど適合が良いと判断する。但し、分布は漸近的に成り立つものであるから、0/1 形式のデータでは分布の形状はこの形にならないので注意を要する。

逸脱度と同様に最適値からのずれを表す統計量に以下のピアソン χ^2 統計量がある。

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)} \sim \chi^2(N-p) \tag{2.20}$$

これは逸脱度と漸近的に同じ指標であるが、逸脱度と比べてこちらの方が分布によく適合するという意見もある。

モデルに意味があるかどうかの検定では、以下の尤度比 χ^2 統計量が使われる。

$$\begin{aligned} C &= 2[l(\hat{\mathbf{p}};\mathbf{y}) - l(\tilde{p};\mathbf{y})] \\ &= 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{\hat{y}_i}{n_i\tilde{p}}\right) + (n_i - y_i)\log\left(\frac{n_i - \hat{y}_i}{n_i - n_i\tilde{p}}\right)\right] \sim \chi^2(p-1) \end{aligned} \tag{2.21}$$

ここに $l(\tilde{p};\mathbf{y})$ は定数パラメータ 1 つの最小モデルの対数尤度で、パラメータは以下のように推定される。

$$\tilde{p} = \sum_{i=1}^{N} y_i \Big/ \sum_{i=1}^{n} n_i$$

これは、帰無仮説として最小モデルが正しい（回帰式は意味がない）とする検定である。

実測値と推測値の関係を与える指標として、決定係数からの類推である以下の擬似 R^2 も利用される。

$$\tilde{R}^2 = \frac{l(\tilde{p};\mathbf{y}) - l(\hat{\mathbf{p}};\mathbf{y})}{l(\tilde{p};\mathbf{y})} \tag{2.22}$$

さらにプログラムでは、実測値と予測値の相関係数も求めている。

3) 2 項分布モデル

2 項分布のパラメータを説明変数の線形結合で推測する場合、密度関数、対数尤度関数、逸脱度、目的変数の平均と分散は以下のようになる。ここで、対数尤度関数の最後の項はパラメータに依存していないので、計算上は考えないことにする（参考文献[1]の数値に従っている）。

$$f(y_i;p_i) = {}_{n_i}C_{y_i}\, p_i^{y_i}(1-p_i)^{n_i-y_i}$$

$$\begin{aligned} l(y_i;p_i) &= \log[{}_{n_i}C_{y_i}\, p_i^{y_i}(1-p_i)^{n_i-y_i}] = y_i \log p_i + (n_i-y_i)\log(1-p_i) + \log {}_{n_i}C_{y_i} \\ &= y_i[\log p_i - \log(1-p_i)] + n_i \log(1-p_i) + \log {}_{n_i}C_{y_i} \\ &\to y_i[\log p_i - \log(1-p_i)] + n_i \log(1-p_i) \end{aligned}$$

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{n_i p_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i p_i}\right)\right]$$ 判定は～ $\chi^2(N-p)$ で行う。

$$b(p_i) = [\log p_i - \log(1-p_i)],\quad c(p_i) = \log(1-p_i),\quad d(y_i) = \log {}_{n_i}C_{y_i} \to 0$$

$$E[Y_i] = n_i p_i \equiv \mu_i$$

$$V[Y_i] = n_i p_i (1-p_i)$$

ここでは n_i 回の試行に対して、 y_i 回の事象が起こったとしているが、 1 回の試行で起こったか起こらないかにする場合は、 $n_i = 1, y_i = \{0,1\}$ とすればよい。

これまでは、2 項分布に基づく一般論であったが、これ以降は、説明変数との関係を与える連結関数の部分に仮定が入る。連結関数の仮定でよく利用されるモデルが、ロジスティックモデル、プロビットモデル、極値モデル等である。以下に最終的な計算で用いられる式を与えておく。

ロジスティックモデル

$$\eta_i = \log\frac{p_i}{1-p_i} = \sum_{j=1}^{p}\beta_j x_{ij} + \beta_0$$

$$p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{1}{1+e^{-\eta_i}},\quad 1-p_i = \frac{1}{1+e^{\eta_i}}$$

$$\mu_i = n_i p_i = \frac{n_i e^{\eta_i}}{1+e^{\eta_i}},\qquad \frac{\partial \mu_i}{\partial \eta_i} = \frac{n_i e^{\eta_i}}{(1+e^{\eta_i})^2} = n_i p_i (1-p_i)$$

$$U_j = \sum_{i=1}^{N}\frac{(y_i-\mu_i)x_{ij}}{V[Y_i]}\frac{\partial \mu_i}{\partial \eta_i} = \sum_{i=1}^{N}\frac{(y_i-\mu_i)x_{ij}}{n_i p_i(1-p_i)} n_i p_i (1-p_i) = \sum_{i=1}^{N}(y_i-\mu_i)x_{ij}$$

$$(\Im)_{jk} = \sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{V[Y_i]}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \sum_{i=1}^{N}\frac{x_{ij}x_{ik}n_i^2 p_i^2 (1-p_i)^2}{n_i p_i (1-p_i)} = \sum_{i=1}^{N} x_{ij}x_{ik} n_i p_i (1-p_i) \qquad (3.1)$$

プロビットモデル

$$\eta_i = \Phi^{-1}(p_i) = \sum_{j=1}^{p} \beta_j x_{ij} + \beta_0$$

$$p_i = \Phi(\eta_i)$$

$$\mu_i = n_i p_i = n_i \Phi(\eta_i), \qquad \frac{\partial \mu_i}{\partial \eta_i} = \frac{n_i}{\sqrt{2\pi}} \exp(-\eta_i^2/2)$$

$$U_j = \sum_{i=1}^{N} \frac{(y_i - \mu_i) x_{ij}}{V[Y_i]} \frac{\partial \mu_i}{\partial \eta_i} = \sum_{i=1}^{N} \frac{(y_i - \mu_i) x_{ij}}{p_i (1 - p_i)} \frac{\exp(-\eta_i^2/2)}{\sqrt{2\pi}}$$

$$(\Im)_{jk} = \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{V[Y_i]} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 = \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{p_i (1 - p_i)} \frac{n_i \exp(-\eta_i^2)}{2\pi} \tag{3.2}$$

極値モデル

$$\eta_i = \log[-\log(1 - p_i)] = \sum_{j=1}^{p} \beta_j x_{ij} + \beta_0$$

$$p_i = 1 - \exp[-\exp(\eta_i)] \qquad (1 - p_i = \exp[-\exp(\eta_i)])$$

$$\mu_i = n_i p_i, \qquad \frac{\partial \mu_i}{\partial \eta_i} = n_i \frac{\partial p_i}{\partial \eta_i} \exp(\eta_i) \exp[-\exp(\eta_i)]$$

$$U_j = \sum_{i=1}^{N} \frac{(y_i - \mu_i) x_{ij}}{V[Y_i]} \frac{\partial \mu_i}{\partial \eta_i} = \sum_{i=1}^{N} \frac{(y_i - \mu_i) x_{ij}}{p_i (1 - p_i)} \exp(\eta_i) \exp[-\exp(\eta_i)]$$

$$= \sum_{i=1}^{N} \frac{(y_i - \mu_i) x_{ij}}{p_i} \exp(\eta_i)$$

$$(\Im)_{jk} = \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{V[Y_i]} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 = \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{p_i (1 - p_i)} n_i \exp(2\eta_i) \exp[-2\exp(\eta_i)]$$

$$= \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{p_i} n_i (1 - p_i) \exp(2\eta_i) \tag{3.3}$$

このモデルの計算には以下の性質を利用する。 $\lim_{p \to 0} \exp(\eta)/p = 1$

プロビットモデルと極値モデルの場合、$p_i \to 0$ や $p_i \to 1$ のときに、計算機のまるめ誤差や分布関数の近似誤差から、除算のエラーが生じることがある。そのため、プログラムではある程度のところで、これらの極限を止めるようにしている。また最終結果でも対数尤度の計算で同様のことが起こる可能性があるので、同じように極端な値を避けるようにしている。現在のプログラムでは、$0.00000 \leq p_i \leq 0.99999$ の範囲に設定している。

20.2　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－2 値ロジスティック回帰］を選択すると、図 1 のような、2 値ロジスティック回帰分析の実行メニューが表示される。

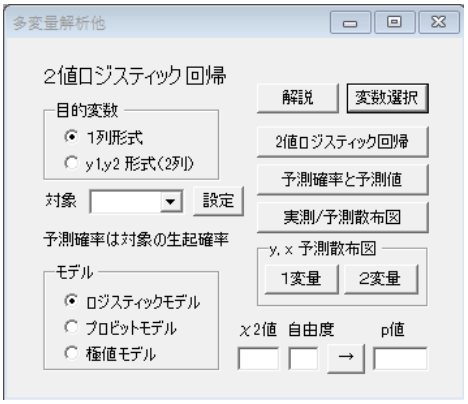


図 1　分析実行メニュー

利用するデータの形式は、「y1,y2 形式（2 列）」と「0/1 形式（1 列）」があり、それぞれ図 2a と図 2b のように、目的変数が 2 列で表されるか、1 列で表されるかの違いである。

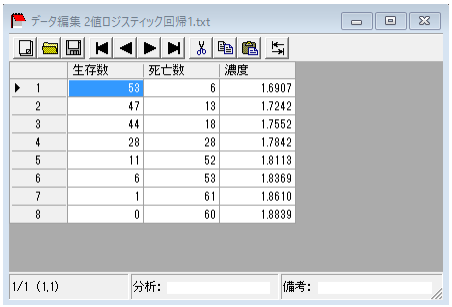


図 2a　目的変数 2 列データ

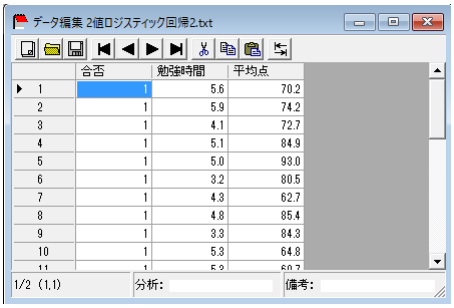


図 2b　目的変数 1 列データ

目的変数が 2 列で表される場合は、事象 1 が何回起きて、事象 2 が何回起きたかの重複のあるデータで、1 列で表される場合は、1 回の試行で事象が起きるかどうかの重複のないデータである。2 列の場合、対象変数と非対象変数を入力し、対象変数をコンボボックスで選択しておく。1 列のデータを起きない回数と起きた回数にして 2 列で表現することも可能である。目的変数が 1 列の場合は、2 列の特別な場合と考えてもよい。以後データ形式を分けて、プログラムの出力について説明する。

図 2a のデータのとき、「ロジスティックモデル」ラジオボタンを選択し、「2 値ロジスティック回帰」ボタンをクリックすると図 3 の結果が表示される。

ロジスティック回帰分析結果

	偏回帰係数	標準化係数	標準偏差	両側確率	2.5%下限	2.5%上限	EXP(b)
▶ 濃度	34.2703	2.3117	2.9121	0.0000	28.5625	39.9781	7.646E14
切片	-60.7175	0.7438	5.1807	0.0000	-70.8716	-50.5633	
対数尤度値	-186.235						
逸脱度D	11.232	自由度	6	上側確率	0.0815	<-n小注意	
ピアソンχ2	10.027	自由度	6	上側確率	0.1235	<-n小注意	
C尤度比	272.970	自由度	1	上側確率	0.0000		
擬似R2	0.423						
実測予測R2	0.989						

図 3　2 値ロジスティック回帰結果

ここでは回帰パラメータの値とその検定値、対数尤度値、逸脱度、目的変数と予測値との相関係数の 2 乗値が表示される。

また、「予測確率と予測値」ボタンをクリックすると、個別の実測値、予測確率、予測値が図 4 のように表示される。

予測確率と予測値

	実測値	予測確率	予測値
▶ 1	6	0.059	3.457
2	13	0.164	9.842
3	18	0.362	22.451
4	28	0.605	33.898
5	52	0.795	50.096
6	53	0.903	53.291
7	61	0.955	59.222
8	60	0.979	58.743

図 4　予測確率と予測値

「実測/予測散布図」をクリックすると、この実測値と予測値が、図 5 のようにプロットされる。

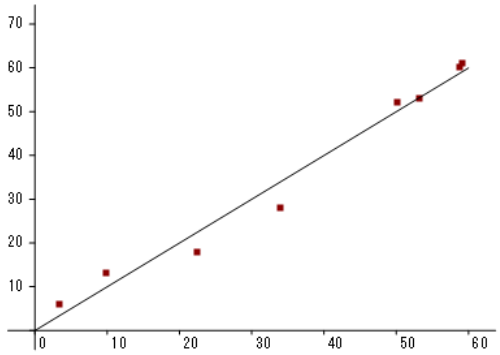


図 5　実測/予測散布図

予測の説明変数が 1 つまたは 2 つの場合、実測値と確率の予測関数（連結関数の逆関数）の関係を表示することができる。ここでは説明変数が 1 つであるので、「y, x 予測散布図」グループボックス内の「1 変量」ボタンをクリックする。結果は図 6 のようになる。

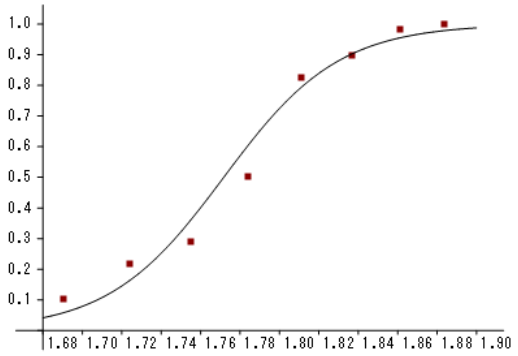


図 6　予測関数とデータ（ロジスティックモデル）

但し、ここでは軸設定を使ってグラフの軸を変更している。

この図と同様に、プロビットモデルと極値モデルの予測関数についても図 7a と図 7b で当てはまりを見てみる。

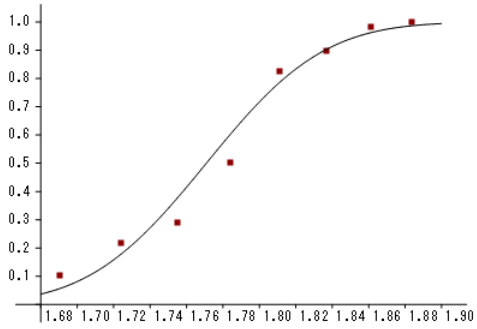


図 7a　プロビットモデル

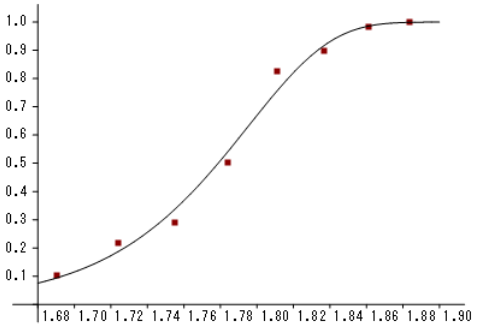


図 7b　極値モデル

これらを比べると極値モデルの当てはまりが良いことが分かる。このことは、「2 値ロジスティック回帰」ボタンで表示される、対数尤度値、逸脱度 D、R^2 の値でも確認できる。

次に図 2b のデータを用いた場合のロジスティックモデルの実行結果を示す。目的変数は「0/1 形式（1 列）」を選択し、「2 値ロジスティック回帰」ボタンをクリックすると図 8 のような結果が表示される。

ロジスティック回帰分析結果

	偏回帰係数	標準化係数	標準偏差	両側確率	2.5%下限	2.5%上限	EXP(b)
▶ 勉強時間	5.0765	5.9238	2.5049	0.0427	0.1670	9.9861	1.602E02
平均点	0.4581	5.2007	0.2231	0.0400	0.0208	0.8955	1.581E00
切片	-52.7491	-20.9681	24.9005	0.0341	-101.5540	-3.9442	
対数尤度値	-5.569						
逸脱度D	11.137	自由度	27	上側確率	0.9969	<-n小注意	
ピアソンχ2	13.422	自由度	27	上側確率	0.9863	<-n小注意	
C尤度比	29.917	自由度	2	上側確率	0.0000		
擬似R2	0.729						
実測予測R2	0.864						
誤判別確率	0を1と	0.059	1を0と	0.077			

図 8　2 値ロジスティック回帰結果

このデータ形式では､以下に述べる､予測による 0/1 の判別についての誤判別確率が追加されている。

また､「予測確率と予測値」ボタンをクリックすると、個体別の実測値、予測確率、予測値が図 9 のように表示される。

予測確率と予測値

	実測値	予測確率	予測値
9	1	0.932	1.000
10	1	0.979	1.000
11	1	0.877	1.000
12	1	1.000	1.000
13	1	0.991	1.000
14	0	0.000	0.000
15	0	0.374	0.000
16	0	0.070	0.000
17	0	0.005	0.000
18	0	0.000	0.000

図 9　予測確率と予測値

ここでは、予測値として、予測確率が 0.5 未満なら 0、予測確率が 0.5 以上なら 1 が与えられている。この予測値と実測値との違いを表すのが、図 8 の誤判別確率である。

「実測/予測散布図」をクリックすると、この実測値と予測確率が、図 10 のように表示される。ここに、図 10 の図 5 との違いは、実測値と予測値の代わりに実測値と予測確率を用いているところである。

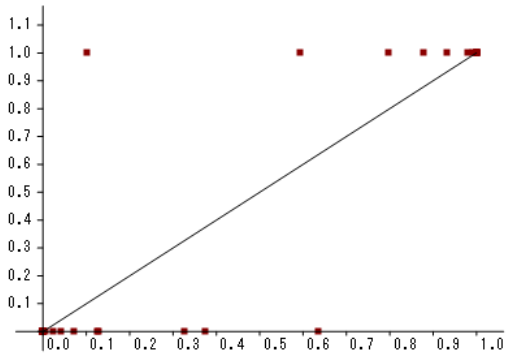


図 10　実測/予測確率散布図

このデータでは説明変数が 2 つであるので、「y, x 予測散布図」グループボックス内の「2 変量」ボタンをクリックする。結果は図 11 のようなグラフになる。

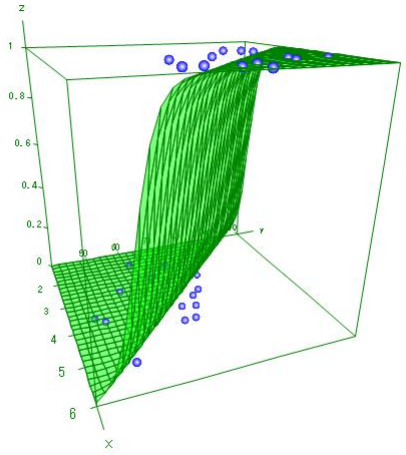


図 11　予測関数とデータ（ロジスティックモデル）

最後に、分析実行メニューの下部に、利用する可能性のある χ^2 分布の確率を求めるボタンを追加しておいた。専用のメニューもあるが、必要に応じて利用してもらいたい。

参考文献

[1] Annette J. Dobson 著, 田中豊他訳, 一般化線形モデル入門 原著第 2 版, 共立出版, 2008.

２１．多値ロジスティック回帰

１．多項分布モデル

多項分布の密度関数、対数尤度関数は以下で与えられる。ここで、対数尤度関数の最後の項はパラメータに依存していないので、計算上は考えないことにする（参考文献[1]の数値に従っている）。

密度関数

$$f(y_i;p_i)=n_i!\prod_{\alpha=1}^{J}\frac{p_{i\alpha}^{y_{i\alpha}}}{y_{i\alpha}!},\quad \sum_{j=1}^{J}y_{ij}=n_i,\quad \sum_{j=1}^{J}p_{ij}=1$$

これより、y_i 及び p_i の中の１つは他の変数で規定される。

対数尤度関数

$$l(y_i;p_i)=\log f(y_i;p_i)=\sum_{\alpha=1}^{J}y_{i\alpha}\log p_{i\alpha}+\log n_i!-\sum_{\alpha=1}^{J}y_{i\alpha}!\to\sum_{\alpha=1}^{J}y_{i\alpha}\log p_{i\alpha}$$

以下この関係を利用して計算過程を考えてみる。

1.1　名義ロジスティック回帰

一般性を失わず、他の変数で規定される定数を $\alpha=J$ とすると、

$$\frac{\partial l_i}{\partial p_{i\alpha}}=\sum_{i=1}^{J-1}y_{i\alpha}\log p_{i\alpha}+y_{iJ}\log p_{iJ}=\frac{y_{i\alpha}}{p_{i\alpha}}-\frac{y_{iJ}}{p_{iJ}}$$

$$E[Y_{i\alpha}]=n_i p_{i\alpha}\equiv\mu_{i\alpha}$$

$$Cov[Y_{i\alpha}Y_{i\beta}]=n_i p_{i\alpha}(\delta_{\alpha\beta}-p_{i\beta})$$

名義尺度ロジスティックモデルは、基準となるカテゴリに対する他のカテゴリのロジットを説明変数の線形結合で推測する。

$$\eta_{i\alpha}=\log\frac{p_{i\alpha}}{p_{i1}}=\sum_{k=1}^{p}\beta_{k\alpha}x_{ik}+\beta_{0\alpha},\quad j=2,\cdots,J$$

$$p_{i\alpha}=p_{i1}e^{\eta_{i\alpha}}$$

$$1=\sum_{\alpha=1}^{J}p_{i\alpha}=p_{i1}+\sum_{\alpha=2}^{J}p_{i\alpha}=p_{i1}[1+\sum_{\alpha=2}^{J}e^{\eta_{i\alpha}}]\quad \text{より、}$$

$$p_{i1}=\frac{1}{1+\sum_{\beta=2}^{J}e^{\eta_{i\beta}}},\quad p_{i\alpha}=\frac{e^{\eta_{i\alpha}}}{1+\sum_{\beta=2}^{J}e^{\eta_{i\beta}}}$$

対数尤度関数

$$l(y_i;p_i)=\log f(y_i;p_i)=\sum_{\alpha=1}^{J}y_{i\alpha}\log p_{i\alpha}+\log n_i!-\sum_{\alpha=1}^{J}y_{i\alpha}!\to\sum_{\alpha=1}^{J}y_{i\alpha}\log p_{i\alpha}$$

以下この関係を利用して計算過程を考えてみる。

$$\frac{\partial\mu_{i\beta}}{\partial\eta_{i\alpha}}=\frac{n_i e^{\eta_{i\beta}}\delta_{\alpha\beta}(1+\sum_{\gamma=1}^{J-1}e^{\eta_{i\gamma}})-n_i e^{\eta_{i\alpha}}e^{\eta_{i\beta}}}{(1+\sum_{\gamma=1}^{J-1}e^{\eta_{i\gamma}})^2}=n_i p_{i\beta}(\delta_{\alpha\beta}-p_{i\alpha})\quad(\alpha\neq J,\beta\neq J)\text{より、}$$

$$\frac{\partial\mu_{i\beta}}{\partial\beta_{j\alpha}}=\frac{\partial\eta_{i\alpha}}{\partial\beta_{j\alpha}}\frac{\partial\mu_{i\beta}}{\partial\eta_{i\alpha}}=x_{ij}\frac{\partial\mu_{i\beta}}{\partial\eta_{i\alpha}}=x_{ij}n_i p_{i\beta}(\delta_{\alpha\beta}-p_{i\alpha})\quad\alpha\neq J,\beta\neq J$$

以上より、

$$U_j^{\alpha}=\sum_{i=1}^{N}\frac{\partial l_i}{\partial\beta_{j\alpha}}=\sum_{i=1}^{N}\sum_{\beta=1}^{J-1}\sum_{\gamma=1}^{J-1}\frac{\partial\mu_{i\beta}}{\partial\beta_{j\alpha}}\frac{\partial p_{i\gamma}}{\partial\mu_{i\beta}}\frac{\partial l_i}{\partial p_{i\gamma}}$$

$$=\sum_{i=1}^{N}\sum_{\beta=1}^{J-1}\sum_{\gamma=1}^{J-1}x_{ij}n_i p_{i\beta}(\delta_{\alpha\beta}-p_{i\alpha})\frac{\delta_{\beta\gamma}}{n_i}\left(\frac{y_{i\gamma}}{p_{i\gamma}}-\frac{y_{iJ}}{p_{iJ}}\right)$$

$$=\sum_{i=1}^{N}\sum_{\beta=2}^{J}x_{ij}p_{i\beta}(\delta_{\alpha\beta}-p_{i\alpha})\left(\frac{y_{i\beta}}{p_{i\beta}}-\frac{y_{iJ}}{p_{iJ}}\right)=\sum_{i=1}^{N}x_{ij}(y_{i\alpha}-n_i p_{i\alpha})$$

$$\Im_{jk}^{\alpha}=-\frac{\partial U_j^{\alpha}}{\partial\beta_{k\alpha}}=-\frac{\partial}{\partial\beta_{k\alpha}}\sum_{i=1}^{N}x_{ij}(y_{i\alpha}-n_i p_{i\alpha})=\sum_{i=1}^{N}x_{ij}n_i\sum_{\beta=1}^{J-1}\frac{\partial\mu_{i\beta}}{\partial\beta_{k\alpha}}\frac{\partial p_{i\alpha}}{\partial\mu_{i\beta}}$$

$$=\sum_{i=1}^{N}x_{ij}n_i\sum_{\beta=1}^{J-1}x_{ik}n_i p_{i\beta}(\delta_{\alpha\beta}-p_{i\alpha})\frac{\delta_{\alpha\beta}}{n_i}=\sum_{i=1}^{N}x_{ij}x_{ik}n_i p_{i\alpha}(1-p_{i\alpha})$$

これらのスコアベクトルと情報ベクトルより、$\beta_{j\alpha}$（$\alpha\neq 1$）は推定される。

最適値からのずれを表す、逸脱度、ピアソンのχ^2統計量及び、最小モデルからのずれを表す、尤度比χ^2統計量は以下のようになる。

$$D=2\sum_{i=1}^{N}\sum_{\alpha=1}^{J}y_{i\alpha}\log\frac{y_{i\alpha}}{n_i\hat{p}_{i\alpha}}\sim\chi^2((J-1)(N-p))$$

$$\chi^2=\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{(y_{ij}-\hat{y}_{ij})^2}{n_i\hat{p}_{ij}}\sim\chi^2((J-1)(N-p))$$

$$C=2\sum_{i=1}^{N}\sum_{\alpha=1}^{J}y_{i\alpha}\log\frac{\hat{y}_{i\alpha}}{n_i\tilde{p}_{\alpha}}\sim\chi^2((J-1)(p-1)),\quad\tilde{p}_{\alpha}=\sum_{i=1}^{N}y_{i\alpha}\Big/\sum_{i=1}^{N}n_i$$

ピアソンのχ^2統計量は逸脱度と漸近的に同じ指標であるが、逸脱度と比べてこちらの方が分布によ

く適合するという意見もある。ここではn_i回の試行に対して、y_i回の事象が起こったとしているが、１回の試行で起こったか起こらないかにする場合は、$n_i=1, y_i=\{0,1\}$とする。但し、分布はデータ数が無限大のときの極限であるので、注意が必要である。

1.2　順序ロジスティック回帰

順序ロジスティック回帰には、累積ロジットモデル、隣接カテゴリ―ロジットモデル、連続比ロジットモデルなどがあるが、ここでは最も扱いやすく、プログラムで取り入れている累積ロジットモデルについて説明する。他のモデルについては、プログラムに導入次第報告する。

累積ロジットモデル

累積ロジットモデルでは、以下の比の対数を線形関数で予測する。

$$\frac{p_1}{p_2+\cdots+p_J}=e^{\eta_1},\quad \frac{p_1+p_2}{p_3+\cdots+p_J}=e^{\eta_2},\quad \cdots,\quad \frac{p_1+\cdots+p_{J-1}}{p_J}=e^{\eta_{J-1}}$$

これは、連続した複数のカテゴリーの出現確率と残りのカテゴリーの出現確率のオッズ比を説明変数の線形関数で予測することに相当する。

上の関係を以下のように書き換え、

$$\frac{p_2+\cdots+p_J}{p_1}=e^{-\eta_1},\quad \frac{p_3+\cdots+p_J}{p_1+p_2}=e^{-\eta_2},\quad \cdots,\quad \frac{p_J}{p_1+\cdots+p_{J-1}}=e^{-\eta_{J-1}}$$

$q_\alpha=p_1+p_2+\cdots+p_\alpha$と定義すると、以下の関係が示される。

$1-p_1=p_2+\cdots+p_J=p_1e^{-\eta_1}$　より、$p_1=\dfrac{e^{\eta_1}}{1+e^{\eta_1}}=q_1$

$p_2=p_1e^{-\eta_1}-(p_1+p_2)e^{-\eta_2}$　より、

$$p_2=\frac{1}{1+e^{-\eta_2}}-\frac{1}{1+e^{-\eta_1}},\quad p_1+p_2=\frac{e^{\eta_2}}{1+e^{\eta_2}}=q_2$$

$p_3=(p_1+p_2)e^{-\eta_2}-(p_1+p_2+p_3)e^{-\eta_3}$　より、

$$p_3=\frac{1}{1+e^{-\eta_3}}-\frac{1}{1+e^{-\eta_2}},\quad p_1+p_2+p_3=\frac{e^{\eta_3}}{1+e^{\eta_3}}=q_3$$

同様にして、

$$p_{J-1}=\frac{1}{1+e^{-\eta_{J-1}}}-\frac{1}{1+e^{-\eta_{J-2}}},\quad p_1+\cdots+p_{J-1}=\frac{e^{\eta_{J-1}}}{1+e^{\eta_{J-1}}}=q_{J-1}$$

また、

$$p_J=1-(p_1+\cdots+p_{J-1})=1-\frac{1}{1+e^{-\eta_{J-1}}}=\frac{1}{1+e^{\eta_{J-1}}}=1-q_{J-1}$$

これらより、q_α について考えれば、各カテゴリー α について独立に、q_α と $1-q_\alpha$ の 2 項分布として β_α の値を推定できることが分かる。そのためこれは 2 値ロジスティック回帰の拡張として捉えることができ、各カテゴリ p_α（$\alpha=1,2,\cdots,J$）については以下のように与えることができる。

$$p_1=q_1,\ p_\alpha=q_\alpha-q_{\alpha-1},\ p_J=1-q_{J-1}$$

1.3　プログラムの利用法

メニュー［分析－多変量解析等－判別手法－多値ロジスティック回帰］を選択すると図 1 のような多値ロジスティック回帰分析の分析実行メニューが表示される。

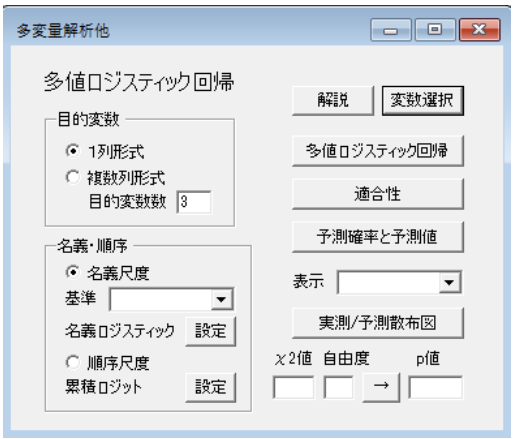


図 1　分析実行メニュー

複数列形式のデータの例を図 2 に示す。

	重要でない	重要	とても重要	性別	年齢	職業
1	26	12	7	0	0	0
2	9	21	15	0	1	0
3	5	14	41	0	0	1
4	40	17	8	1	0	0
5	17	15	12	1	1	0
6	8	15	18	1	0	1

図 2　複数列形式のデータ

「目的変数」グループボックスの「複数列形式」を選択し、変数選択ですべての変数を選択し、「名義ロジスティック」の設定から図 3 のように基準に「重要でない」を選択する。

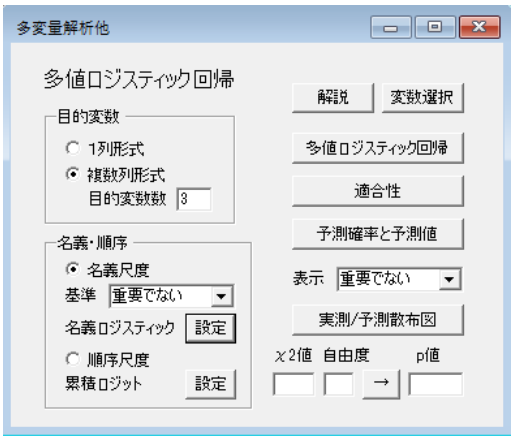


図 3　複数列目的変数の名義ロジスティック設定

ここでは、「重要でない」カテゴリーの確率で、他のカテゴリーの確率を割った対数オッズを説明変数の線形関数で推定することになる。

「多値ロジスティック回帰」ボタンをクリックすると図 4 のような分析結果が表示される。

log(確率/重要でない確率)の線形予測

	偏回帰係数	標準化係数	標準偏差	両側確率	2.5%下限	2.5%上限	EXP(b)
▶ 重要/重要でない							
性別	-0.3881	-0.2126	0.2528	0.1247	-0.8835	0.1073	6.783E-01
年齢	1.1283	0.5826	0.3064	0.0002	0.5277	1.7289	3.090E00
職業	1.5877	0.8199	0.3108	0.0000	0.9785	2.1969	4.892E00
切片	-0.5908	-0.0616	0.2595	0.0228	-1.0993	-0.0823	
とても重要/重要でない							
性別	-0.8130	-0.4453	0.2710	0.0027	-1.3442	-0.2818	4.435E-01
年齢	1.4781	0.7633	0.3640	0.0000	0.7647	2.1916	4.385E00
職業	2.9167	1.5062	0.3463	0.0000	2.2380	3.5955	1.848E01
切片	-1.0391	-0.0668	0.3072	0.0007	-1.6412	-0.4370	

図 4　対数オッズの推定

ここでは、オッズ比推定の偏回帰係数、標準化偏回帰係数、偏回帰係数の標準誤差、偏回帰係数が 0 となる検定確率、偏回帰係数の下限と上限、説明変数単位量の変化によるオッズ比の変化量が表示される。

「適合性」ボタンをクリックすると、図 5 のように各種の適合性指標が表示される。

適合性

▶ 対数尤度値	-290.351				
逸脱度D	3.939	自由度	4	上側確率	0.4144
ピアソンχ2	3.927	自由度	4	上側確率	0.4160
C尤度比	77.842	自由度	6	上側確率	0.0000
擬似R2	0.118				
実測予測R2	0.981				

図 5　適合性指標

「予測確率と予測値」ボタンをクリックすると、図 6 のような結果が表示される。

予測確率と予測値

	重要でない	予測確率	予測値	重要	予測確率	予測値	とても重要	予測確率	予測値
▶	26	0.524	23.589	12	0.290	13.066	7	0.185	8.345
	9	0.235	10.556	21	0.402	18.069	15	0.364	16.375
	5	0.098	5.855	14	0.264	15.865	41	0.638	38.280
	40	0.652	42.411	17	0.245	15.934	8	0.102	6.655
	17	0.351	15.444	15	0.408	17.931	12	0.241	10.625
	8	0.174	7.145	15	0.320	13.134	18	0.505	20.721

図 6　予測確率と予測値

これには 3 つのカテゴリーについての実測値、予測確率、予測値が表示される。「表示変数」を 1 つ選んで、「実測/予測散布図」ボタンをクリックすると、図 7 のような散布図が表示される。

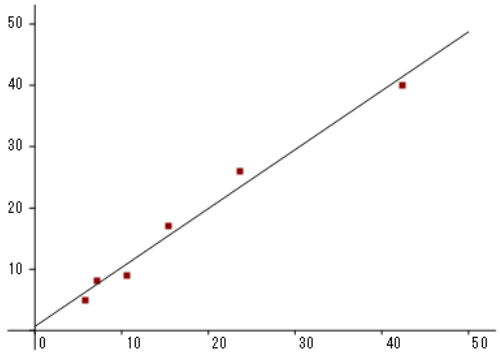


図 7　実測/予測散布図

同じデータを順序尺度として、順序ロジスティックの累積ロジットモデルで分析すると図 8 のような結果を得る。

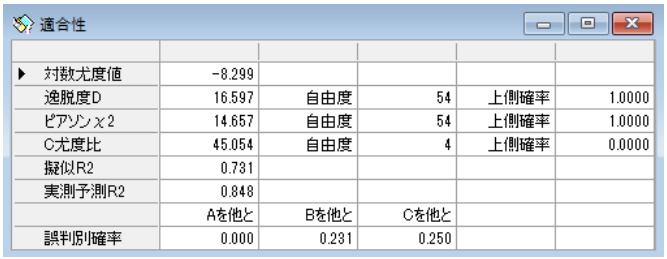
log(累積確率/他累積確率)の線形予測

	偏回帰係数	標準化係数	標準偏差	両側確率	2.5%下限	2.5%上限	EXP(b)
▶ 重要でない/重要～							
性別	0.5723	0.3135	0.2708	0.0346	0.0416	1.1031	1.772E00
年齢	-1.2597	-0.6505	0.3064	0.0000	-1.8602	-0.6591	2.838E-01
職業	-2.2490	-1.1614	0.3577	0.0000	-2.9501	-1.5479	1.055E-01
切片	0.0746	-0.6751	0.2492	0.7647	-0.4139	0.5631	
～重要/とても重要							
性別	0.5930	0.3248	0.2710	0.0286	0.0619	1.1241	1.809E00
年齢	-0.9711	-0.5015	0.3638	0.0076	-1.6841	-0.2581	3.787E-01
職業	-2.1137	-1.0915	0.3462	0.0000	-2.7923	-1.4351	1.208E-01
切片	1.5266	0.8221	0.3087	0.0000	0.9216	2.1317	

図 8　累積ロジットモデルでの結果

これは最初が「重要でない」を「重要」と「とても重要」を足したカテゴリーで割った対数オッズ、次が「重要でない」と「重要」を足したカテゴリーを「とても重要」で割った対数オッズについての説明変数の線形関数での推定である。

最後に目的変数が同じファイル 2 頁目の「1 列形式」（ファイルは異なる）で与えられる場合、「適合性」の結果に図 9 のように誤判別確率の値が表示される。


適合性

対数尤度値	-8.299				
逸脱度D	16.597	自由度	54	上側確率	1.0000
ピアソンχ2	14.657	自由度	54	上側確率	1.0000
C尤度比	45.054	自由度	4	上側確率	0.0000
擬似R2	0.731				
実測予測R2	0.848				
	Aを他と	Bを他と	Cを他と		
誤判別確率	0.000	0.231	0.250		

図 9　1 列形式の場合の適合性結果

参考文献

[1] Annette J. Dobson 著, 田中豊他訳, 一般化線形モデル入門 原著第 2 版, 共立出版, 2008.

２２．K-平均法

K-平均法は、非階層的なクラスター分析の代表的な手法の１つで、多数のデータでも高速に分類できる特徴を持っている。データ $x_{i\lambda}$ は λ 番目（$\lambda=1,\cdots,N$）の個体の i 番目（$i=1,\cdots,p$）の変数を表している。K-平均法はこの個体をある決められた K 個のクラスターに分類する。ここではプログラム中で使ったこの手法の手順を示しておく。

データはそのままでも標準化してもよいが、データの大きさや単位が異なる場合は標準化して使用する方がすべての変数を同等に扱える。ここでは標準化したデータも $x_{i\lambda}$ で表すことにする。

K-平均法は以下の方法によってクラスター構成を行う。

①データの中から K 個のデータを乱数によって選び出し、それをクラスターのシードにして、他のデータを最も近いシードに配置し、K 個のクラスターを構成する。

②各クラスターの重心を新たなクラスターのシードとして、クラスターを再配置する。前回のクラスターと新しいクラスターの構成が異なれば再配置をもう一度繰り返し、同じならば終了する。

この方法は簡単で、高速であるが、結果は最初の乱数に依存することが多い。そのため、階層的クラスター分析の Ward 法で用いられる within group error の考え方を取り入れ、その総和 E の最も小さいものを最良の候補とする。

メニュー［分析－多変量解析－クラスター分析－K-平均法］を選択すると図１のような分析メニューが表示される。

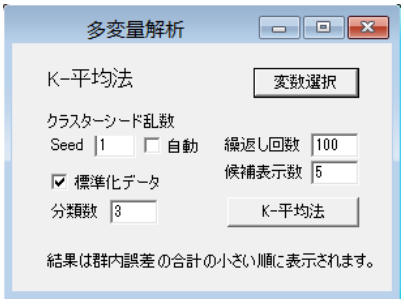


図１　分析メニュー

例としてクラスター分析 2.txt のデータを用いて、「分類数」を 3 にし、「K-平均法」のボタンをクリックすると「候補表示数」に示された 5 個のクラスター分類の候補が図 2 のように表示される。ここでは、「標準化データ」のチェックボックスにチェックを入れ、データを標準化した後、計算を実行している。また、クラスター分類は最初のシードの設定を変えながら「繰返し回数」100 回行い、異なった解のうち、within group error の総和の小さい順に表示されている。

	候補1	候補2	候補3	候補4	候補5
亀本	1	1	1	1	1
藤田	2	2	2	2	2
三井	3	3	3	2	2
松井	2	2	1	3	3
村社	3	3	3	2	2
田中	3	3	3	2	2
佐藤	3	3	3	2	2
増川	1	1	1	3	3
福井	1	1	1	1	1
三好	1	1	1	1	3
芝田	2	1	1	3	3
細川	3	3	3	2	2
門田	2	2	2	2	2
尾崎	2	2	2	2	2
奥田	3	3	3	2	2
群内誤差合計	19.554	20.548	20.717	22.556	23.401

図 2　結果表示

この表示では、欠損値などで計算不可能な部分は空欄として表示されるので、個体数は順番通りに表示され、グリッドエディタにコピーして分類データとして活用することもできる。

参考文献

２３．生存時間分析

生存時間分析は中途打ち切りを含むデータから死亡危険率や生存確率分布を予測する分析手法である。この分析は生物の生存時間だけでなく、機械の故障までの時間などにも利用できる。そのため、死亡という言葉は、あるイベントが発生するまでの時間とした方が的を射ているが、ここでは慣例的に使われてきた死亡や生存という言葉を使うことにする。

1. 生存時間分析の基礎

時刻 $t=0$ に $l(0)$ 個の個体があり、死亡で時刻 t に個体数が $l(t)$ 個になっているものとする。時刻 t からの単位時間の間に死亡する割合 $p(t)=-\dfrac{dl(t)}{dt}$ は、以下で与えられると仮定する。

$$-\frac{dl(t)}{dt}=\mu(t)l(t)$$

ここに $\mu(t)$ は時刻 t における死力という。

上式を時刻 t と時刻 $t+h$ の間で定積分すると以下の関係を得る。

$$\log l(t+h)-\log l(t)=-\int_t^{t+h}\mu(\tau)d\tau=-\int_0^h\mu(t+\tau)d\tau$$

これより、

$$l(t+h)=l(t)\exp\left[-\int_0^h\mu(t+\tau)d\tau\right]$$

ここで、 $p(h;t)=\exp\left[-\int_0^h\mu(t+\tau)d\tau\right]$ とおくと、 $p(h;t)$ は時間 $t\sim t+h$ の間の期間生存率と呼ばれる。この期間生存率は、以下のようになる。

$$p(h;t)=\frac{l(t+h)}{l(t)}$$

同様にして、期間死亡率 $q(h;t)$ も以下のように与えられる。

$$q(h;t)=1-p(h;t)=\frac{l(t)-l(t+h)}{l(t)}\equiv\frac{d(h;t)}{l(t)}$$

ここに $d(h;t)$ は期間死亡数を表す。

特に、 $h=1$ とした区間生存率、区間死亡率を単に時刻 t での生存率 $p(t)$ 、死亡率 $q(t)$ という。

時刻 t 以降の生存時間の合計 $T(t)$ を個体の数で割った $e(t)$ を平均余命という。

$$e(t)=\int_t^\infty l(\tau)d\tau\Big/l(t)=T(t)/l(t)$$

また、 $t=0$ での平均余命を平均寿命という。

死亡の発生までの時間を確率変数 T とする確率分布を考え、その密度関数を $f(t)$ 、分布関数を $F(t)$ とすると、これらには以下の関係がある。分布関数 $F(t)$ は累積死亡関数である。

$$F(t) = P(0 \le T \le t) = \int_0^t f(\tau) d\tau$$

これに対して、時刻 t まで生きる確率を表す関数を累積生存関数 $S(t)$ といい、以下で表す。

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(t) dt$$

時刻 t における死亡発生危険率をハザード関数（故障率関数）$\lambda(t)$ といい、以下のように定義する。

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t)$$

死亡率 $q(t)$ は以下のように定義されるが、

$$q(t) = \int_t^{t+1} f(\tau) d\tau \Big/ S(t)$$

時間の分割が小さい場合は、近似的にハザード関数の積分としても表される。

$$q(t) \simeq \int_t^{t+1} \lambda(\tau) d\tau$$

このハザード関数を積分した累積ハザード関数 $\Lambda(t)$ は以下のように定義される。

$$\Lambda(t) = \int_0^t \lambda(\tau) d\tau = -\log S(t)$$

逆に累積生存関数は、以下のように表される。

$$S(t) = e^{-\Lambda(t)}$$

累積生存関数は $t \to \infty$ で $S(t) \to 0$ であるから、累積ハザード関数は $t \to \infty$ で $\Lambda(t) \to \infty$ でなければならない。

生存時間分布には、主に指数分布とワイブル分布が仮定される。

指数分布の確率密度関数は以下で与えられる。

$$f(t) = \lambda e^{-\lambda t} \quad （t \ge 0）$$

分布関数と累積生存関数はそれぞれ以下で与えられる。

$$F(t) = 1 - e^{-\lambda t}, \quad S(t) = e^{-\lambda t}, \quad （t \ge 0）$$

確率変数の平均、分散、標準偏差はそれぞれ以下で与えられる。

$$E[T] = \frac{1}{\lambda}$$

$$V[T] = \frac{1}{\lambda^2}$$

$$\sigma = \sqrt{V[T]} = \frac{1}{\lambda}$$

ハザード関数は定数で与えられる。

$$\lambda(t)=\frac{f(t)}{S(t)}=\frac{\lambda e^{-\lambda t}}{e^{-\lambda t}}=\lambda$$

ワイブル分布の確率密度関数は以下で与えられる。

$$f(t)=(a/b)\left(t/b\right)^{a-1}\exp\left[-\left(t/b\right)^{a}\right] \qquad (t\geq 0)$$

分布関数と累積生存関数はそれぞれ以下で与えられる。

$$F(t)=1-\exp\left[-\left(t/b\right)^{a}\right],\quad S(t)=\exp\left[-\left(t/b\right)^{a}\right] \qquad (t\geq 0)$$

確率変数の平均、分散、標準偏差はそれぞれ以下で与えられる。

$$E[T]=b\,\Gamma\left(1+1/a\right)$$

$$V[T]=b^{2}\left[\Gamma(2+1/a)-\Gamma(1+1/a)\right]^{2}$$

$$\sigma=\sqrt{V[T]}$$

ハザード関数は以下で与えられる。

$$\lambda(t)=\frac{f(t)}{S(t)}=\frac{(a/b)\left(t/b\right)^{a-1}\exp\left[-\left(t/b\right)^{a}\right]}{\exp\left[-(t/b)^{a}\right]}=(a/b)\left(t/b\right)^{a-1}=at^{a-1}/b^{a}$$

実際のハザード関数は、初期段階で値が大きく、しばらく時間が経つと安定期に入り、最終的な段階でまた値が大きくなる。安定期では指数分布が使われ、初期段階ではワイブル分布がよく利用される。最終段階ではどちらの分布もあまり当てはまりが良くないと言われている。

2. Kaplan-Meier 推定と log-rank 検定

観測対象 $\lambda=1,\cdots,N$ に対して、生存時間を $t_\lambda=0$ から $t_\lambda=T_\lambda$（打ち切りのないデータ）、$t_\lambda=0$ から $t_\lambda=T_\lambda^{+}$（打ち切りのあるデータ、実際のデータでは 17+ 等と表記）とする。この終了時刻 T_λ を 0 から順番に並べた時刻を $t_0=0,t_1,\cdots,t_m$（同一のものもある）とし、t_m ですべて死亡および打ち切りが確認されたものとする。これに対して、一定の時間間隔で時刻を取る方法もある。各時点での生存数を l_i、$t_i<t\leq t_{i+1}$ の間に死亡した数を d_i、打ち切りになった数を w_i とする。これらを使って、死亡のリスクにさらされた数を $r_i=l_i-w_i/2$ とする。

死亡の期間発生率 q_i と期間生存率 p_i は以下で与えられる。

$$q_i=d_i/r_i,\quad p_i=1-q_i$$

累積生存関数 S_i、密度関数 f_i、ハザード関数 λ_i は以下のように計算される。

$$S_i=\prod_{k=0}^{i-1}p_k,\quad f_i=q_iS_i/(t_i-t_{i-1}),\quad \lambda_i=f_i/S_i=q_i$$

このような累積生存関数の推定法を Kaplan-Meier の product-limit 推定法という。累積生存関数 S_i のばらつきを表す標準誤差 $S.E.[S_i]$ は近似的に以下で与えられることが知られている。

$$S.E.[S_i] = S_{i-1}\sqrt{\sum_{k=1}^{i-1}\frac{d_k}{l_k(l_k-d_k)}} \qquad (i \geq 2)$$

期間内の生存時間 μ_i は以下で与えられる。

$$\mu_i = S_i(t_i - t_{i-1})$$

指数分布やワイブル分布の見極めは、累積ハザード関数に関する以下の関係を利用し、グラフが直線になるか否かで判断することができる。

指数分布　　$-\log S(t) = \lambda t$

ワイブル　　$\log(-\log S) = a\log(t/b) = a\log t - a\log b$

指数分布やワイブル分布のパラメータの最小 2 乗推定は、以下の式によって与えられる。

指数分布　$S(t) = e^{-\lambda t}$

$$\lambda = -\sum_{i=0}^{m-1} t_i \log S_i \Big/ \sum_{i=0}^{m-1} t_i^2$$

ワイブル分布　$S(t) = \exp\left[-(t/b)^a\right]$

$t_i' = \log t_i$，　$S_i' = \log(-\log S_i)$　として、

$$a = \sum_{i=1}^{m-1}(t_i' - \bar{t}')(S_i' - \bar{S}') \Big/ \sum_{i=1}^{m-1}(t_i' - \bar{t}')^2 \quad , \quad b = \exp\left[-(\bar{S}' - a\bar{t}')/a\right]$$

分類数 G の個体群について、生存時間データの差の検定を行うには以下の性質を用いる。第 r 分類群の t_i 時点での期間死亡数を d_i^r 、生存数を l_i^r として

$$O_r = \sum_{i=0}^{m-1} d_i^r \quad , \quad E_r = \sum_{i=0}^{m-1} l_i^r (d_i/l_i) \quad , \quad \text{ここに、} l_i = \sum_{r=1}^{G} l_i^r \ , \ d_i = \sum_{r=1}^{G} d_i^r$$

を計算し、以下の近似的な関係を用いて群間の差を検定する。

$$\chi^2 = \sum_{r=1}^{G}\frac{(O_r - E_r)^2}{E_r} \sim \chi^2_{G-1}$$

この検定を Peto & Peto の log-rank 検定という。

3. パラメータの最尤推定

3.1 指数分布に基づく最尤推定

最初に通常の指数分布の最尤推定を考える。指数分布の確率密度関数と分布関数は以下で与えられ

る。

$$f(t)=\lambda\exp(-\lambda t) \qquad (t\geq 0)$$

$$S(t)=\exp(-\lambda t) \qquad (t\geq 0)$$

指数分布の最尤推定で、尤度 $L(\lambda)$ は以下で与えられる。

$$L(\lambda)=\prod_{i=1}^{N} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

ここで、打ち切りデータと非打ち切りデータをそれぞれ $\delta_i=0,1$ としている。

ハザード関数は以下で与えられる。

$$\lambda(t)=\frac{f(t)}{S(t)}=\lambda$$

対数尤度は以下となる。

$$\log L(\lambda)=\sum_{i=1}^{N}\left[\delta_i \log\lambda(t_i)+\log S(t_i)\right]$$

$$=\sum_{i=1}^{N}\left[\delta_i \log\lambda-\lambda t_i\right]$$

対数尤度を微分してスコアベクトルに相当するものを作成するが、この場合はスカラーである。これを仮にスコアと呼ぶ。

$$\frac{\partial}{\partial\lambda}\log L=\sum_{i=1}^{N}\left[\delta_i/\lambda-t_i\right]=\frac{1}{\lambda}\sum_{i=1}^{N}\delta_i-\sum_{i=1}^{N}t_i=0$$

$$\lambda=\sum_{i=1}^{N}\delta_i\Big/\sum_{i=1}^{N}t_i$$

スコアをもう一度微分して、情報行列 $\Im$ に相当するものを作成する。この場合もスカラーである。

$$\Im=-\frac{\partial^2}{\partial\lambda^2}\log L=\frac{1}{\lambda^2}\sum_{i=1}^{N}\delta_i$$

この逆数は、推定値の分散を与える。

3.2 ワイブル分布に基づく最尤推定

最初に通常のワイブル分布の最尤推定を考える。ワイブル分布の確率密度関数と分布関数は以下で与えられる。

$$f(t)=(a/b)\left(t/b\right)^{a-1}\exp\left[-\left(t/b\right)^{a}\right] \qquad (t\geq 0)$$

$$S(t)=\exp\left[-\left(t/b\right)^{a}\right] \qquad (t\geq 0)$$

ワイブル分布の最尤推定で、尤度 $L(a,b)$ は以下で与えられる。

$$L(a,b)=\prod_{i=1}^{N} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

ここで、打ち切りデータと非打ち切りデータをそれぞれ $\delta_i = 0,1$ としている。

ハザード関数は以下で与えられる。

$$\lambda(t)=\frac{f(t)}{S(t)}=\frac{(a/b)\left(t/b\right)^{a-1}\exp\left[-\left(t/b\right)^{a}\right]}{\exp\left[-(t/b)^{a}\right]}=(a/b)\left(t/b\right)^{a-1}=at^{a-1}b^{-a}$$

対数尤度は以下となる。

$$\begin{aligned}\log L(a,b)&=\sum_{i=1}^{N}\left[\delta_i\log\lambda(t_i)+\log S(t_i)\right]\\&=\sum_{i=1}^{N}\left[\delta_i\log\left(at_i^{a-1}b^{-a}\right)-t_i^{a}b^{-a}\right]\\&=\sum_{i=1}^{N}\left[\delta_i\log\left(at_i^{a-1}e^{\beta}\right)-t_i^{a}e^{\beta}\right]\\&=\sum_{i=1}^{N}\left[\delta_i\left(\log a+(a-1)\log t_i+\beta\right)-t_i^{a}e^{\beta}\right]\end{aligned}$$

ここで、$b^{-a}=e^{\beta}$（$b=e^{-\beta/a}$，$e^{\beta}=b^{-a}\rightarrow\exp({}^{t}\mathbf{x}\boldsymbol{\beta})$ に相当）としている。

これを微分して、スコアベクトル $\mathbf{U}$ と情報行列 $\mathfrak{I}$ をもとめると以下となる。

$$\boldsymbol{\beta}'=\begin{pmatrix}a\\ \beta\end{pmatrix},\qquad \mathbf{U}=\begin{pmatrix}\partial\log L/\partial a\\ \partial\log L/\partial\beta\end{pmatrix},\qquad \mathfrak{I}=-\begin{pmatrix}\partial^2\log L/\partial a^2 & \partial^2\log L/\partial a\partial\beta\\ \partial^2\log L/\partial a\partial\beta & \partial^2\log L/\partial\beta^2\end{pmatrix}$$

ここに、

$$\frac{\partial}{\partial a}\log L=\sum_{i=1}^{N}\left[\delta_i(1/a+\log t_i)-\log t_i\, t_i^{a}e^{\beta}\right]$$

$$\frac{\partial}{\partial\beta}\log L=\sum_{i=1}^{N}\left[\delta_i-t_i^{a}e^{\beta}\right]$$

$$\frac{\partial^2}{\partial a^2}\log L=-\sum_{i=1}^{N}\left[\delta_i/a^2+(\log t_i)^2 t_i^{a}e^{\beta}\right]$$

$$\frac{\partial}{\partial a\partial\beta}\log L=-\sum_{i=1}^{N}\log t_i\, t_i^{a}e^{\beta}$$

$$\frac{\partial^2}{\partial\beta^2}\log L=-\sum_{i=1}^{N}t_i^{a}e^{\beta}$$

これらを用いてニュートン・ラフソン法でパラメータの推定を行う。

$$\boldsymbol{\beta}'^{(m+1)}=\boldsymbol{\beta}'^{(m)}+(\mathfrak{I}^{(m)})^{-1}\mathbf{U}^{(m)}$$

ここに右肩の添え字はニュートン・ラフソン法のループの段階を表している。

この情報行列の逆行列の対角成分はパラメータの分散を与える。

3.3 混合分布に基づく最尤推定

混合分布の最尤推定で、尤度 $L(\lambda)$ は以下で与えられる。

$$L(\lambda)=\prod_{i=1}^{N} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

K 種混合分布では、それぞれの密度関数を $f_k(t)$、分布関数を $S_k(t)$ として、全体の密度関数と分布関数は以下となる。ここに、π_k は分布の重ね合わせの確率である。

$$f(t)=\sum_{k=1}^{K}\pi_k f_k(t), \qquad S(t)=\sum_{k=1}^{K}\pi_k S_k(t)$$

混合分布の最尤推定で、尤度 $L(\boldsymbol{\theta},\boldsymbol{\pi})$ は以下で与えられる。

$$L(\boldsymbol{\theta},\boldsymbol{\pi})=\prod_{i=1}^{N} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}=\prod_{i=1}^{N}\left(\sum_{k=1}^{K}\pi_k f_k(t_i)\right)^{\delta_i}\left(\sum_{k=1}^{K}\pi_k S_k(t_i)\right)^{1-\delta_i}$$

ここで、打ち切りデータと非打ち切りデータをそれぞれ $\delta_i=0,1$ としている。

対数尤度は以下となる。

$$\begin{aligned}
\log L(\boldsymbol{\theta},\boldsymbol{\pi}) &= \sum_{i=1}^{N}\left[\delta_i \log\sum_{k=1}^{K}\pi_k f_k(t_i)+(1-\delta_i)\log\sum_{k=1}^{K}\pi_k S_k(t_i)\right] \\
&= \sum_{i=1}^{N}\left[\delta_i \log\sum_{k=1}^{K} q_k^{(i)}\frac{\pi_k f_k(t_i)}{q_k^{(i)}}+(1-\delta_i)\log\sum_{k=1}^{K} q_k^{(i)}\frac{\pi_k S_k(t_i)}{q_k^{(i)}}\right] \\
&\geq \sum_{i=1}^{N}\left[\sum_{k=1}^{K} q_k^{(i)}\delta_i \log\frac{\pi_k f_k(t_i)}{q_k^{(i)}}+\sum_{k=1}^{K} q_k^{(i)}(1-\delta_i)\log\frac{\pi_k S_k(t_i)}{q_k^{(i)}}\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\left[\delta_i \log\frac{\pi_k f_k(t_i)}{q_k^{(i)}}+(1-\delta_i)\log\frac{\pi_k S_k(t_i)}{q_k^{(i)}}\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\left[\delta_i \log f_k(t_i)+(1-\delta_i)\log S_k(t_i)+\log\pi_k-\log q_k^{(i)}\right]
\end{aligned}$$

この $q_k^{(i)}$ について、$\sum_{k=1}^{K} q_k^{(i)}=1$ の条件をつけて右辺を最大化するために、ラグランジュの未定定数法を用いる。

$$\frac{\partial}{\partial q_k^{(i)}}\left[\log L(\boldsymbol{\theta},\boldsymbol{\pi})-\sum_{i=1}^{N}\eta_i\left(\sum_{k=1}^{K}q_k^{(i)}-1\right)\right]$$
$$=\delta_i\log\frac{\pi_k f_k(t_i)}{q_k^{(i)}}+(1-\delta_i)\log\frac{\pi_k S_k(t_i)}{q_k^{(i)}}+1-\eta_i$$
$$=\log\frac{\pi_k f_k(t_i)^{\delta_i}S_k(t_i)^{1-\delta_i}}{q_k^{(i)}}+1-\eta_i=0$$

これより、

$$q_k^{(i)}=e^{1-\eta_i}\pi_k f_k(t_i)^{\delta_i}S_k(t_i)^{1-\delta_i}=\frac{\pi_k f_k(t_i)^{\delta_i}S_k(t_i)^{1-\delta_i}}{\sum_{k=1}^{K}\pi_k f_k(t_i)^{\delta_i}S_k(t_i)^{1-\delta_i}}$$

これを書き換えて、以下のようにすることもできる。

$$q_k^{(i)}=\pi_k f_k(t_i)\Big/\sum_{k=1}^{K}\pi_k f_k(t_i)\qquad\text{for }\ \delta_i=1$$

$$q_k^{(i)}=\pi_k S_k(t_i)\Big/\sum_{k=1}^{K}\pi_k S_k(t_i)\qquad\text{for }\ \delta_i=0$$

この $q_k^{(i)}$ を群 k への帰属度という。

この尤度関数をパラメータで微分して 0 と置き、パラメータの推定を行うが、$\sum_{k=1}^{K}\pi_k=1$ の条件をつけるために、ラグランジュの未定定数法を用いる。

$$\frac{\partial}{\partial\pi_j}\left[\log L(\boldsymbol{\theta},\boldsymbol{\pi})-\eta\left(\sum_{k=1}^{K}\pi_k-1\right)\right]=\sum_{i=1}^{N}q_j^{(i)}\Big/\pi_j-\eta=0$$

より、

$$\pi_k=\frac{1}{\eta}\sum_{i=1}^{N}q_k^{(i)},\qquad\sum_{k=1}^{K}\pi_k=\frac{1}{\eta}\sum_{k=1}^{K}\sum_{i=1}^{N}q_k^{(i)}=\frac{1}{\eta}\sum_{i=1}^{N}\sum_{k=1}^{K}q_k^{(i)}=\frac{1}{\eta}\sum_{i=1}^{N}1=\frac{N}{\eta}=1$$

となり、以下の関係を得る。

$$\pi_k=\frac{1}{N}\sum_{i=1}^{N}q_k^{(i)}$$

他のパラメータについては具体的な関数形を用いて考える。

混合指数分布に基づく最尤推定

指数分布の確率密度関数と分布関数の以下の具体的な表式を代入すると

$$f_k(t)=\lambda_k\exp(-\lambda_k t),\qquad S_k(t)=\exp(-\lambda_k t)$$

対数尤度は以下のようになる。

$$\log L(\boldsymbol{\lambda},\boldsymbol{\pi}) \geq \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\left[\delta_i\left(\log\lambda_k - \lambda_k t_i\right) - (1-\delta_i)\lambda_k t_i\right]$$
$$= \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\left[\delta_i \log\lambda_k - \lambda_k t_i\right]$$

これより、群 k への帰属度は以下となる。

$$q_k^{(i)} = \pi_k\lambda_k \exp(-\lambda_k t_i) \Big/ \sum_{k=1}^{K}\pi_k\lambda_k \exp(-\lambda_k t_i) \qquad \text{for } \delta_i = 1$$

$$q_k^{(i)} = \pi_k \exp(-\lambda_k t_i) \Big/ \sum_{k=1}^{K}\pi_k \exp(-\lambda_k t_i) \qquad \text{for } \delta_i = 0$$

$$\log L(\boldsymbol{\lambda},\boldsymbol{\pi}) \geq \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\left[\delta_i\left(\log\lambda_k - \lambda_k t_i\right) - (1-\delta_i)\lambda_k t_i + \log\pi_k - \log q_k^{(i)}\right]$$
$$= \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\left[\delta_i \log\lambda_k - \lambda_k t_i + \log\pi_k - \log q_k^{(i)}\right]$$

これを微分して、スコアベクトルを求め、それを 0 とする。

$$\frac{\partial}{\partial\lambda_j}\log L = \sum_{i=1}^{N} q_j^{(i)}\left(\delta_i/\lambda_j - t_i\right) = \frac{1}{\lambda_j}\sum_{i=1}^{N} q_j^{(i)}\delta_j - \sum_{i=1}^{N} q_j^{(i)} t_j = 0$$

これより、

$$\lambda_j = \sum_{i=1}^{N} q_j^{(i)}\delta_j \Big/ \sum_{i=1}^{N} q_j^{(i)} t_j$$

スコアをもう一度微分して、情報行列 $\mathfrak{I}$ に相当するものを作成する。

$$\mathfrak{I}_{jk} = -\frac{\partial^2}{\partial\lambda_j\partial\lambda_k}\log L = \frac{\delta_{jk}}{\lambda_j^2}\sum_{i=1}^{N} q_j^{(i)}\delta_i$$

この逆行列の対角成分は、推定値の分散を与える。

混合ワイブル分布に基づく最尤推定

K 種混合ワイブル分布では、以下となる。

$$f(t) = \sum_{k=1}^{K}\pi_k f_k(t) = \sum_{k=1}^{K}\pi_k a_k t^{a_k-1} b_k^{-a_k} \exp(-t^{a_k} b_k^{-a_k}) = \sum_{k=1}^{K}\pi_k a_k t^{a_k-1} e^{\beta_k} \exp(-t^{a_k} e^{\beta_k})$$

$$S(t) = \sum_{k=1}^{K}\pi_k S_k(t) = \sum_{k=1}^{K}\pi_k \exp(-t^{a_k} b_k^{-a_k}) = \sum_{k=1}^{K}\pi_k \exp(-t^{a_k} e^{\beta_k})$$

混合ワイブル分布の対数尤度は以下となる。

$$\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi}) \geq \sum_{i=1}^{N}\sum_{k=1}^{K} q_k^{(i)}\left[\delta_i\left(\log a_k + (a_k-1)\log t_i + \beta_k\right) - t_i^{a_k}e^{\beta_k} + \log \pi_k - \log q_k^{(i)}\right]$$

これより、群 k への帰属度は以下となる。

$$q_k^{(i)} = \frac{\pi_k a_k t^{a_k-1} e^{\beta_k}\exp(-t^{a_k}e^{\beta_k})}{\sum_{k=1}^{K}\pi_k a_k t^{a_k-1} e^{\beta_k}\exp(-t^{a_k}e^{\beta_k})} \qquad \text{for } \delta_i = 1$$

$$q_k^{(i)} = \frac{\pi_k \exp(-t^{a_k}e^{\beta_k})}{\sum_{k=1}^{K}\pi_k \exp(-t^{a_k}e^{\beta_k})} \qquad \text{for } \delta_i = 0$$

ここで、$b_k^{-a_k} = e^{\beta_k}$ （$b_k = e^{-\beta_k/a_k}$ に相当）としている。

$$\frac{\partial}{\partial a_j}\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi}) = \sum_{i=1}^{N} q_j^{(i)}\left[\delta_i\left(1/a_j + \log t_i\right) - \log t_i\, t_i^{a_j}e^{\beta_j}\right]$$

$$\frac{\partial}{\partial \beta_j}\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi}) = \sum_{i=1}^{N} q_j^{(i)}\left[\delta_i - t_i^{a_j}e^{\beta_j}\right]$$

$$\frac{\partial^2}{\partial a_j \partial a_k}\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi}) = -\delta_{jk}\sum_{i=1}^{N} q_j^{(i)}\left[\delta_i/a_j^2 + (\log t_i)^2\, t_i^{a_j}e^{\beta_j}\right]$$

$$\frac{\partial^2}{\partial a_j \partial \beta_k}\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi}) = -\delta_{jk}\sum_{i=1}^{N} q_j^{(i)}\log t_i\, t_i^{a_j}e^{\beta_j}$$

$$\frac{\partial^2}{\partial \beta_j \partial \beta_k}\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\pi}) = -\delta_{jk}\sum_{i=1}^{N} q_j^{(i)} t_i^{a_j}e^{\beta_j}$$

3.4 比例ハザードモデル

比例ハザードモデルはハザート関数に対して以下の仮定を行う。

$$\lambda(t\,|\,\mathbf{x},\boldsymbol{\beta}) = \lambda_0(t)\exp({}^t\mathbf{x}\boldsymbol{\beta}) \quad \text{ここに、} {}^t\mathbf{x}\boldsymbol{\beta} = \sum_{i=1}^{p} x_i\beta_i + \beta_0$$

Cox の比例ハザードモデルでは $\lambda_0(t)$ と定数項 β_0 について議論しないが、ワイブル比例ハザードモデルでは

$$\lambda(t\,|\,\mathbf{x},\boldsymbol{\beta}) = (a/b)\left(t/b\right)^{a-1} = at^{a-1}b^{-a} = at^{a-1}\exp({}^t\mathbf{x}\boldsymbol{\beta})$$

として、時間に関してワイブル分布のハザード関数を仮定する。

Cox の比例ハザードモデル

Cox の比例ハザードモデルでは、尤度関数に対して近似的な部分尤度関数を考えて処理を行う。その

対数尤度は以下で与えられる[3]。

$$\log L'(\boldsymbol{\beta}) = \sum_{i=0}^{m-1}\left[\sum_{j\in D_i} {}^t\mathbf{x}_j\boldsymbol{\beta} - d_i \log\sum_{j\in R_i}\exp({}^t\mathbf{x}_j\boldsymbol{\beta})\right]$$

ここに、$\boldsymbol{\beta}$ は定数項を除いた偏回帰係数ベクトル、D_i は $t_i < t \le t_{i+1}$ で亡くなった個体の集合、R_i は時刻 t_i で生存が確認されている個体の集合である。これを最大化するようにニュートン・ラフソン法を使って $\boldsymbol{\beta}$ を求める。ここではそのための準備として以下の値を示しておく。

$$\mathbf{U} \equiv \frac{\partial}{\partial\boldsymbol{\beta}}\log L'(\boldsymbol{\beta}) = \sum_{i=1}^{m-1}\left[\sum_{i\in D_i}\mathbf{x}_j - d_i\sum_{j\in R_i} w_j\mathbf{x}_j \Big/ \sum_{j\in R_i} w_j\right]$$

$$\mathfrak{I} \equiv -\frac{\partial^2}{\partial\boldsymbol{\beta}\partial{}^t\boldsymbol{\beta}}\log L'(\boldsymbol{\beta}) = \sum_{i=1}^{m-1} d_i\left[\sum_{j\in R_i} w_j\mathbf{x}_j\,{}^t\mathbf{x}_j \Big/ \sum_{j\in R_i} w_j - \sum_{j\in R_i} w_j\mathbf{x}_j \sum_{j\in R_i} w_j\,{}^t\mathbf{x}_j \Big/ \left(\sum_{j\in R_i} w_j\right)^2\right]$$

ここに $w_j = \exp({}^t\mathbf{x}_j\boldsymbol{\beta})$

この $\mathbf{U}$ をスコアベクトル、$\mathfrak{I}$ を情報行列という。$\boldsymbol{\beta}$ の推定値は以下の計算を繰り返して求める。

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathfrak{I}^{(m)})^{-1}\mathbf{U}^{(m)}$$

ここに右肩の添え字はニュートン・ラフソン法のループの段階を表している。

ワイブル比例ハザードモデル

ワイブル比例ハザードモデルは、ハザード関数に対して以下の仮定を行う。

$$\lambda(t) = \frac{f(t)}{S(t)} = (a/b)\left(t/b\right)^{a-1} = at^{a-1}/b^a = at^{a-1}\exp({}^t\mathbf{x}\boldsymbol{\beta})$$

通常のワイブル分布との関係は以下である。

$$b^{-a} = e^{\beta} \to \exp({}^t\mathbf{x}\boldsymbol{\beta}) \quad (\beta \to {}^t\mathbf{x}\boldsymbol{\beta} \equiv \sum_{i=1}^{p} x_i\beta_i + \beta_0)$$

これより、$b = \exp(-{}^t\mathbf{x}\boldsymbol{\beta}/a)$ であるから、$\mu \equiv E[T] = b\,\Gamma\left(1+1/a\right)$ より、

$$\eta \equiv {}^t\mathbf{x}\boldsymbol{\beta} = -a\log b = -a\log\left(\mu/\Gamma(1+1/a)\right)$$

となり、右辺が一般化線形モデルの連結関数となる。

この関係を用いて、累積生存関数と密度関数を求めると以下となる。

$$S(t) = \exp\left[-\left(t/b\right)^a\right] = \exp\left[-t^a b^{-a}\right] = \exp\left[-t^a\exp({}^t\mathbf{x}\boldsymbol{\beta})\right]$$

$$f(t) = -at^{a-1}\exp({}^t\mathbf{x}\boldsymbol{\beta})\exp\left[-t^a\exp({}^t\mathbf{x}\boldsymbol{\beta})\right]$$

打ち切りデータと非打ち切りデータをそれぞれ $\delta_i = 0,1$ と区別し、尤度を求めると以下となる。添え字 i について、ここでは個体の番号として使っている。

$$L(\alpha,\boldsymbol{\beta}) = \prod_{i=1}^{N} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

さらに、対数尤度は以下となる。

$$\begin{aligned}\log L(\alpha,\boldsymbol{\beta}) &= \sum_{i=1}^{N}\left[\delta_i \log\lambda(t_i) + \log S(t_i)\right] \\ &= \sum_{i=1}^{N}\left[\delta_i \log\left(a t_i^{a-1}\exp({}^t\mathbf{x}_i\boldsymbol{\beta})\right) - t_i^a \exp({}^t\mathbf{x}_i\boldsymbol{\beta})\right] \\ &= \sum_{i=1}^{N}\left[\delta_i \left(\log a + (a-1)\log t_i + {}^t\mathbf{x}_i\boldsymbol{\beta}\right) - t_i^a \exp({}^t\mathbf{x}_i\boldsymbol{\beta})\right]\end{aligned}$$

これを微分すると

$$\begin{aligned}\frac{\partial}{\partial a}\log L &= \sum_{i=1}^{N}\left[\delta_i(1/a + \log t_i) - \log t_i t_i^a \exp({}^t\mathbf{x}_i\boldsymbol{\beta})\right] \\ \frac{\partial}{\partial \boldsymbol{\beta}}\log L &= \sum_{i=1}^{N}\left[\delta_i\mathbf{x}_i - \mathbf{x}_i t_i^a \exp({}^t\mathbf{x}_i\boldsymbol{\beta})\right] \\ \frac{\partial^2}{\partial a^2}\log L &= \sum_{i=1}^{N}\left[-\delta_i/a^2 - (\log t_i)^2 t_i^a \exp({}^t\mathbf{x}_i\boldsymbol{\beta})\right] \\ \frac{\partial^2}{\partial a\partial \boldsymbol{\beta}}\log L &= -\sum_{i=1}^{N}(\log t_i)\mathbf{x}_i t_i^a \exp({}^t\mathbf{x}_i\boldsymbol{\beta}) \\ \frac{\partial^2}{\partial \boldsymbol{\beta}\partial {}^t\boldsymbol{\beta}}\log L &= -\sum_{i=1}^{N}\mathbf{x}_i\,{}^t\mathbf{x}_i t_i^a \exp({}^t\mathbf{x}_i\boldsymbol{\beta})\end{aligned}$$

これらを用いてスコアベクトル $\mathbf{U}$ と情報行列 $\mathfrak{I}$ を以下のように定義する。

$$\boldsymbol{\beta}' = \begin{pmatrix} a \\ \boldsymbol{\beta} \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} \partial\log L/\partial a \\ \partial\log L/\partial\boldsymbol{\beta} \end{pmatrix}, \quad \mathfrak{I} = -\begin{pmatrix} \partial^2\log L/\partial a^2 & \partial^2\log L/\partial a\partial {}^t\boldsymbol{\beta} \\ \partial^2\log L/\partial a\partial\boldsymbol{\beta} & \partial^2\log L/\partial\boldsymbol{\beta}\partial {}^t\boldsymbol{\beta} \end{pmatrix}$$

これらを用いてニュートン・ラフソン法でパラメータの推定を行う。

$$\boldsymbol{\beta}'^{(m+1)} = \boldsymbol{\beta}'^{(m)} + (\mathfrak{I}^{(m)})^{-1}\mathbf{U}^{(m)}$$

ここに右肩の添え字はニュートン・ラフソン法のループの段階を表している。

求められたパラメータを使って、個人の予想寿命を以下のように求めることができる。

$$\mu \equiv E[T] = b\,\Gamma\left(1+1/a\right) = \exp(-{}^t\mathbf{x}\boldsymbol{\beta}/a)\Gamma\left(1+1/a\right)$$

この値を実際の寿命と比較することで相関係数等を求めることもできる。

混合ワイブル比例ハザードモデル

K 種混合ワイブル比例ハザードモデルでは以下を仮定する。

$$f(t)=\sum_{k=1}^{K}\pi_k f_k(t)=\sum_{k=1}^{K}\pi_k a_k t^{a_k-1}\exp({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k)\exp(-t^{a_k}\exp({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k))$$

$$S(t)=\sum_{k=1}^{K}\pi_k S_k(t)=\sum_{k=1}^{K}\pi_k \exp(-t^{a_k}\exp({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k))$$

通常のワイブル分と比較すると、ここでは以下を仮定している。

$$b_k^{-a_k}=e^{\beta_k}\rightarrow\exp({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k)\quad(\beta_k\rightarrow{}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k\equiv\sum_{i=1}^{p}x_i\beta_i+\gamma_k)$$

これより、$b_k\rightarrow\exp\left[-({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k)/a_k\right]$ であるから、

$$\mu\equiv E[T]=\sum_{k=1}^{K}\pi_k b_k\Gamma\left(1+1/a_k\right)$$

となる。連結関数については、以下の関数の逆関数である。

$$\mu=\sum_{k=1}^{K}\pi_k\exp\left[-({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k)/a_k\right]\Gamma\left(1+1/a_k\right)$$

$$=\sum_{k=1}^{K}\pi_k\exp\left[-(\eta+\gamma_k)/a_k\right]\Gamma\left(1+1/a_k\right)$$

混合ワイブル分布の対数尤度は以下となる。

$$\log L(\mathbf{a},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\pi})$$

$$\geq\sum_{i=1}^{N}\sum_{k=1}^{K}q_k^{(i)}\left[\delta_i\left(\log a_k+(a_k-1)\log t_i+{}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_k\right)-t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_k)+\log\pi_k-\log q_k^{(i)}\right]$$

これより、群 k への帰属度は以下となる。

$$q_k^{(i)}=\frac{\pi_k a_k t_i^{a_k-1}\exp({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k)\exp(-t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_k))}{\sum_{k=1}^{K}\pi_k a_k t_i^{a_k-1}\exp({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k)\exp(-t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_k))}\qquad\text{for }\delta_i=1$$

$$q_k^{(i)}=\frac{\pi_k\exp(-t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_k))}{\sum_{k=1}^{K}\pi_k\exp(-t_i^{a_k}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_k))}\qquad\text{for }\delta_i=0$$

ここで、$b_k^{-a_k}\rightarrow\exp({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k)$（$b_k\rightarrow\exp\left[-({}^t\mathbf{x}\boldsymbol{\beta}+\gamma_k)/a_k\right]$）としている。

対数尤度をパラメータで微分すると

$$\frac{\partial}{\partial a_j}\log L=\sum_{i=1}^{N}q_j^{(i)}\left[\delta_i\left(1/a_j+\log t_i\right)-\log t_i\, t_i^{a_j}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_j)\right]$$

$$\frac{\partial}{\partial\gamma_j}\log L=\sum_{i=1}^{N}q_j^{(i)}\left[\delta_i-t_i^{a_j}\exp({}^t\mathbf{x}_i\boldsymbol{\beta}+\gamma_j)\right]$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log L = \sum_{i=1}^{N} \sum_{k=1}^{K} q_k^{(i)} \left[\delta_i \mathbf{x}_i - \mathbf{x}_i t_i^{a_k} \exp({}^t \mathbf{x}_i \boldsymbol{\beta} + \gamma_j) \right]$$

$$\frac{\partial^2}{\partial a_j \partial a_k} \log L = -\delta_{jk} \sum_{i=1}^{N} q_j^{(i)} \left[\delta_i / a_j^2 + (\log t_i)^2 \, t_i^{a_j} \exp({}^t \mathbf{x}_i \boldsymbol{\beta} + \gamma_j) \right]$$

$$\frac{\partial^2}{\partial a_j \partial \gamma_k} \log L = -\delta_{jk} \sum_{i=1}^{N} q_j^{(i)} \log t_i \, t_i^{a_j} \exp({}^t \mathbf{x}_i \boldsymbol{\beta} + \gamma_j)$$

$$\frac{\partial^2}{\partial a_j \partial \boldsymbol{\beta}} \log L = -\sum_{i=1}^{N} q_j^{(i)} \log t_i \, \mathbf{x}_i t_i^{a_j} \exp({}^t \mathbf{x}_i \boldsymbol{\beta} + \gamma_j)$$

$$\frac{\partial^2}{\partial \gamma_j \partial \gamma_k} \log L = -\delta_{jk} \sum_{i=1}^{N} q_j^{(i)} t_i^{a_j} \exp({}^t \mathbf{x}_i \boldsymbol{\beta} + \gamma_j)$$

$$\frac{\partial^2}{\partial \gamma_j \partial \boldsymbol{\beta}} \log L = -\sum_{i=1}^{N} q_j^{(i)} \mathbf{x}_i t_i^{a_j} \exp({}^t \mathbf{x}_i \boldsymbol{\beta} + \gamma_j)$$

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial {}^t \boldsymbol{\beta}} \log L = -\sum_{i=1}^{N} \sum_{k=1}^{K} q_k^{(i)} \mathbf{x}_i \, {}^t \mathbf{x}_i t_i^{a_k} \exp({}^t \mathbf{x}_i \boldsymbol{\beta} + \gamma_k)$$

これより、スコアベクトル $\mathbf{U}$ と情報行列 $\mathfrak{I}$ を以下のように定義する。

$$\boldsymbol{\beta}' = \begin{pmatrix} \mathbf{a} \\ \boldsymbol{\gamma} \\ \boldsymbol{\beta} \end{pmatrix}, \qquad \mathbf{U} = \begin{pmatrix} \partial \mathrm{l} \circ \mathrm{g} / \partial \mathbf{a} \\ \partial \mathrm{l} \circ \mathrm{g} / \partial \boldsymbol{\gamma} \\ \partial \mathrm{l} \circ \mathrm{g} / \partial \boldsymbol{\beta} \end{pmatrix},$$

$$\mathfrak{I} = -\begin{pmatrix} \partial^2 \log L / \partial \mathbf{a} \partial^t \mathbf{a} & \partial^2 \log L / \partial \mathbf{a} \partial^t \boldsymbol{\gamma} & \partial^2 \log L / \partial \mathbf{a} \partial^t \boldsymbol{\beta} \\ \partial^2 \log L / \partial \boldsymbol{\gamma} \partial^t \mathbf{a} & \partial^2 \log L / \partial \boldsymbol{\gamma} \partial^t \boldsymbol{\gamma} & \partial^2 \log L / \partial \boldsymbol{\gamma} \partial^t \boldsymbol{\beta} \\ \partial^2 \log L / \partial \boldsymbol{\beta} \partial^t \mathbf{a} & \partial^2 \log L / \partial \boldsymbol{\beta} \partial^t \boldsymbol{\gamma} & \partial^2 \log L / \partial \boldsymbol{\beta} \partial^t \boldsymbol{\beta} \end{pmatrix}$$

これらを用いてニュートン・ラフソン法でパラメータの推定を行う。

$$\boldsymbol{\beta}'^{(m+1)} = \boldsymbol{\beta}'^{(m)} + (\mathfrak{I}^{(m)})^{-1} \mathbf{U}^{(m)}$$

ここに右肩の添え字はニュートン・ラフソン法のループの段階を表している。

4. プログラムの利用法

メニュー［分析－多変量解析他－生存時間分析］を選択すると、図 1 のような分析実行メニューが表示される。

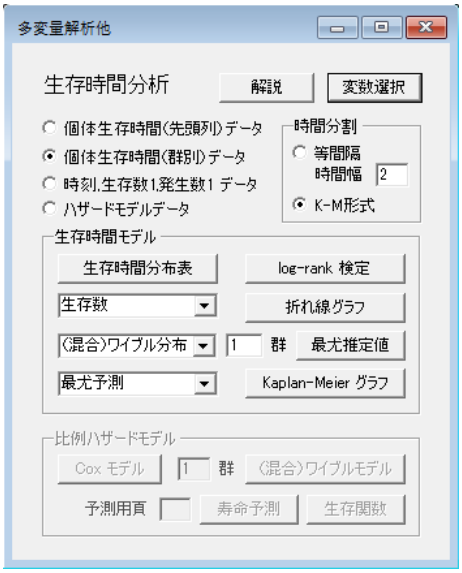


図 1　生存時間分析実行メニュー

この分析のデータ形式は大きく分けて 3 種類ある。1 つは個体の生存時間を元にしたデータで、先頭列で分類される形式とすでに群別に並べられている形式に分けられる。これらの形式は基本統計のデータ形式に類似している。次に、すでに生命表に近い形式になっているデータである。これは、観測時刻、その時点での生存個体数、その時点より後で次の時点までに死亡する期間発生数が、すでに表の形式になっているデータである。生存個体数と期間発生数は複数組入力が可能である。詳しくはサンプルを見てもらいたい。最後は、ハザードモデルデータで、重回帰分析などと同様の形式である。最初と最後の形式で、通常のデータと異なる部分は、観測の打ち切りデータが含まれる点である。打ち切りデータは、観測を打ち切られた時点の数値の後ろに+記号を付けて表す。観測が打ち切られた際の扱いは、生存数から打ち切られたデータ数の半分を引いて、死亡リスクに晒されたデータ数として処理している[1]。

最初に図 2 の単独データを元に説明をする。

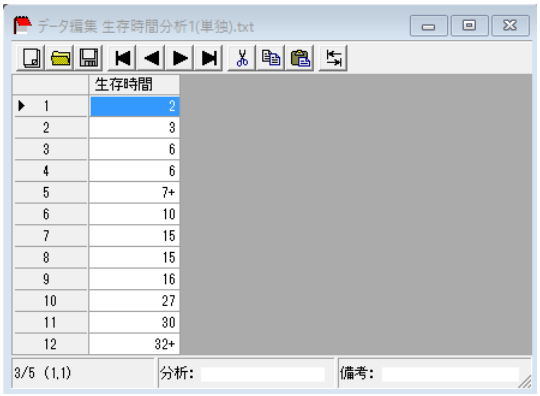
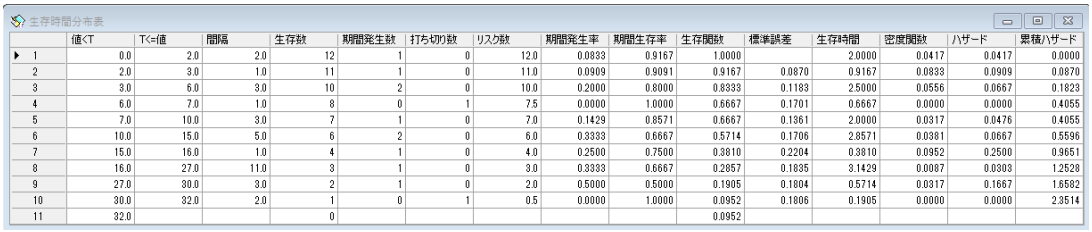


図 2　単独データ（生存時間分析 1(単独).txt 3 頁目）

このデータでは、2 個体が観測を打ち切られている。

「個体生存時間(群別)データ」ラジオボタンを選択し、変数選択を実行して、「生存時間分布表」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。



生存時間分布表

	値<T	T<=値	間隔	生存数	期間発生数	打ち切り数	リスク数	期間発生率	期間生存率	生存関数	標準誤差	生存時間	密度関数	ハザード	累積ハザード
1	0.0	2.0	2.0	12	1	0	12.0	0.0833	0.9167	1.0000		2.0000	0.0417	0.0417	0.0000
2	2.0	3.0	1.0	11	1	0	11.0	0.0909	0.9091	0.9167	0.0870	0.9167	0.0833	0.0909	0.0870
3	3.0	6.0	3.0	10	2	0	10.0	0.2000	0.8000	0.8333	0.1183	2.5000	0.0556	0.0667	0.1823
4	6.0	7.0	1.0	8	0	1	7.5	0.0000	1.0000	0.6667	0.1701	0.6667	0.0000	0.0000	0.4055
5	7.0	10.0	3.0	7	1	0	7.0	0.1429	0.8571	0.6667	0.1361	2.0000	0.0317	0.0476	0.4055
6	10.0	15.0	5.0	6	2	0	6.0	0.3333	0.6667	0.5714	0.1706	2.8571	0.0381	0.0667	0.5596
7	15.0	16.0	1.0	4	1	0	4.0	0.2500	0.7500	0.3810	0.2204	0.3810	0.0952	0.2500	0.9651
8	16.0	27.0	11.0	3	1	0	3.0	0.3333	0.6667	0.2857	0.1835	3.1429	0.0087	0.0303	1.2528
9	27.0	30.0	3.0	2	1	0	2.0	0.5000	0.5000	0.1905	0.1804	0.5714	0.0317	0.1667	1.6582
10	30.0	32.0	2.0	1	0	1	0.5	0.0000	1.0000	0.0952	0.1806	0.1905	0.0000	0.0000	2.3514
11	32.0			0						0.0952					

図 3　生存時間分布表結果

図 3 では、様々な指標が区切られた時点毎に表示されている。ここで特に大切な指標は、「生存関数」と「ハザード」である。これらはそれぞれ、その時点まで生存している確率とその時点での死亡の危険率の意味を持つ。

図 3a の生存時間分布表の中で、生存数、累積生存関数、ハザード関数、累積ハザード関数については、コンボボックスで設定して、「折れ線グラフ」ボタンをクリックすると表示される。ここでは累積生存関数とハザード関数についてのグラフを図 4a と図 4b に示す。

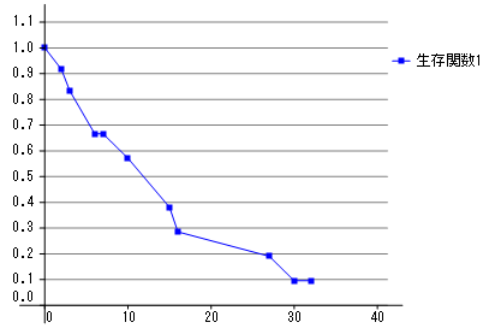


図 4a　生存関数

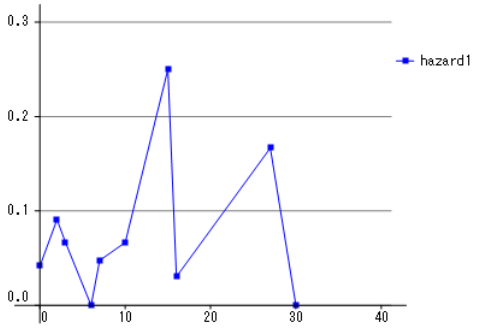


図 4b　ハザード関数

また、同じコンボボックスで「指数分布確認」または「ワイブル分布確認」を選択すると、図 5a と図 5b のような図が表示される。

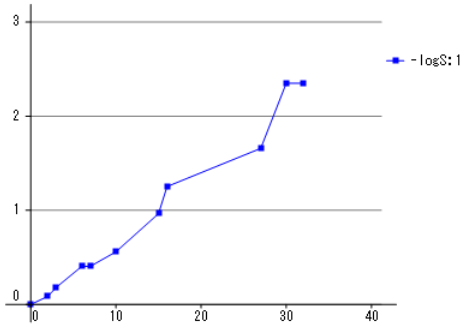


図 5a　累積生存関数

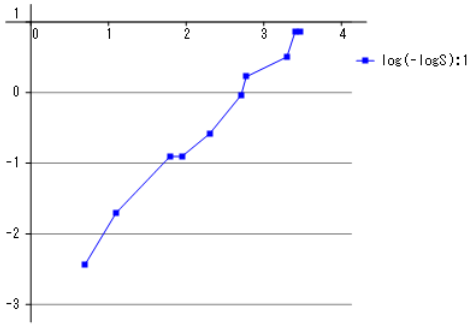


図 5b　ハザード関数

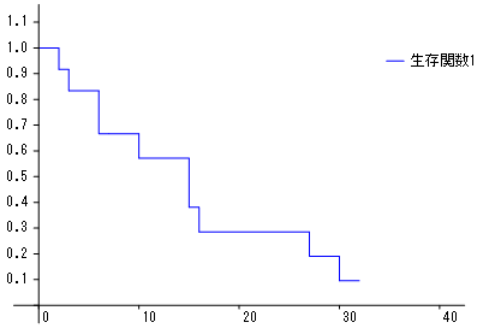
生存時間が指数分布またはワイブル分布に従うならば、それぞれの累積生存関数の時間依存性からこの点列は直線状に並ぶ。指数分布はワイブル分布の特殊な場合であるので、指数分布が成り立つ場合はワイブル分布も成り立つ。

分布の確認の場合、「折れ線グラフ」をクリックすると、上図と共に分布の当てはまりの良さを示す、図 6a や図 6b のような指標も表示される。

図 6a　指数分布の指標

図 6b　ワイブル分布の指標

生存時間関数の Kaplan-Meier 推定のグラフは、「Kaplan-Meier グラフ」ボタンをクリックして表示される。その際、左のコンボボックスで指定して、指数分布またはワイブル分布の予想曲線を描くこともできる。予想曲線のないグラフと、ワイブル分布の予想曲線を付けて描いたグラフを図 7a と図 7b に示す。



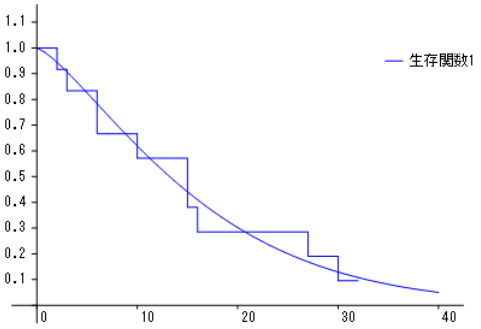


図 7a　Kaplan-Meier 生存関数グラフ　図 7b　予想曲線付き Kaplan-Meier グラフ

これらの予想曲線では最小 2 乗法によるものと最尤法によるものとが選択できる。上図は最尤法によるものである。

また、予想曲線は混合指数分布や混合ワイブル分布についても表示することができる。その際は分布を選んだコンボボックスの右のテキストボックスで混合する数を指定する。図 8 に 2 群の混合ワイブル分布による予測曲線を付けた Kaplan-Meier グラフを表示する。サンプルでは 2 つの時期に危険度が高くなっている。

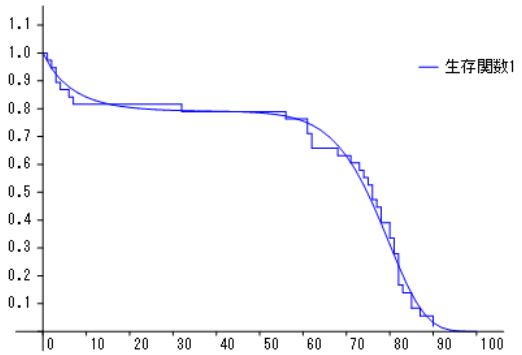


図 8　2 群混合分布による予測（生存時間分析 1(単独).txt 8 頁目）

このパラメータの値については、上と同じ設定で「最尤推定値」ボタンをクリックすると、図 9 のように表示される。

混合ワイブル分布推定結果

	推定値	標準偏差	5%下限	5%上限
▶ 生存時間	R	0.993	R^2	0.986
出現確率1	0.790			
a1	10.545	0.000	10.545	10.545
b1=exp(-β/a)	80.564			
β1	-46.283	0.000	-46.283	-46.283
出現確率2	0.210			
a2	0.937	0.000	0.937	0.937
b2=exp(-β/a)	6.964			
β2	-1.819	0.000	-1.819	-1.819

図 9　2 群混合ワイブル予測（生存時間分析 1(単独).txt 8 頁目）

ここでは表示されていないが、混合がない場合には、右端に最小 2 乗推定による推定値も表示される。

複数群の生存時間分布表は、先頭列で群分けデータ（生存時間分析 2(2 群比較).txt）または群別データを元に図 10 のように縦に並べて表示される。

生存時間分布表

	値<T	T<=値	間隔	生存数	期間発生数	打ち切り数	リスク数	期間発生率	期間生存率	生存関数	標準誤差	生存時間	密度関数	ハザード	累積ハザード
1	0.0	1.0	1.0	12	0	0	12.0	0.0000	1.0000	1.0000		1.0000	0.0000	0.0000	0.0000
2	1.0	2.0	1.0	12	1	0	12.0	0.0833	0.9167	1.0000	0.0000	1.0000	0.0833	0.0833	0.0000
3	2.0	3.0	1.0	11	1	0	11.0	0.0909	0.9091	0.9167	0.0870	0.9167	0.0833	0.0909	0.0870
4	3.0	6.0	3.0	10	2	0	10.0	0.2000	0.8000	0.8333	0.1183	2.5000	0.0556	0.0667	0.1823
5	6.0	7.0	1.0	8	1	0	8.0	0.1250	0.8750	0.6667	0.1701	0.6667	0.0833	0.1250	0.4055
6	7.0	9.0	2.0	7	0	0	7.0	0.0000	1.0000	0.5833	0.1627	1.1667	0.0000	0.0000	0.5390
7	9.0	10.0	1.0	7	1	0	7.0	0.1429	0.8571	0.5833	0.1423	0.5833	0.0833	0.1429	0.5390
8	10.0	15.0	5.0	6	2	0	6.0	0.3333	0.6667	0.5000	0.1684	2.5000	0.0333	0.0667	0.6931
9	15.0	16.0	1.0	4	1	0	4.0	0.2500	0.7500	0.3333	0.2041	0.3333	0.0833	0.2500	1.0986
10	16.0	22.0	6.0	3	0	0	3.0	0.0000	1.0000	0.2500	0.1667	1.5000	0.0000	0.0000	1.3863
11	22.0	27.0	5.0	3	1	0	3.0	0.3333	0.6667	0.2500	0.1250	1.2500	0.0167	0.0667	1.3863
12	27.0	30.0	3.0	2	1	0	2.0	0.5000	0.5000	0.1667	0.1614	0.5000	0.0278	0.1667	1.7918
13	30.0	32.0	2.0	1	1	0	1.0	1.0000	0.0000	0.0000	0.1596	0.0000	0.0000	0.5000	0.0000
14	32.0			0						0.0000					
1	0.0	1.0	1.0	9	4	0	9.0	0.4444	0.5556	1.0000		1.0000	0.4444	0.4444	0.0000
2	1.0	2.0	1.0	5	1	0	5.0	0.2000	0.8000	0.5556	0.2981	0.5556	0.1111	0.2000	0.5878
3	2.0	3.0	1.0	4	2	0	4.0	0.5000	0.5000	0.4444	0.2070	0.4444	0.2222	0.5000	0.8109
4	3.0	6.0	3.0	2	0	0	2.0	0.0000	1.0000	0.2222	0.2772	0.6667	0.0000	0.0000	1.5041
5	6.0	7.0	1.0	2	0	0	2.0	0.0000	1.0000	0.2222	0.1386	0.2222	0.0000	0.0000	1.5041
6	7.0	9.0	2.0	2	1	0	2.0	0.5000	0.5000	0.2222	0.1386	0.4444	0.0556	0.2500	1.5041
7	9.0	10.0	1.0	1	0	0	1.0	0.0000	1.0000	0.1111	0.2095	0.1111	0.0000	0.0000	2.1972
8	10.0	15.0	5.0	1	0	0	1.0	0.0000	1.0000	0.1111	0.1048	0.5556	0.0000	0.0000	2.1972
9	15.0	16.0	1.0	1	0	0	1.0	0.0000	1.0000	0.1111	0.1048	0.1111	0.0000	0.0000	2.1972
10	16.0	22.0	6.0	1	1	0	1.0	1.0000	0.0000	0.0000	0.1048	0.0000	0.0000	0.1667	0.0000
11	22.0	27.0	5.0	0						0.0000					
12	27.0	30.0	3.0												
13	30.0	32.0	2.0												
14	32.0														

図 10　2 群の生存時間分布表

これ以外に、もっと群の違いを比較できる方法を考えて行きたい。

複数群の累積生存関数と Kaplan-Meire 累積生存関数グラフを図 11 と図 12 に示す。

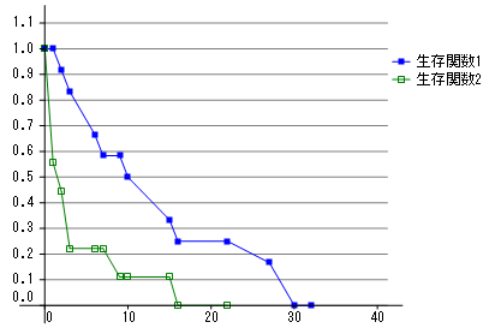


図 11　2 種類の累積生存関数グラフ

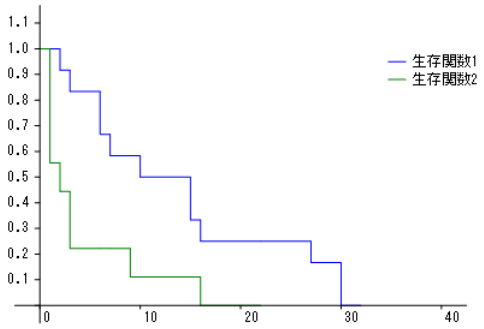


図 12　2 種類の Kaplan-Meier グラフ

複数群の累積生存関数間の差の log-rank 検定結果は、「log-rank 検定」ボタンをクリックすると図 13 のように表示される。

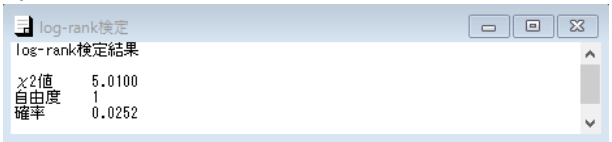


図 13　log-rank 検定結果

最後に、比例ハザードモデルの分析結果について示しておく。データは図 14 のような重回帰分析などと同じデータ形式である。

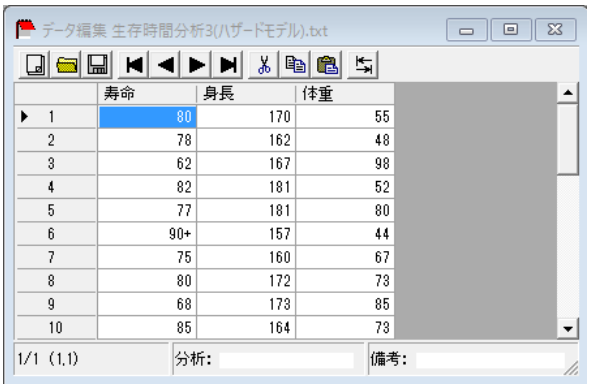


図 14　比例ハザードモデルデータ（生存時間分析 3(ハザードモデル).txt）

ハザードモデルでは Cox 比例ハザードモデルと Weibull 比例ハザードモデルを組み込んでいる。ハザード関数について、2 つのモデルとも以下の形を仮定する。

$$\lambda(t \mid \mathbf{x}, \boldsymbol{\beta}) = \lambda_0(t)\exp({}^t\mathbf{x}\boldsymbol{\beta}) \quad \text{ここに、} {}^t\mathbf{x}\boldsymbol{\beta} = \sum_{i=1}^{p} x_i\beta_i + \beta_0$$

Cox 比例ハザードモデルは $\lambda_0(t)$ や β_0 の推定は行わないが、分布の形に依存しない利点がある。Weibull ハザードモデルでは、時間部分にワイブル分布を仮定し、その 1 つのパラメータを説明変数で推定するという一般化線形モデルの形式を採用している。

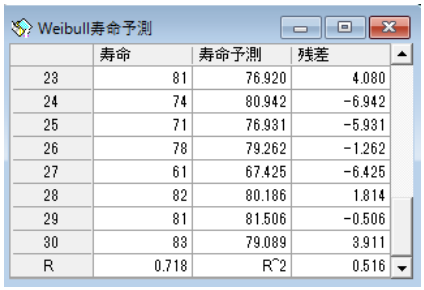
$$\lambda(t \mid \mathbf{x}, \boldsymbol{\beta}) = (a/b)\left(t/b\right)^{a-1} = at^{a-1}b^{-a} = at^{a-1}\exp({}^t\mathbf{x}\boldsymbol{\beta})$$

「Cox モデル」ボタンをクリックした結果を図 12 に、「Weibull モデル」ボタンをクリックした結果を図 15 に示す。

図 15　Cox 比例ハザードモデル結果

図 16　Weibull 比例ハザードモデル

最後に Weibull 比例ハザードモデルが予想する生存時間の平均値と実際の観測値との比較を行ってみる。「寿命予測」ボタンをクリックすると図 17a と図 17b の結果が示される。



Weibull寿命予測

	寿命	寿命予測	残差
23	81	76.920	4.080
24	74	80.942	-6.942
25	71	76.931	-5.931
26	78	79.262	-1.262
27	61	67.425	-6.425
28	82	80.186	1.814
29	81	81.506	-0.506
30	83	79.089	3.911
R	0.718	R^2	0.516

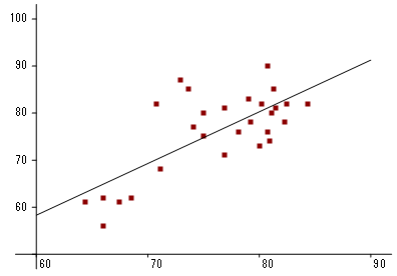


図 17a　寿命予測図　　17b　実測/予測散布図

これには非打ち切りデータのみが用いられている。また、寿命予測の結果の最後に、予測値と実測値の相関係数の値とその 2 乗の値を表示している。

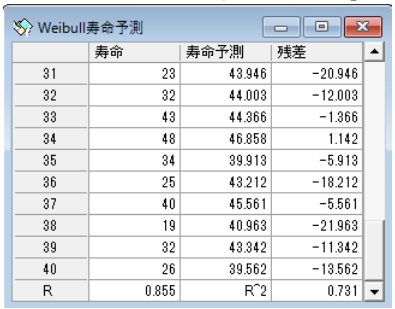
2 種混合ワイブルハザードモデルの場合、比例ハザードモデルの中の「群」テキストボックスに 2 を入れて、「(混合) ワイブルモデル」ボタンをクリックする。図 18 に結果を示す。

複合ワイブル比例ハザードモデル結果

	偏回帰係数	標準偏差	両側確率	2.5%下限	2.5%上限	EXP(b)
▶ 身長	0.0274	0.7407	0.9705	-1.4244	1.4792	1.028E00
体重	0.0526	0.7407	0.9434	-1.3992	1.5043	1.054E00
要因	5.2765	0.7407	0.0000	3.8248	6.7283	1.957E02
出現確率1	0.2497					
a1	29.7025	5.6218		18.6837	40.7212	
γ1	-140.1803	25.9021		-190.9484	-89.4123	
出現確率2	0.7503					
a2	7.3947	1.0264		5.3829	9.4065	
γ2	-40.1365	6.2861		-52.4574	-27.8157	

図 18　混合ワイブルハザードモデル（生存時間分析 3(ハザードモデル).txt 2 頁目）

このモデルによる実測・予測値と重相関係数 R の値、及びそのグラフを表示するには、「予測用頁」テキストボックスを空欄のまま、「寿命予測」ボタンをクリックする。結果は図 19 のようになる。



Weibull寿命予測

	寿命	寿命予測	残差
31	23	43.946	-20.946
32	32	44.003	-12.003
33	43	44.366	-1.366
34	48	46.858	1.142
35	34	39.913	-5.913
36	25	43.212	-18.212
37	40	45.561	-5.561
38	19	40.963	-21.963
39	32	43.342	-11.342
40	26	39.562	-13.562
R	0.855	R^2	0.731

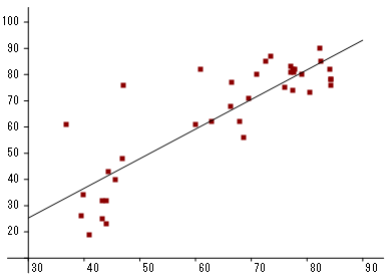


図 19　混合モデルによる実測・予測値

このモデルと混合ワイブル分布の Kaplan-Meier 推定とを比較してみる。寿命予測するページを現在のページ（空欄も可）にして「生存関数」ボタンをクリックし、各個体の生存関数を描画すると図 20 のようになる。また混合ワイブル分布を使った Kaplan-Meier 推定は図 21 のようになる。

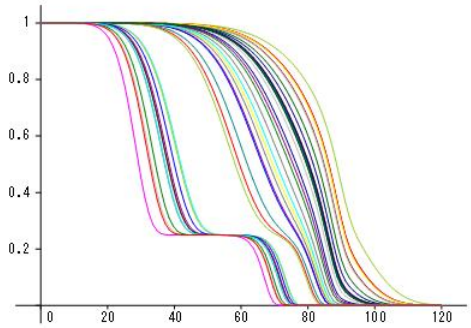


図 20　各個体の生存関数

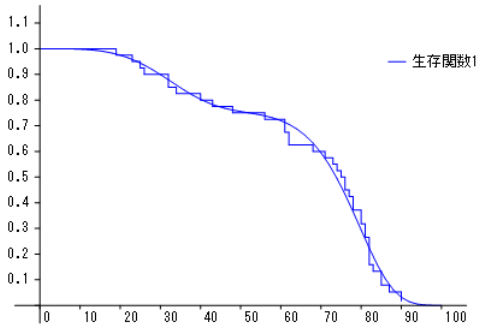


図 21　混合ワイブル分布による推定

このグラフの関係は、図 20 の曲線の平均を取ると、図 22 のように、図 21 の形になる。

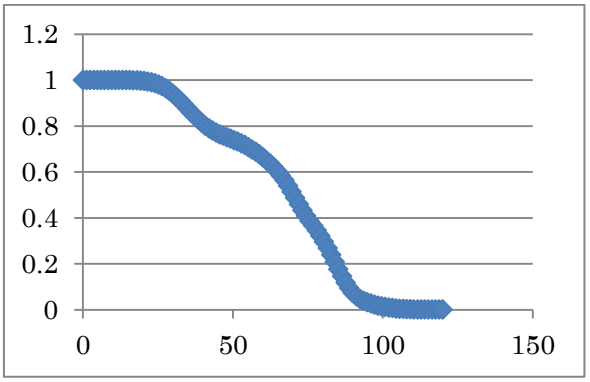


図 22　各個体の生存関数の平均

推定するデータを別頁にするときは、「予測用頁」テキストボックスにデータのある頁番号を入力し、「寿命予測」ボタンをクリックする。

推定するデータを別頁にするときは、「予測用頁」テキストボックスにデータのある頁番号を入力し、「寿命予測」ボタンをクリックする。

参考文献

[1] 打波守, Excel で学ぶ生存時間解析, オーム社, 2005.

[2] 柳井晴夫, 高木廣文編著, 多変量解析ハンドブック, 現代数学社, 1986.

[3] Annete J. Dobson, 田中豊他訳, 一般化線形モデル入門 原著第 2 版, 共立出版, 2008.