福山平成大学経営学部紀要
第 17 号（2021）、*－**頁

College Analysis を使い易くする追加機能３

福井　正康[*1]・細川　光浩[*2]

[*1] 福山平成大学経営学部経営学科
[*2] 福山平成大学大学教育センター

要旨：我々は社会システム分析に用いられる様々な手法を統合化したプログラム College Analysis を作成し、新しい分析を追加するごとに報告してきたが、ここでは既存のプログラムの拡張として、順位和検定への同順位処理の追加、直交表実験計画法へのコンジョイント分析の追加、重回帰分析の予測値への標準化残差とてこ比の追加、回帰式の比較に関する共分散分析の拡張などについて報告する。

キーワード：順位和検定の同順位補正、コンジョイント分析、てこ比、共分散分析

1.　はじめに

この報告は、これまで College Analysis で作成してきた既存のプログラムの拡張の報告である。分析手法が増えるわけではないが、これまで出来なかった改良や利用者からの要望など、重要なプログラムも含まれている。

今回の改良では 2 章の順位和検定の同順位補正が特に重要である。統計分析で、検定のプログラムを作成して以来、順位和検定に同順位補正を含めていないことが、我々には大きな気がかりであった。今回は同順位和補正に加えて Kruskal-Wallis 検定に Yates 補正も加えることができた。これにより、2 群の検定と 3 群以上の検定のつながりが分かり易くなった。

コンジョイント分析は経営分野で使われることのある分析であるが、これは直交表による比較を基本としている。そのため直交表分散分析のプログラムに追加した。また、これまでの重回帰分析では、回帰係数などの検定には力を入れてきたが、予測値の誤差には触れてこなかった。今回この点を改良し、予測値に、標準化残差と不適切なデータを見分ける「てこ比」を追加した。また、バイアスのかかった 1 元配置分散分析の問題で、バイアスを除去する 1 つの方法として知られる共分散分析を、回帰式の定数項の比較の問題として追加した。

2.　順位和検定の同順位補正

College Analysis のノンパラメトリック検定において、これまで Wilcoxon の順位和検定と

符号付き順位和検定には Yates 補正を付け、Kruskal -Wallis 検定では何も補正を付けていなかった。今回まず Kruskal-Wallis 検定に Yates 補正を付け、これらすべてに同順位補正を加えた[1]。

データへの順位の割り当てで、同じ数値には順位の平均を割り当てて来たが、同順位の多いデータでは同順位補正により、精度の向上が期待できる。我々はこの 2 つの補正を図 1 のように変数選択の画面に加え、利用者が自由に選べるようにした。

図 1　Yates 補正と同順位補正

以下、検定の種類別にこれらの補正の定義を与えておく。

Wilcoxon の符号付順位和検定

データ x_i 、指定値 μ 、 $z_i = x_i - \mu$

$|z_i|$ の昇順に 0 を除いて順位 r_i を付け、 z_i の正負で 2 群に分類する。

各群の順位和 R_r, R_s の中で小さい方を選択し、 $R = \min(R_r, R_s)$ とする。

標本数が少ないとき（$z_i \neq 0$ の例数<10）

参考文献 [1] の数表を利用する。

標本数が多いとき（$z_i \neq 0$ の例数≥ 10、τ_i は小さい方から i 番目の同順位データの数）

$$Z = \frac{R - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \sim N(0,1)$$

$$\to \frac{|R - n(n+1)/4| - 1/2}{\sqrt{n(n+1)(2n+1)/24}} \left[1 - \frac{1}{2n(n+1)(2n+1)} \sum_{i=1}^{e} \tau_i (\tau_i^2 - 1) \right]^{-1/2}$$

Wilcoxon 符号付順位和検定において、同順位の場合は順位平均を用いているが、上は補正なしの場合、下は Yates の連続補正と同順位の補正を加えた場合である。プログラムのデフォルトは両方の補正を加えたものとしている。

Wilcoxon の順位和検定

標本数 n_1, n_2 （$n_1 \leq n_2$）、標本 x_i^1, x_j^2

標本の昇順に順位 r_i を付け、標本数の少ない群の順位和を求める。

$$W = \sum_{i=1}^{n_1} r_i$$

標本数が少ない場合（$n_2 \le 20$）

参考文献 [1] の数表を利用する。

標本数が多い場合（$n_2 > 20$，τ_i は小さい方から i 番目の同順位データの数）

$$Z = \frac{W - n_1(N+1)/2}{\sqrt{n_1 n_2 (N+1)/12}} \sim N(0,1)$$

$$\to \frac{|W - n_1(N+1)/2| - 1/2}{\sqrt{n_1 n_2 (N+1)/12}} \left[1 - \frac{1}{N(N^2-1)} \sum_{i=1}^{e} \tau_i (\tau_i^2 - 1) \right]^{-1/2}$$

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

2) Kruskal-Wallis 検定

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

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

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

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

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

3. コンジョイント分析

コンジョイント分析は直交表分散分析と数量化 I 類を合わせた分析手法である。直交表分散分析では直交表によって実験の組み合わせを考え実験計画を立てるが[2]、コンジョイント分析ではアンケートの中で商品の特徴を効率よく組み合わせるために直交表が使われる。回答者はこのように特徴が組み合わされた商品に対して効用値（点数でも好きな順位でもよい）を付ける。分析では、効用値を目的変数にして数量化 I 類を実行し、各質問項目がどのように効用値に影響しているかを見る。コンジョイント分析が数量化 I 類と異なるところは、交互作用を容易に取り入れられるところである。

回答者が複数の場合、データを後ろに加えて行くが、通常の直交表分散分析では同じ質問項目の平均を使って分析を実行するため、多くのデータを集めることは意味がない。実際、直交表分散分析は実験回数を減らすことが目的なので、複数組のデータということは考えない。しかし、コンジョイント分析は数量化 I 類（0/1 データに変換して重回帰分析）の処理を行うので、パラメータの精度はデータ数に応じて良くなり、特徴が見つけやすくなる。

コンジョイント分析を含む直交表実験計画法の実行画面を図 1 に示す。

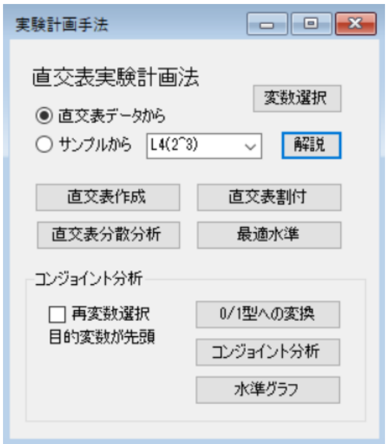

図 1　直交表実験計画法実行画面

直交表実験計画法 2.txt のデータを利用して、コンジョイント分析のプログラムを見て行こう。図 2 にそのデータを表示する。

データ編集 直交表実験計画法2.txt

L8(2^7)	A	B	A*B	D		A*C	C	data
1	1	1	1	1	1	1	1	20
2	1	1	1	2	2	2	2	22
3	1	2	2	1	1	2	2	25
4	1	2	2	2	2	1	1	19
5	2	1	2	1	2	1	2	27
6	2	1	2	2	1	2	1	24
7	2	2	1	1	2	2	1	19
8	2	2	1	2	1	1	2	22

1/9 (1,1)　分析　備考

図 2　直交表実験計画法のデータ 1 （直交表実験計画法 2.txt）

この場合、データ（目的変数）の選択順は直交表分散分析と同じく最後にする。直交表分散分析の結果を図 3 に、コンジョイント分析の結果を図 4 に示す。

直交表分散分析結果

要因	平方和S	自由度	平均平方V	F値	P値
A	4.500	1	4.500	2.250	0.3743
B	8.000	1	8.000	4.000	0.2952
A*B	18.000	1	18.000	9.000	0.2048
D	2.000	1	2.000	1.000	0.5000
A*C	0.500	1	0.500	0.250	0.7048
C	24.500	1	24.500	12.250	0.1772
誤差	2.000	1	2.000		
Total	59.500	7			

図 3　直交表分散分析の結果

カテゴリウェイト

	重回帰ウェイ	重回帰確率	基準化ウェイ	基準化確率	結合変数	結合確率
A:1	0.0000		-0.7500	0.3743	A	0.3743
A:2	1.5000	0.3743	0.7500	0.3743	B	0.2952
B:1	0.0000		1.0000	0.2952	A*B	0.2048
B:2	-2.0000	0.2952	-1.0000	0.2952	D	0.5000
A*B:1	0.0000		-1.5000	0.2048	A*C	0.7048
A*B:2	3.0000	0.2048	1.5000	0.2048	C	0.1772
D:1	0.0000		0.5000	0.5000		
D:2	-1.0000	0.5000	-0.5000	0.5000		
A*C:1	0.0000		-0.2500	0.7048		
A*C:2	0.5000	0.7048	0.2500	0.7048		
C:1	0.0000		-1.7500	0.1772		
C:2	3.5000	0.1772	1.7500	0.1772		
定数項	19.5000	0.0431	22.2500	0.0143		
重相関	0.983	寄与率	0.966			
有効性F値	4.7917	自由度	6,1	p値	0.1006	

図 4　コンジョイント分析の結果

コンジョイント分析の結果は、数量化 I 類の重回帰ウェイトと基準化ウェイトを用いており、どの選択肢が結果を上げるのか、下げるのか、などがよく分かる。また「結合確率」によって、直交表分散分析と同様に、各変数の重要性も分かる。ここで、基準化ウェイトの係数が 0 となることを検定する「基準化確率」や 1 つの変数を構成する分けられた複数変数の係数が同時に 0 になることを検定する「結合確率」は、結合仮説の検定の以下の関係式を用いて求めている。

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

ここに、 q は同時制約の数である。

「水準グラフ」ボタンをクリックすると、この基準化ウェイトの変動を図 5 のようにグラフ化することができる。

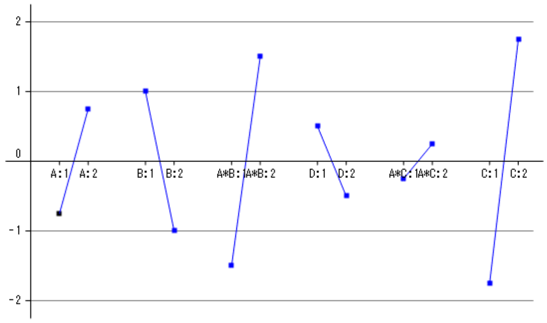

図 5　基準化ウェイト

次にもう少し複雑な例として、1 つの変数を 2 つ以上の列で表す例を考える。図 6 にそのデータを与えておく。

データ編集 直交表実験計画法2.txt

	A	B	C	A*B	A*B	A*C	A*C	D	F	A*D	A*D			data
1	1	1	1	1	1	1	1	1	1	1	1	1	1	14
2	1	1	2	1	1	2	3	2	3	2	3	3	2	15
3	1	1	3	1	1	3	2	3	2	3	2	2	3	16
4	1	2	1	2	3	1	1	2	2	2	3	2	3	23
5	1	2	2	2	3	2	3	3	1	3	2	1	1	21
6	1	2	3	2	3	3	2	1	3	1	1	3	2	22
7	1	3	1	3	2	1	1	3	3	3	2	3	2	14
8	1	3	2	3	2	2	3	1	2	1	1	2	3	17
9	1	3	3	3	2	3	2	2	1	2	3	1	1	15
10	2	1	1	2	2	2	2	1	1	2	2	2	2	21
11	2	1	2	2	2	3	1	2	3	3	1	1	3	22
12	2	1	3	2	2	1	3	3	2	1	3	3	1	20

3/9 (1,1)　分析　備考 p162上

図 6　直交表実験計画法のデータ 2 （直交表実験計画法 2.txt）

このデータから求められたコンジョイント分析の結果を図 7 に示す。

カテゴリウェイト

	重回帰ウェイ	重回帰確率	基準化ウェイ	基準化確率	結合変数	結合確率
A:1	0.0000		−0.8889	0.1657	A	0.0464
A:2	2.8889	0.0337	2.0000	0.0190	B	0.0318
A:3	−0.2222	0.8190	−1.1111	0.1018	C	0.2740
B:1	0.0000		−0.7778	0.2127	A*B	0.1639
B:2	3.0000	0.0300	2.2222	0.0133	A*C	0.0552
B:3	−0.6667	0.5042	−1.4444	0.0513	D	0.7014
C:1	0.0000		1.0000	0.1296	F	0.6516
C:2	−1.5556	0.1624	−0.5556	0.3497	A*D	0.6610
C:3	−1.4444	0.1874	−0.4444	0.4450		
A*B:1	0.0000		−1.0000	0.1296		
A*B:2	2.5556	0.0483	1.5556	0.0415		
A*B:3	0.4444	0.6507	−0.5556	0.3497		
A*B:1	0.0000		−0.6667	0.2731		
A*B:2	0.5556	0.5743	−0.1111	0.8428		

図 7　コンジョイント分析結果

また基準化ウェイトをグラフにしたものを図 8 に示す。

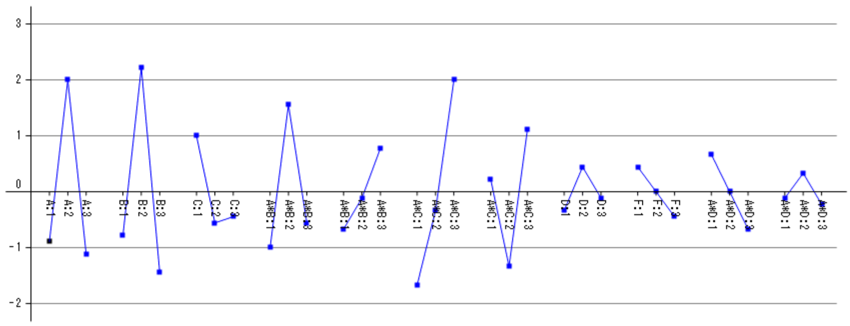

図 8　基準化ウェイト

この分析メニューでは、不要な変数を取り除いて選択し、通常の数量化 I 類のように、目的変数を先頭に選んで分析することもできる。その際は、コンジョイント分析グループボックス内の「再変数選択」チェックボックスにチェックを入れる。

4.　重回帰分析とてこ比

最初にここで使う記号を簡単に整理しておく。

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

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

ここに、

$$\mathbf{y}(N\times 1)\text{，}\mathbf{Z}=\begin{pmatrix}\mathbf{1} & \mathbf{X}(N\times k)\end{pmatrix}\text{，}\mathbf{d}'=\begin{pmatrix}b_0 & \mathbf{b}'(1\times k)\end{pmatrix}\text{，}\mathbf{u}(N\times 1)\sim N(\mathbf{0},\boldsymbol{\Sigma}_u)$$

標準化残差とてこ比について

College Analysis の重回帰分析では､回帰係数の標準誤差などは詳しく与えられているが、予測値の誤差については表示されていなかった。今回の改良では、データの個別の正当性を見るために、予測値の標準化残差とてこ比を出力に付け加えた[3]。

重回帰分析で、最小 2 乗法を用いると、偏回帰係数の推定値 $\hat{\mathbf{d}}$ は以下となる。

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

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

次に、回帰分析の予測値は以下のように与えられるが、

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

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

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

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

これを成分で書き表すと、

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

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

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

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

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

予測値の分布

てこ比に関連して予測値の分布についても見ておく。

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

上の関係から、予測値の共分散は以下で与えられることが分かる。

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

これから、次の結果を得る。

均一分散の場合

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

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

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

不均一分散の場合

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

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

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

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

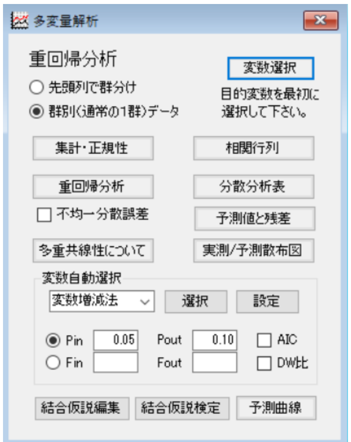

図 1　重回帰分析実行画面

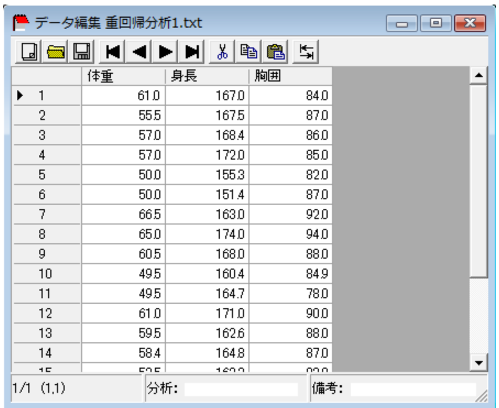

図 2　重回帰分析データ

実行画面の「予測値と残差」ボタンでは、図 3 のように各レコード毎の実測値、予測値、残差、標準化残差、てこ比が表示される。変数名の部分に、目安として与えられるてこ比の平均の 2.5 倍の値が表示されている。

予測値と残差

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

図 3　予測値と残差

5.　共分散分析

共分散分析は交絡因子を除いて群を比較する 1 つの方法である[1]。今、群 i（$i=1,\cdots,k$）における変数 y_i の平均を比較する問題を考える。これは単純な 1 元配置の問題であるが、この変数に影響を及ぼす交絡因子 x_i が存在し、群間でこの交絡因子に偏りが見られるとき、y_i を単純に比較することはできない。しかし、交絡因子の影響が各群同じで、共通のパラメータ a を用いて以下のように与えられるとき、

$$y_i = ax_i + b_i + \varepsilon_i \text{、} E(\varepsilon_i)=0 \text{、} Var(\varepsilon_i)=\sigma^2$$

群間の本質的な違いは切片 b_i の違いに帰着する。以下ではこのような回帰直線で与えられる k 群のデータの勾配係数と切片の差の検定を考える。

k 群の回帰係数の比較では、群別にデータ数を n_i、標本回帰式を $y=a_i x+b_i$ として、まず、以下の関係を利用して勾配係数の比較を行う。

$$F_a = \left[(\Delta_2/\Delta_1)-1\right](N-2k)/(k-1) \sim F_{k-1,\,N-2k} \text{、} \qquad N=\sum_{i=1}^{k} n_i$$

勾配係数が異なるとすると回帰式はそのまま使われ、勾配係数が等しいとすると以下の関係を利用して共分散分析と呼ばれる定数係数の比較を行う。

$$F_b = \left[(\Delta_3/\Delta_2)-1\right](N-k-1)/(k-1) \sim F_{k-1,\,N-k-1}$$

ここで、定数係数が異なるとすると

$$a=\sum_{i=1}^{k} SS_{xyi} \Big/ \sum_{i=1}^{k} SS_{xi} \text{、} \quad b_i = \bar{y}_i - a\bar{x}_i$$

として、回帰式は以下を与える。

$$y=ax+b_1 \text{、} \quad y=ax+b_2$$

定数係数が同じとすると $a=SS_{xy}\big/SS_x$、$b=\bar{y}-a\bar{x}$ として、回帰式は同一に以下で与える。

$$y=ax+b$$

ここに、$i=1,\cdots,k$ として以下の関係を用いる。

$$\bar{x}_i=\frac{1}{n_i}\sum\nolimits_{\lambda=1}^{n}x_{i\lambda}\,,\quad \bar{y}_i=\frac{1}{n_i}\sum\nolimits_{\lambda=1}^{n}y_{i\lambda}$$

$$SS_{xi}=\sum\nolimits_{\lambda=1}^{n}x_{i\lambda}^2-n_i\bar{x}_i^2\,,\quad SS_{yi}=\sum\nolimits_{\lambda=1}^{n}y_{i\lambda}^2-n_i\bar{y}_i^2\text{、}\quad SS_{xyi}=\sum\nolimits_{\lambda=1}^{n}x_{i\lambda}y_{i\lambda}-n_i\bar{x}_i\bar{y}_i$$

$$\Delta_1=\sum\nolimits_{i=1}^{k}\left[SS_{yi}-(SS_{xyi})^2\big/SS_{xi}\right]$$

$$\Delta_2=\sum\nolimits_{i=1}^{k}SS_{yi}-\left(\sum\nolimits_{i=1}^{k}SS_{xyi}\right)^2\Big/\sum\nolimits_{i=1}^{k}SS_{x1}\text{、}\qquad \Delta_3=SS_y-(SS_{xy})^2\big/SS_x$$

回帰式の比較の実行画面とデータを図 1 に、共分散分析と呼ばれる「多群回帰式比較」ボタンをクリックした結果を図 2 に示す。ここでは、すべての比較の結果が出力される。

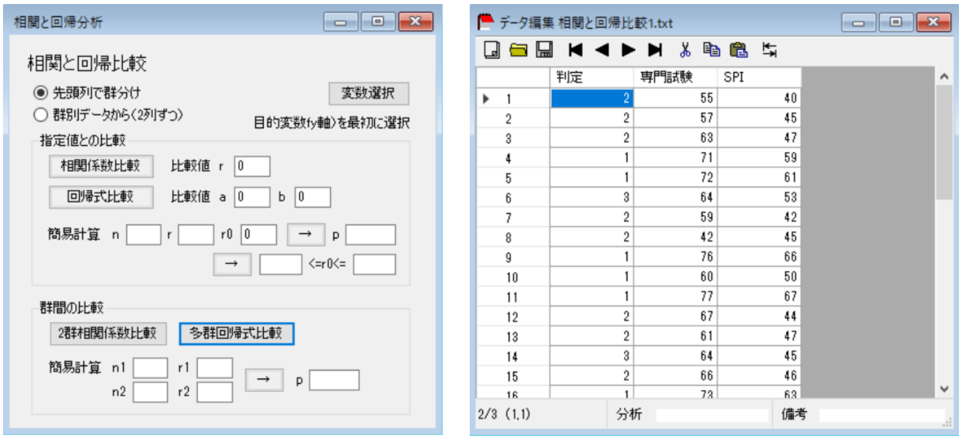

図 1　回帰式の比較の実行画面とデータ

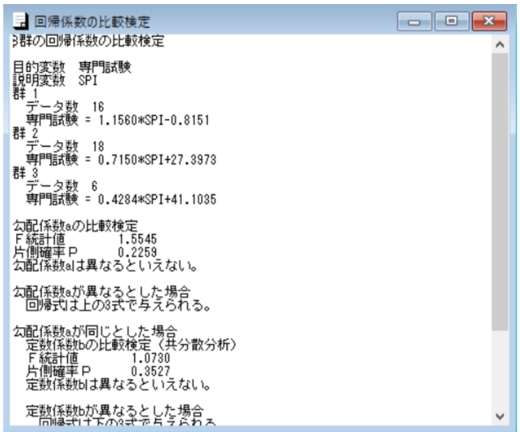

図 2　共分散分析（多群の回帰式比較）実行結果

6.　おわりに

今回追加した Kruskal-Wallis 検定での Yates 補正や順位和検定全般での同順位補正は、以前からの懸案事項であった。授業などで Wilcoxon 順位和検定と Kruskal-Wallis 検定とのつながりを説明する際や同順位が多いデータでの検定の際に常に気に掛かっていた。Yates 補正によって前者は問題がなくなり、後者も同順位補正により検定確率が小さくなることで、以前の結果も間違いでないことが分かった。

コンジョイント分析については、直交表分散分析がデータを最後の列に指定していたので、数量化 I 類のデータ選択の順番と逆になってしまった。利用の際には注意が必要である。また、利用に際して、直交表についての理解がある程度必要で、特に交互作用項について少し調べておく必要がある。

重回帰分析については、予測値以外にも「多重共線性について」ボタンから表示されるリッジ回帰分析他の実行メニューで、交差検証法をこれまでの近似法から厳密な方法に書き換えている。共分散分析は、これまでの 2 群の回帰直線の比較の検定を 3 群以上に拡張したものである。使い方は 2 群の場合と同じで特に問題はないであろう。これ以外に、実験計画法の「等分散性の検定」では、通常 Bartlett 検定を使用するが、特に 2 群の場合には F 検定も使えるようにした。また、量的データの集計の「S-W 検定*」で、複数変数選択の場合の結果の出力順に変更を加えている。

プログラムは一度作れば終わりではない。今後も追加と改良を繰り返して行く。一度作られたプログラムの改良や部分的な機能追加については、このシリーズで紹介して行く。

参考文献

［１］　丹後俊郎，古川俊之監修、「医学への統計学【第 3 版】」、朝倉書店（2013）

［２］　森田浩，今里健一郎，奥村清志、「Excel でここまでできる実験計画法 一元配置実験から直交配列表実験まで」、日本規格協会（2011）

［３］　永田靖，棟近雅彦、「多変量解析法入門」、サイエンス社（2001）