岡本安晴　2002.8

数量化１類

　従属変数（外的基準値という）が連続量で与えられていて、独立変数が質的データである場合において、外的基準値を予測するのに最適な質的データの数量化を考える。i 番目の個体の外的基準値を $y(i)$ で、アイテム j におけるカテゴリ k の値を $n_i(jk)$ で表す[1]。ここで、$n_i(jk)$ の値は次のように定める。

$$n_i(jk) = \begin{cases} 1 & i\text{番目の個体がアイテム}j\text{のカテゴリ}k\text{を選択しているとき} \\ 0 & i\text{番目の個体がアイテム}j\text{のカテゴリ}k\text{を選択していないとき} \end{cases}$$

それぞれのアイテムから１つのカテゴリが選択されるものとする。個体の総数を m 、アイテムの総数を n、アイテム j のカテゴリの数を N_j とおく。

　$n_i(jk)$ は次式を満たす。

$$\sum_{k=1}^{N_j} n_i(jk) = 1$$

　いま、アイテム j におけるカテゴリ k の値に適当な数値 $x(jk)$ を割り当てて外的基準値 $y(i)$ を次式で推定することを考える。

$$y(i) \approx \hat{y}(i) = c + \sum_{j=1}^{n} \sum_{k=1}^{N_j} x(jk) \cdot n_i(jk) \qquad （１）$$

ここで、 c は適当な定数である。

　（１）式による推定の誤差を次式で表す。

$$Q = \sum_{i=1}^{m} \left(y(i) - \hat{y}(i) \right)^2 \qquad （２）$$

（２）式の値を最小にする c は次式

$$c = \bar{y} - \sum_{j=1}^{n} \sum_{k=1}^{N_j} x(jk) \cdot \bar{n}(jk) \qquad （３）$$

で与えられ、 $x(jk)$ は次式を満たすものとして与えられる。

$$\mathbf{Ax} = \mathbf{b} \qquad （４）$$

ここで、

[1] ここでの記法は岩坪（1987）に従うものである。

$$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y(i)$$

であり、行列

$$\mathbf{A} = \left(a(jk, j'k')\right)$$

およびベクトル

$$\mathbf{b} = \left(b(jk)\right)$$

の要素は次式で与えられる。

$$a(jk, j'k') = \frac{1}{m}\sum_{i=1}^{m}\{n_i(jk) - \bar{n}(jk)\}\{n_i(j'k') - \bar{n}(j'k')\}$$

$$b(jk) = \frac{1}{m}\sum_{i=1}^{m}\{y(i) - \bar{y}\}\{n_i(jk) - \bar{n}(jk)\}$$

ここで、

$$\bar{n}(jk) = \frac{1}{m}\sum_{i=1}^{m} n_i(jk)$$

である。

（4）式の解を、岩坪（1987）はムーア・ペンローズ逆行列 $\mathbf{A}^{+}$ を用いて

$$\mathbf{x} = \mathbf{A}^{+}\mathbf{b}$$

で与え、議論を展開している。

（4）式の 1 つの解において、同じアイテム内のカテゴリについて定数を加えたものも（4）式の解である（証明については付録を参照）。各アイテムごとの $x(jk)$ の平均値が 0 になるように定数を加える（標準化）。アイテム j における平均値は次式で与えられる。

$$\frac{1}{m}\sum_{i}\left\{\sum_{k} n_i(jk)x(jk)\right\} = \sum_{k}\left[\left\{\frac{1}{m}\sum_{i} n_i(jk)\right\}x(jk)\right]$$

$$= \sum_{k}\bar{n}(jk)x(jk)$$

上式の値を $x(jk)$ から引いた値

$$x^{*}(jk) = x(jk) - \sum_{k}\bar{n}(jk)x(jk)$$

は平均値が 0 になる。すなわち次式が成り立つ。、

$$\frac{1}{m}\sum_{i}\left\{\sum_{k} n_i(jk)x^{*}(jk)\right\}$$

$$= \frac{1}{m}\sum_{i}\left[\sum_{k} n_i(jk)\left\{x(jk) - \sum_{k}\bar{n}(jk)x(jk)\right\}\right]$$

$$= \frac{1}{m}\sum_i \left\{ \sum_k n_i(jk)x(jk) \right\} - \frac{1}{m}\sum_i \left[\left\{ \sum_k \bar{n}(jk)x(jk) \right\} \cdot \sum_k n_i(jk) \right]$$

$$= \sum_k \bar{n}(jk)x(jk) - \frac{1}{m}\sum_i \left[\left\{ \sum_k \bar{n}(jk)x(jk) \right\} \times 1 \right]$$

$$= \sum_k \bar{n}(jk)x(jk) - \sum_k \bar{n}(jk)x(jk)$$

$$= 0$$

標準化された解 $\mathbf{x}^* = \left(x^*(jk)\right)$ による外的基準値 $y(i)$ の推定値 $\hat{y}(i)$ は次式で与えられる。

$$\hat{y}(i) = \bar{y} + \sum_{j=1}^{n}\sum_{k=1}^{N_j} x^*(jk) \cdot n_i(jk)$$

プログラム

数量化１類による分析を行うプログラム PQ1.dpr を用意した。このプログラムを実行すると図１のフォームが表示される。

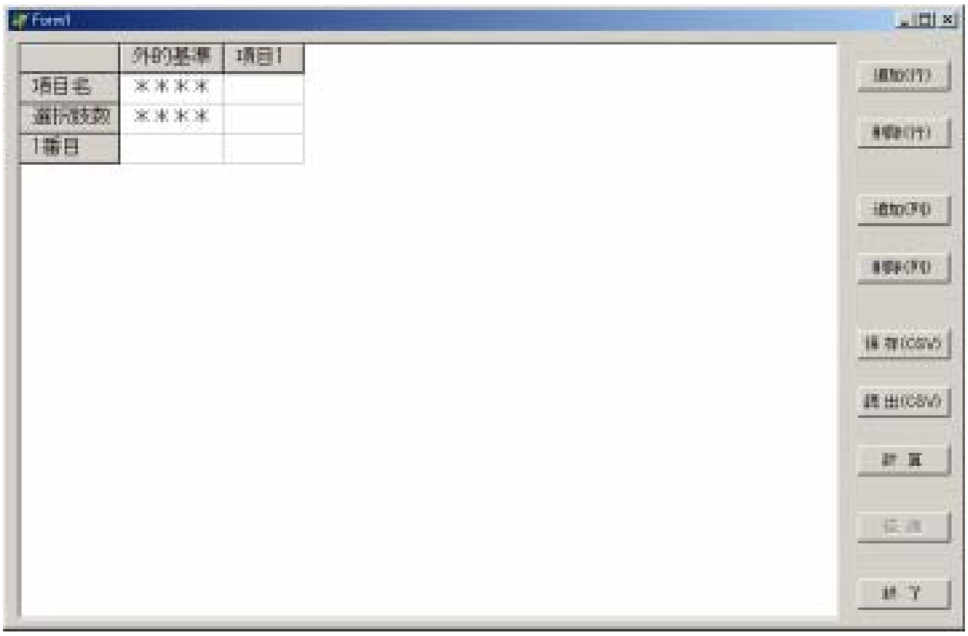


図１　実行開始時のフォーム

「追加」「削除」ボタンをクリックして、行および列を必要な数だけ増やす。図２はデータの設定例である。

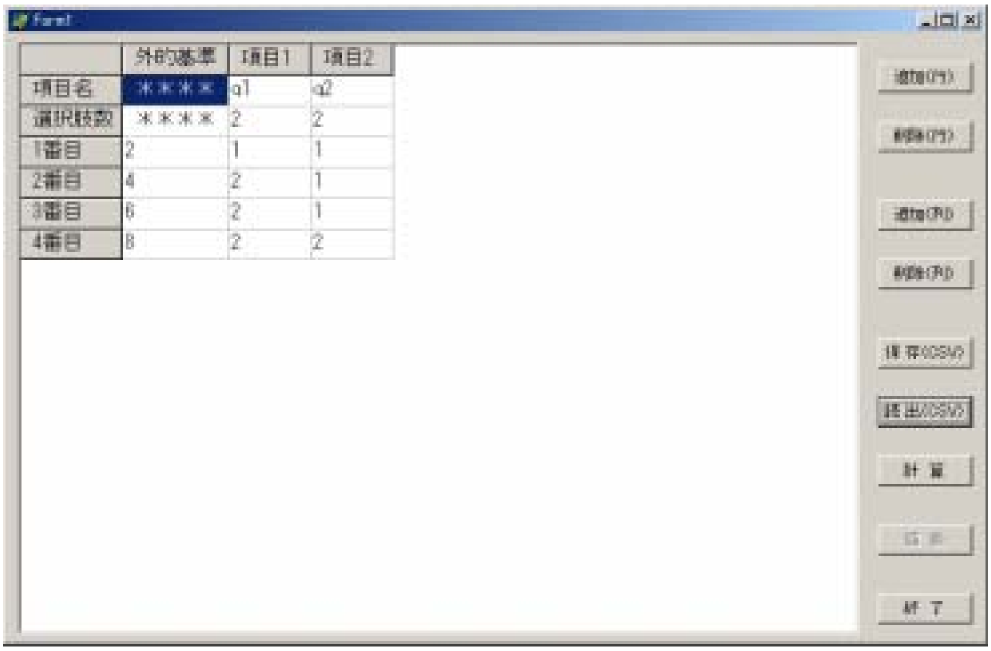
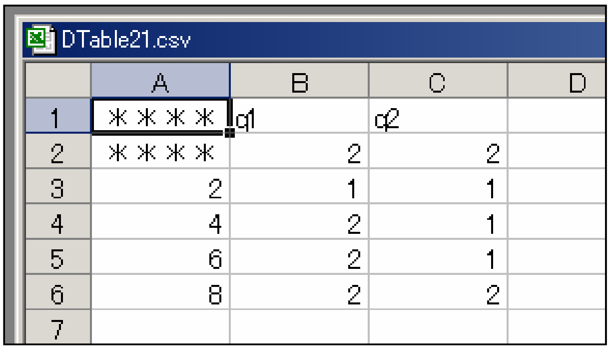


図２　データの設定

外的基準の下の 2 行に「＊＊＊＊」が設定されているが、このセルは入力の対象とならない。４行目から１番目のデータが設定されている。外的基準の欄には外的基準の値 $y(i)$ を設定する。項目欄は、２行目にその項目（アイテム）のラベル、３行目にその項目内のカテゴリの数を設定する。４行目からそれぞれの項目に対して選択されたカテゴリを設定する。カテゴリは項目ごとに１から順番に整数値を割り当て、その選択されたカテゴリの数値を設定する。カテゴリ数が N_j であれば１から N_j までの範囲の整数値を設定する。設定したデータは「保存（ＣＳＶ）」ボタンをクリックするとＣＳＶ形式で保存される。保存したデータは「読出（ＣＳＶ）」ボタンのクリックで読み込むことができる。データはＣＳＶ形式で保存されるので、Excel などで開くこともできる。図２のデータを保存したものを Excel で開くと図３のようになる。



DTable21.csv

	A	B	C	D
1	＊＊＊＊	q1	q2	
2	＊＊＊＊	2	2	
3	2	1	1	
4	4	2	1	
5	6	2	1	
6	8	2	2	
7				

図３　Excel で開いたデータ

逆に、図３の形式でデータを Excel で用意したものを CSV 形式で保存すれば、図１のフォームの「読出（CSV)」ボタンのクリックで読み込むことができる。

　データの設定後、「計算」ボタンをクリックすると図４のファイル名設定のダイアログボックスが表示される。

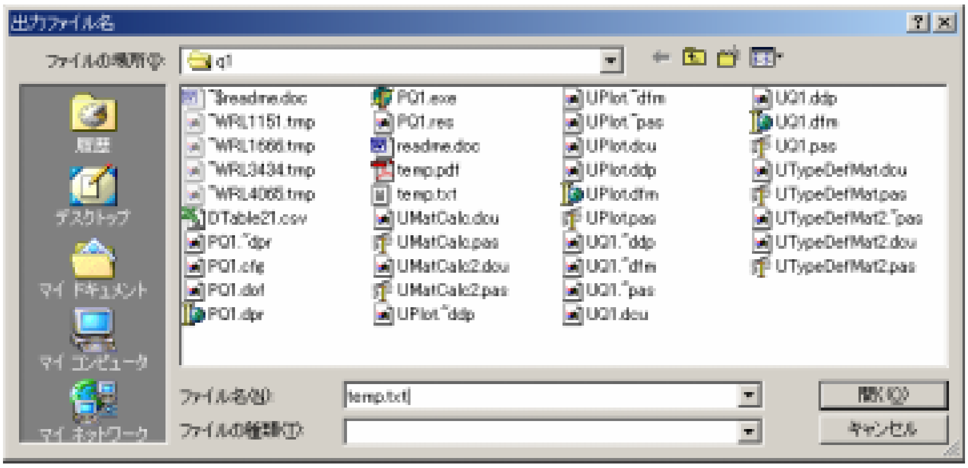


図４　出力用ファイル名の設定

図４で設定した名前のファイルに計算結果がテキストファイルとして書き出される。テキストファイルなので、プログラムの実行終了後エディタなどで開いて見ることができる。名前の設定後、「開く」ボタンをクリックすると計算が始まる。

　計算が終了すると図５のフォームになる。「描画」ボタンがイネーブルになっている。「描画」ボタンをクリックすると図６のように散布図が表示される。散布図には点$\left(y(i), \hat{y}(i)\right)$が青い小円でプロットされている。横軸の目盛り「MinY =」と「MaxY」は$y(i)$の最小値と最大値の位置を表す。

図５　計算終了時のフォーム

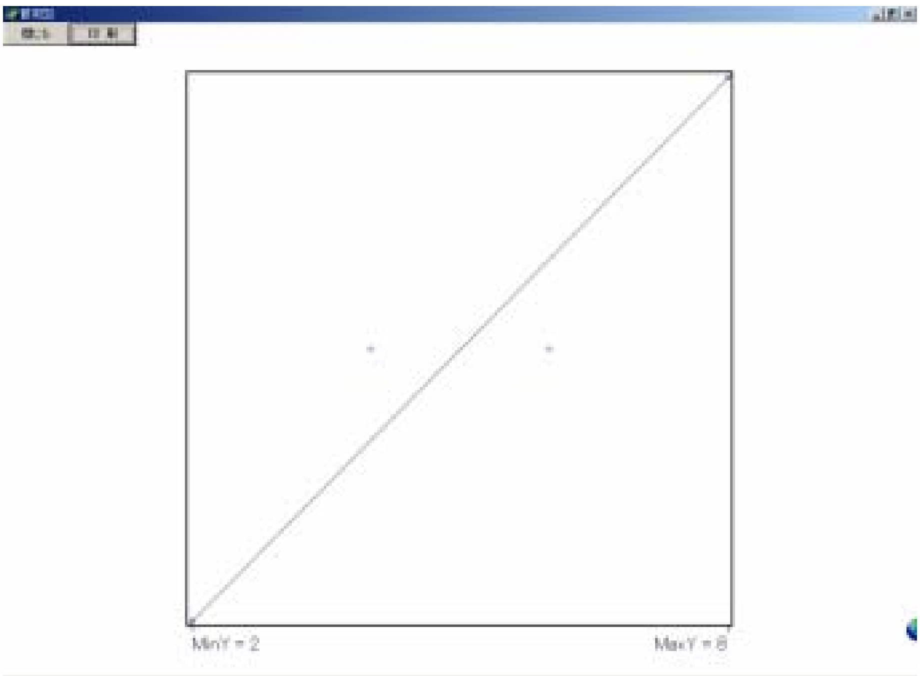


図６　散布図

図６の画面左上の「印刷」ボタンをクリックすると、散布図がプリンタに出力される。図

６の「閉じる」ボタン、あるいは図５の「終了」ボタンのクリックでプログラムの実行が終了する。

実行終了後、図４で設定した名前のテキストファイルを開くと、図２のデータの場合はリスト１のような出力になっている。

リスト１　出力例

```
Data...
No. of subjects = 4
No. of items    = 2
No. of categories
Item-1 = 2,  Item-2 = 2,

Values =
    1         2 10 10
    2         4 01 10
    3         6 01 10
    4         8 01 01

x =
    x(1,1) = -2.25
    x(1,2) = 0.75
    x(2,1) = -0.75
    x(2,2) = 2.25

y_hat =
       y_hat(1) = 2
       y_hat(2) = 5
       y_hat(3) = 5
       y_hat(4) = 8

外的基準値 y とその推定値 y_hat の相関係数 = 0.9486833
```

まず、入力データの内容が書き出されている。「Valus =」に続く行に固体別にデータ値が出力されている。各行は個体の通し番号がまず書き出され、続いて外的基準値 $y(i)$ が書き出されている。その後、各項目（アイテム）の選択カテゴリが１－０の形式で出力されている。項目間は空白で区切られている。例えば、２番目の個体の場合、外的基準値の値「４」に続いて「０１」とあるのは第１項目は第２カテゴリが選択されていることを表す。続いて、空白を挟んで「１０」とあるのは第２項目では第１カテゴリが選択されていることを表す。

入力データの出力の後、項目カテゴリの標準化された最適値 $x^*(jk)$ が書き出され、続いて外的基準値の推測値 $\hat{y}(i)$ が書き出されている。

最後に、外的基準値 $y(i)$ とその推測値 $\hat{y}(i)$ の相関係数が書き出されている。リスト１の場合、約 0.95 と高い値である。

岡本安晴　2002.8

付　録

$\mathbf{x} = \left(x(jk)\right)$が（４）式の解であるとき、アイテム j' のカテゴリ k' に $c_{j'}$ を加えたものを

$$\mathbf{x}_1 = \left(x_1(jk)\right)$$

とおく。

$$x_1(jk) = \begin{cases} x(jk) + c_{j'} & j = j' \text{のとき} \\ & \\ x(jk) & j \neq j' \text{のとき} \end{cases}$$

である。このとき

$$\mathbf{x}_1 = \mathbf{x} + \mathbf{d}$$

とおけば、

$$\mathbf{d}' = \begin{bmatrix} 0 & \cdots & 0 & \underbrace{c_{j'} \quad \cdots \quad c_{j'}}_{\text{項目}j'\text{のカテゴリ}} & 0 & \cdots & 0 \end{bmatrix}$$

となっている。よって

$$\mathbf{Ax}_1 = \mathbf{Ax} + \mathbf{Ad}$$

となる。$\mathbf{Ad}$ の (jk) 行目の値は次式で与えられる。

$$\sum_{k'=1}^{N_{j'}} a(jk, j'k') \cdot c_{j'} = c_{j'} \cdot \sum_{k'=1}^{N_{j'}} \left[\frac{1}{m} \sum_i \left\{ n_i(jk) - \bar{n}(jk) \right\} \left\{ n_i(j'k') - \bar{n}(j'k') \right\} \right]$$

$$= \frac{c_{j'}}{m} \sum_i \left[\left\{ n_i(jk) - \bar{n}(jk) \right\} \cdot \sum_{k'} \left\{ n_i(j'k') - \bar{n}(j'k') \right\} \right]$$

ここで、

$$\sum_{k'} \left\{ n_i(j'k') - \bar{n}(j'k') \right\} = \sum_{k'} n_i(j'k') - \sum_{k'} \bar{n}(j'k')$$

$$= 1 - \sum_{k'} \left\{ \frac{1}{m} \sum_i n_i(j'k') \right\}$$

$$= 1 - \frac{1}{m} \sum_i \left\{ \sum_{k'} n_i(j'k') \right\}$$

$$= 1 - \frac{1}{m} \sum_{i=1}^{m} 1$$

$$= 1 - \frac{1}{m} \times m$$

$$= 0$$

すなわち、

$$\mathbf{Ad} = \mathbf{0}$$

となる。よって、

$$\mathbf{Ax}_1 = \mathbf{Ax} = \mathbf{b}$$

の成り立つことが分る。

参考文献

岩坪秀一　1987　「数量化法の基礎」　朝倉書店