数量化プログラムの使い方と応用例

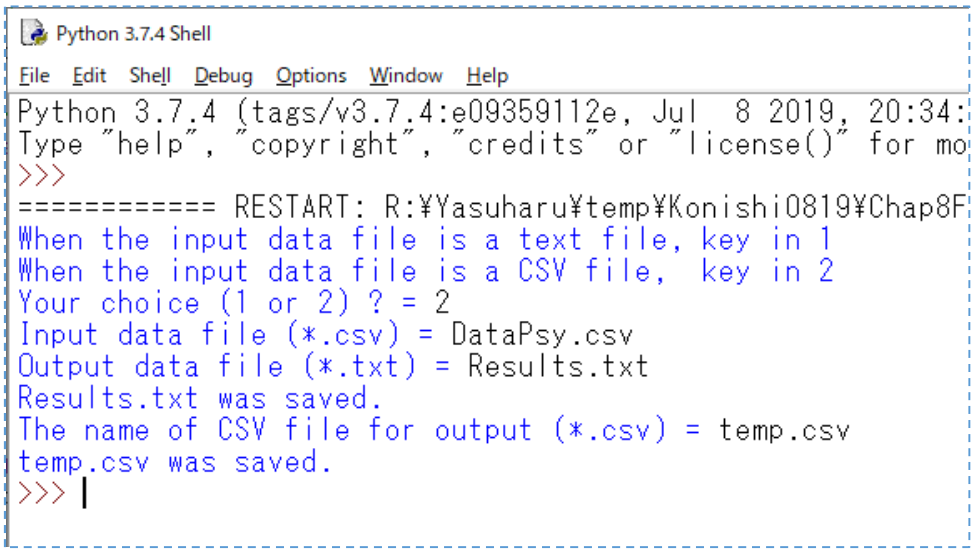
共通数量化は、機械学習におけるカテゴリ変数の前処理になっているが、カテゴリの数値化においてデータの分布が反映されている。本書第 8 章における共通数量化プログラム使用による分析例を分析の各ステップに分けて説明する。

ステップ 1 （フォルダ：Quantification）

フォルダ Quantification ファイル DataPsy.csv は以下のようになっている。

	A	B	C	D	E	F	G	H
1	ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7
2	0	3	3	3	3	3	0	0
3	39D	3	1	3	3	1	10	0
4	40D	3	2	2	2	1	40	0
5	41D	3	2	2	2	1	0	90
6	42D	2	1	2	2	1	60	80
7	43D	2	2	2	2	1	80	0

上のデータは、拙著表 8.2. 1 を入力したものである。このデータに対してプログラム Main.py を実行して、上記のファイルを入力ファイルとする



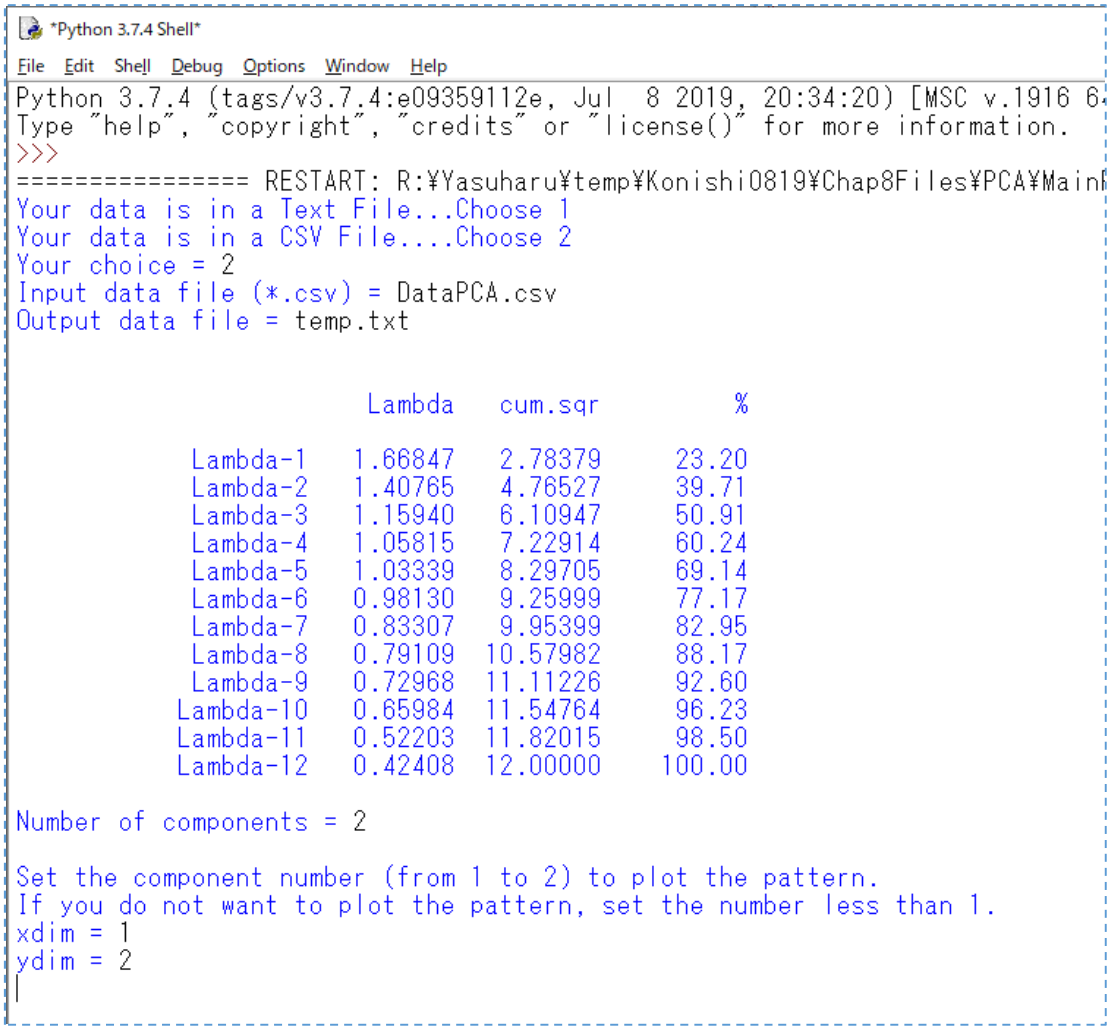
CSV 形式のファイルを、次のステップ 2 での入力ファイルとする。名前は拡張子として.csv を付けて適当に設定する

ステップ 2 （フォルダ：PCA）

ステップ 1 で作成したファイル temp.csv を Excel で開くと以下のようになっている。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	Q1-0	Q1-1	Q2-0	Q2-1	Q3-0	Q3-1	Q4-0	Q4-1	Q5-0	Q5-1	Q6	Q7	
2	39D	0.79057	1.27475	-1.48663	0.83062	1.22508	1.35371	1.69367	0.68177	-0.58248	-0.07445	10	0	
3	40D	0.79057	1.27475	0.77081	0.17566	-0.87731	0.05524	-0.65002	0.14803	-0.58248	-0.07445	40	0	
4	41D	0.79057	1.27475	0.77081	0.17566	-0.87731	0.05524	-0.65002	0.14803	-0.58248	-0.07445	0	90	
5	42D	-1.26491	0	-1.48663	0.83062	-0.87731	0.05524	-0.65002	0.14803	-0.58248	-0.07445	60	80	
6	43D	-1.26491	0	0.77081	0.17566	-0.87731	0.05524	-0.65002	0.14803	-0.58248	-0.07445	80	0	
7	44D	0.79057	-1.27475	0.77081	0.17566	-0.87731	0.05524	-0.65002	0.14803	-0.58248	-0.07445	80	80	

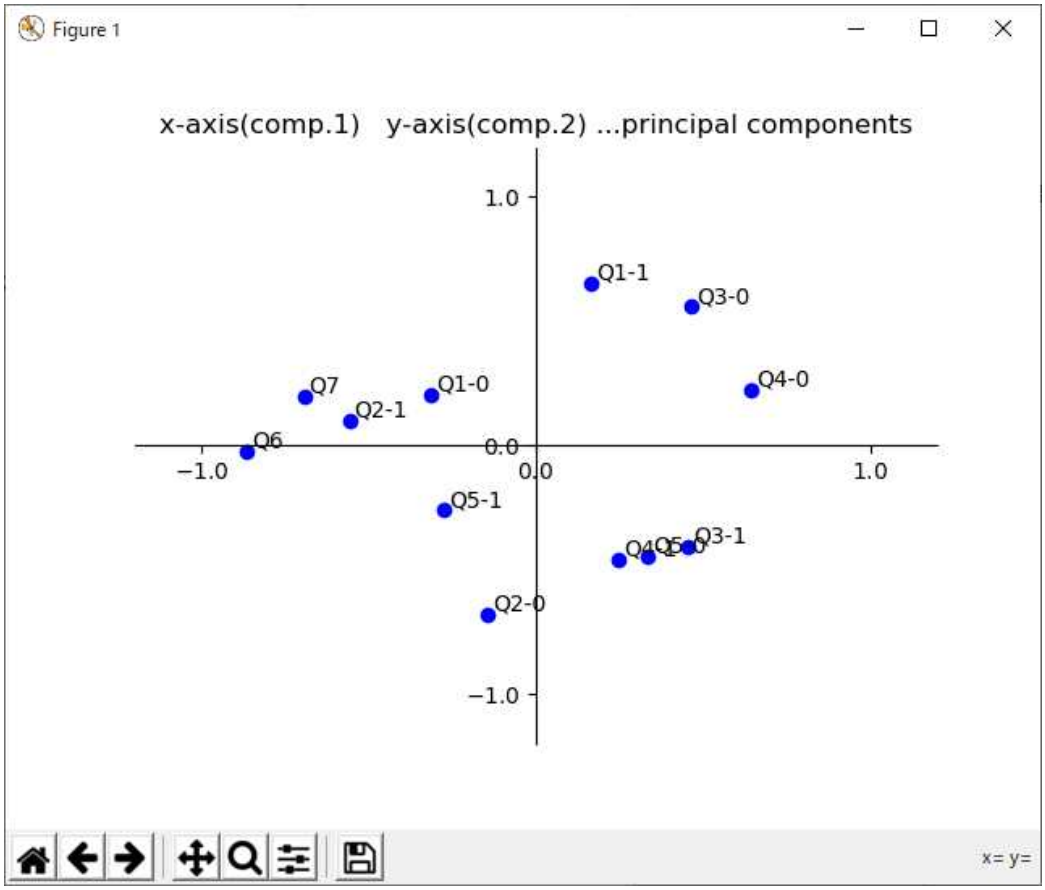
このファイルをフォルダ PCA にファイル名 DataPCA.csv として保存する。フォルダ PCA のファイル MainPrg.py を実行する。下図のようにファイル名などを設定する。



xdim = 1

ydim = 2

と設定すると、下図のマップが表示される。主成分分析プログラムの使い方は拙著第 7 章「主成分分析」を参照されたい。



拙著図 8.2.4 と比べると左右逆になっているが、これは固有値ベクトルが正負逆のものがとれるからである。拙著を用意していた時のライブラリと現在のライブラリとでバージョンが異なることによる。いずれも正しい解である。

上の布置から、Q2-1 と Q6、Q4-0 と Q6 の関係を調べる（拙著解説参照）。そのための入力ファイルをフォルダ SimpleRegression に作成する。

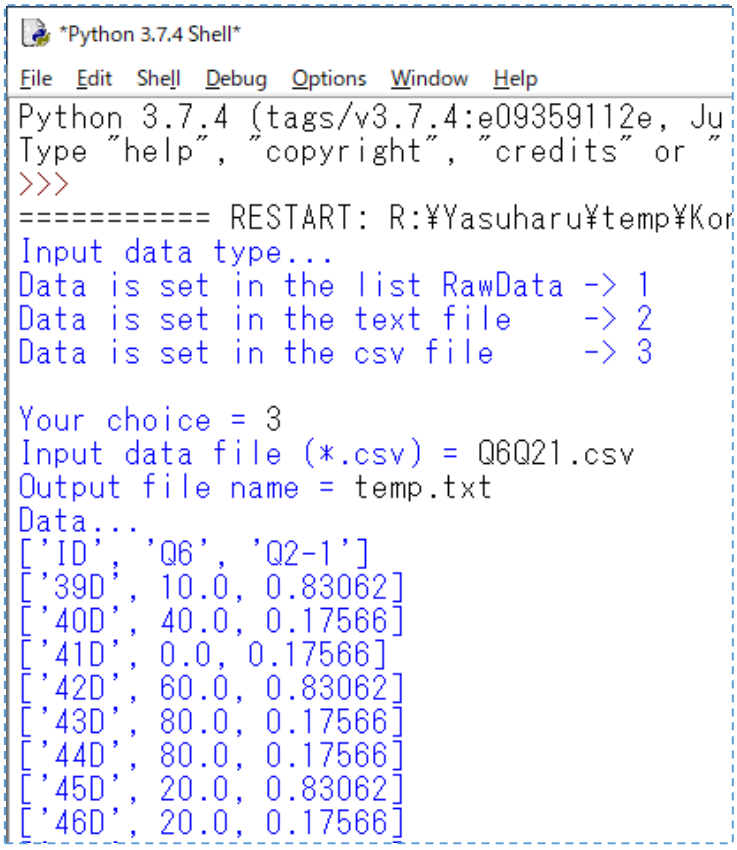
ステップ 3（フォルダ：SimpleRegression）

ステップ 1 での出力ファイル temp.csv を開く。

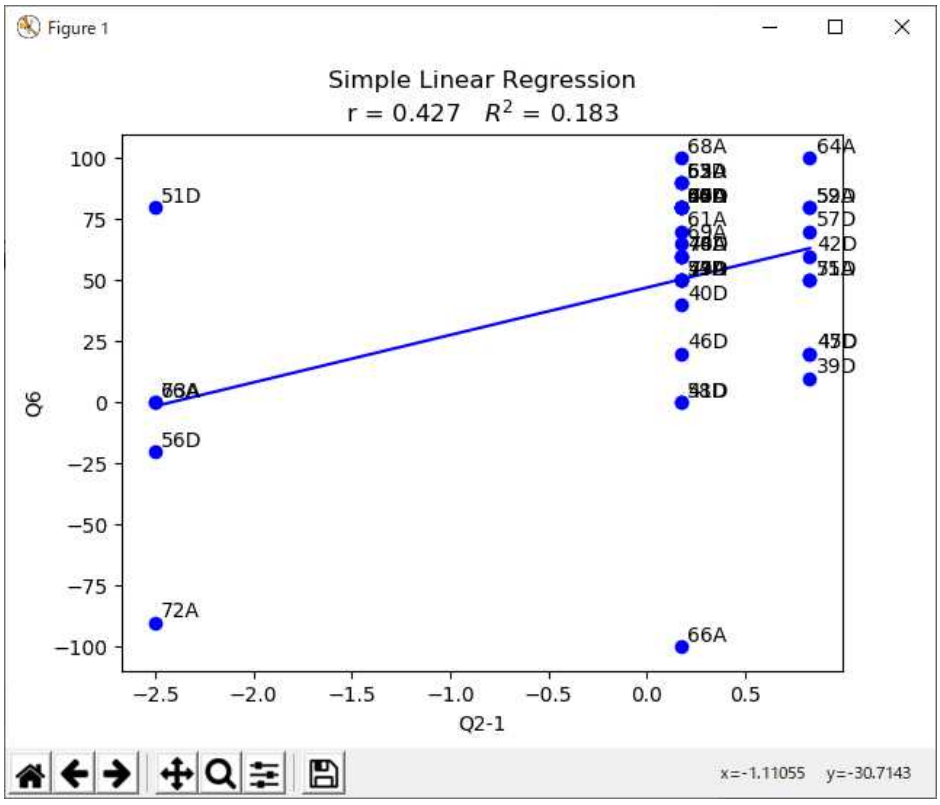
変数 Q6 と Q2-1 を下図のように選び、フォルダ SimpleRegression にファイル名 Q6Q21.csv として保存する。

	A	B	C	D
1	ID	Q6	Q2-1	
2	39D	10	0.83062	
3	40D	40	0.17566	
4	41D	0	0.17566	
5	42D	60	0.83062	
6	43D	80	0.17566	
7	44D	80	0.17566	
8	45D	20	0.83062	
9	46D	20	0.17566	

上のファイルを入力データとして、フォルダ SimpleRegression のプログラム Main.py を実行する。



上図のように入力ファイルなどを設定すると下図の回帰直線が描かれる。



次に、ステップ 1 での出力ファイル temp.csv を開き、変数 Q6 と Q4-0 を下図のように選び、フォルダ SimpleRegression にファイル名 Q6Q40.csv として保存する。

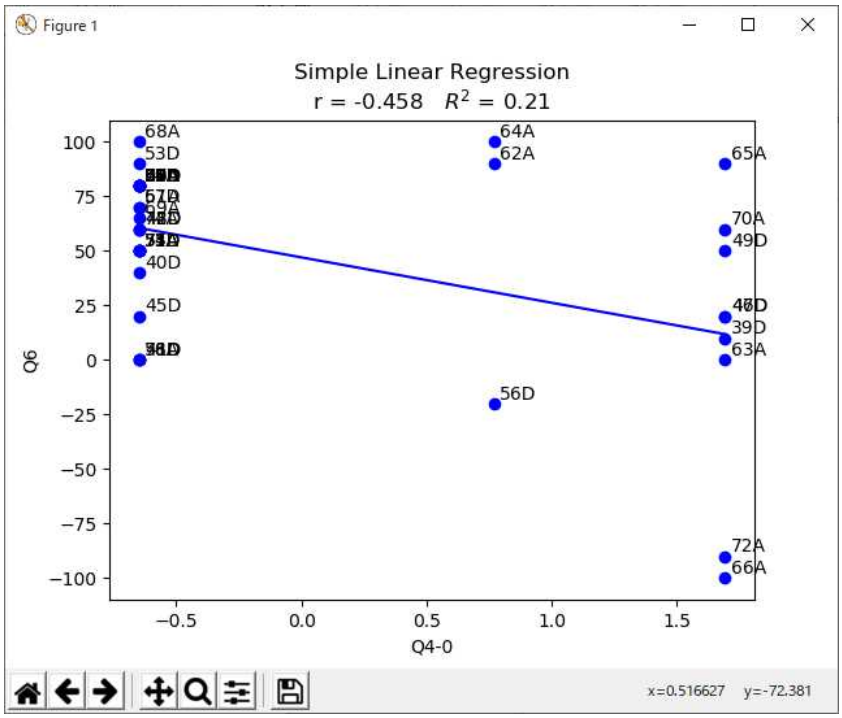
	A	B	C	D
1	ID	Q6	Q4-0	
2	39D	10	1.69367	
3	40D	40	-0.65002	
4	41D	0	-0.65002	
5	42D	60	-0.65002	
6	43D	80	-0.65002	
7	44D	80	-0.65002	
8	45D	20	-0.65002	
9	46D	20	1.69367	

上のファイルを入力データとして、フォルダ SimpleRegression のプログラム Main.py を実行する。

```
*Python 3.7.4 Shell*
File  Edit  Shell  Debug  Options  Window  Help
Python 3.7.4 (tags/v3.7.4:e09359112e, Jul  8 2
Type "help", "copyright", "credits" or "licens
>>>
========== RESTART: R:¥Yasuharu¥temp¥KonishiC
Input data type...
Data is set in the list RawData -> 1
Data is set in the text file    -> 2
Data is set in the csv file     -> 3

Your choice = 3
Input data file (*.csv) = Q6Q40.csv
Output file name = temp.txt
Data...
['ID', 'Q6', 'Q4-0']
['39D', 10.0, 1.69367]
['40D', 40.0, -0.65002]
```

上図の設定を行うと、下図の回帰曲線が描かれる。



ステップ 4 （フォルダ：CrossTable）

カテゴリ変数 Q1 と Q2 の関係を見る。変数 Q1 と Q2 を選ぶ理由については、本書の説明を参照されたい。

まず、ステップ 1 におけるフォルダ Quantification のデータファイル DataPsy.csv から下図のファイルを作成して、ファイル名 DataQ1Q2.csv としてフォルダ CrossTable に保存する。

	A	B	C
1	Q1	Q2	
2	3	1	
3	3	2	
4	3	2	
5	2	1	
6	2	2	
7	1	2	

上のデータを入力ファイルとしてフォルダ CrossTable のプログラム CrossTable.py を実行すると以下のようになる。

```
>>>
============ RESTART: R:¥Yasuharu¥temp¥Konishi0819¥Chap8Files¥
n1 =  3    n2 =  3
Input data file (*.csv) = DataQ1Q2.csv
Cnt =
 [[1, 1, 10], [3, 2, 10], [6, 2, 4]]


Cross Table
            psy=clin     other  psy>clin       Sum
      elem         1         1        10        12
    junior         3         2        10        15
    senior         6         2         4        12
       Sum        10         5        24        39
>>>
```