Median Polish 法を用いたデータ解析の特性に関する考察

2013SE089 小林優
指導教員：松田眞一

1　はじめに

Median Polish 法はジョン・チューキー氏によって提唱された方法であり，多因子の様々な要因を重要性を調べるためのデータの分析方法とされている．二元配置分散分析と Median Polish 法の二つの方法を用いて，はずれ値を含むデータの比較を行い，Median Polish 法の性質について明らかにしていく．

2　事例研究について

研究事例をいくつか行ったがここでは紙面の都合上 1 つだけ紹介する．Web[1] のデータを用いて，47 都道府県の県別人口と県庁所在地人口の都道府県人口に対する割合から都道府県の平均地価を求めた．ここでは県庁所在地の人口÷都道府県の人口を県庁所在地人口率という言葉で定義する．47 都道府県別人口は 600 万人以上，600 万人-200 万人，200 万人以下の大，中，小で区分けし，県庁所在地人口率は 40% 以上，40%-20%，20% 以下の大，中，小で区分けした．そして図 1 のように各区分に入る 9 都道府県を取り出した．表 1 は縦は人口，横は県庁所在地人口率に対し，それぞれの都道府県の平均地価をあらわしている．

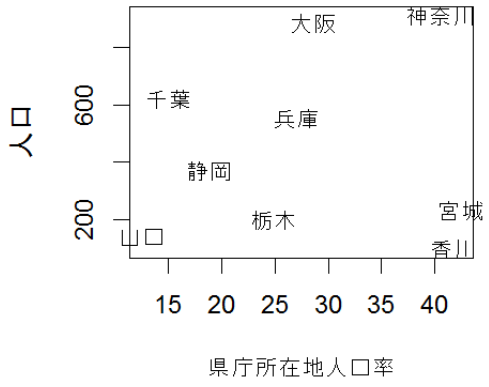

図 1　人口と県庁所在地人口率の関係

表 1　平均地価

	大	中	小
大	23.91	23.60	10.71
中	8.15	13.50	8.76
小	4.77	4.15	3.47

3　二元配置分散分析について

二元配置分散分析とは，二つの因子 A と B が存在し個々のデータのもっている情報を縮約して 1 つの値にした統計値への影響を調査する手法である．ここでは繰り返しのないデータを使用する．一般的な手法であるため一般式などは省略する．事例研究の平均地価を二元配置分散分析を用いて得られた残差が表 2 である．

表 2　二元配置分散分析を用いて得られた残差

	大	中	小
大	3.4511	1.6678	−5.1189
中	−3.0389	0.8378	2.2011
小	−0.4122	−2.5056	2.9178

これから県庁所在地人口率が小，人口が大である千葉県が少し外れ値になっていることがわかる．

4　Median Polish 法について

Median Polish 法は竹内 [2] に基づき事例研究のデータで説明する．R では medpolish 関数で計算が可能である．

表 3　Median Polish 法の計算手順その 1

23.6	0.31	0.00	−12.89
8.76	−0.61	4.74	0.00
4.15	0.62	0.00	−0.68
	0.31	0.00	−0.68
0.00	0.00	0.00	−12.21
0.68	−0.92	4.74	0.68
0.00	0.31	0.00	0.00

表 4　Median Polish 法の計算手順その 2

0.00	0.00	0.00	−12.21
0.68	−1.6	4.06	0.00
0.00	0.31	0.00	0.00
	0.00	0.00	0.00
0.00	0.00	0.00	−12.21
0.00	−1.6	4.06	0.00
0.00	0.31	0.00	0.00

1. 表 1 の各行の中央値（行中央値）を求め，表 3 左上に記入する．

2. 行ごとの各データから行中央値を引いた残差を求め，表 3 右上に記入する．
3. 手順 2 で求めた残差について列ごとの中央値 (列効果) を求め，表 3 右中央に記入する．
4. 手順 3 で求めた残差から列効果を引き，表 3 の右下に記入する．
5. 手順 4 で求めた残差について行ごとの中央値 (行効果) を求め，表 3 の左下に記入する．このとき表 3 の左下のすべての数値が 0 にならない場合，表 3 の右下の数値を元のデータとして用いて，この手順 1〜手順 5 すべての数値が 0 になるまで繰り返す．
(ここでは 0 になっていないので繰り返す.)
6. 表 3 と表 4 の列効果成分同士，行効果成分同士を足して新たな列効果と行効果を得る．このとき新たに得た行効果の中央値を総中央値とする．
7. 行効果から総中央値を引いて新たな行効果を得る．
この事例研究のデータでは総中央値が 9.44，行効果は上から 14.16，0，−5.29，列効果は 0.31，0，−0.68 となる．

事例研究の平均地価を Median Polish 法を用いて得られた結果，二元配置分散分析以上に千葉県が大幅に外れ値を取ることが分かった．

5　シミュレーション

事例研究から Median Polish 法が外れ値を含むデータに対して有効的な手段ではないかと考察されたが，実際にどのくらいの大きさの外れ値に有効な手段であるかをモンテカルロ法でシミュレーションを行った．

6　プログラムについて

小，中，大の効果 (今回は 0，1，2 で固定) を A,B の要素として，それぞれに与え，3 行 3 列の 9 分割したデータを作成し，平均 0，標準偏差 0.1 の正規乱数を誤差として加えた．その行列にさらに外れ値を乗せて，二元配置分散分析と Median Polish 法を同時に 100 回ずつ行い，その平均を求めるプログラムを製作した．外れ値を与える位置は 3 行 3 列の行列の角である (1,1) 成分，辺である (1,2) 成分，真ん中である (2,2) 成分の 3 種類に分けた．外れ値は 0.1 から 2.0 まで 0.1 ずつ増やした．

7　考察

(1,1) 成分，(1,2) 成分，(2,2) 成分の 3 つのパターンとも Median Polish 法の B 中が 0 をとり, 基準をあわせるために二元配置分散分析の A 中が 0 となるように基準化した．その結果の内，(1,1) 成分と (2,2) 成分について，横軸に外れ値の大きさ，縦軸に推定された効果の大きさを示したグラフが図 2,3 である．
二元配置分散分析は線形的にプロットされていくパターンか 0 をプロットし続けるパターン，Median Polish 法は全体的に疎らにプロットされていくパターンが多いことがわかった．しかし例外も存在する．図 2 の二元配置分散分析 A 大は，まだらにプロットされている. これは A 中が 0 となるように基準化し，A 大にも影響が及んだためだと考察された．また，図 2 の Median Polish 法 A 小も例外であり，x 軸の 1 付近で点の動き方が変化している．これは Median Polish 法は中央値を用いており，差が 1 で A 中と中央値が入れ替わるためだと考察された．

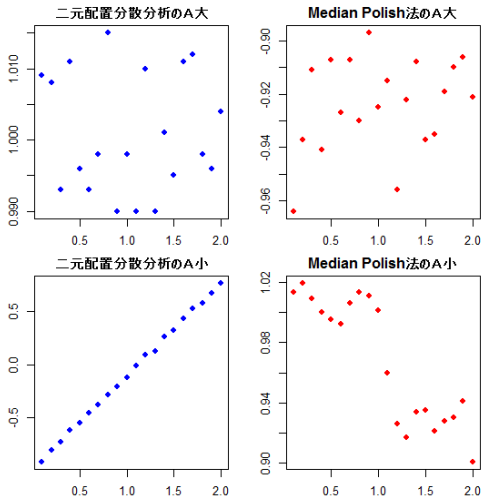

図 2　(1,1) 成分の結果

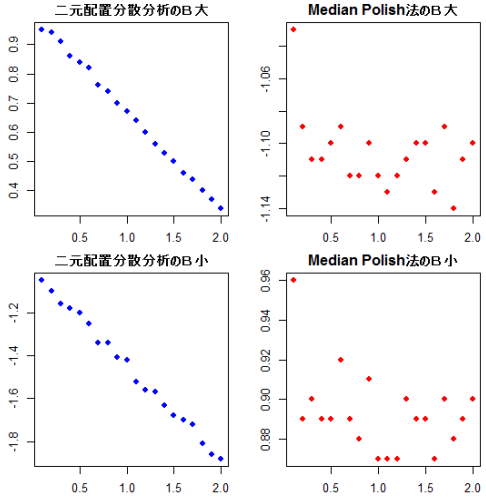

図 3　(2,2) 成分の結果

8　おわりに

シミュレーションを行った事で，例外も存在するが基本的には外れ値を含むデータに対し Median Polish 法は疎らに分散される事がわかった．

参考文献

[1] 観光庁：『統計情報・白書』,http://www.mlit.go.jp/kankocho/siryou/toukei/ranking.html, 2014.(2016/6 閲覧)
[2] 竹内啓：『統計学辞典』, 東洋経済新報社,1990.