

正準相関係数

1. 目的

2つのデータ列 $\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n$ の相関係数 r は以下のように計算できます。

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad \text{として}$$
$$r = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}}$$

一方 正準相関係数では、2群のデータ列間の相関を考慮します。
今 2群のデータ列が以下にあるとします。

- $\{x_i^1\}_{i=1}^n \cdots \{x_i^p\}_{i=1}^n$
- $\{y_i^1\}_{i=1}^n \cdots \{y_i^q\}_{i=1}^n$

第1群はp個のデータ列で、第2群はq個のデータ列です。

第1群、第2群それぞれのデータ列について以下のような、線形結合を考慮します。

- $\{X_i\}_{i=1}^n = a_1 \{x_i^1\}_{i=1}^n + a_2 \{x_i^2\}_{i=1}^n + \cdots + a_p \{x_i^p\}_{i=1}^n$
- $\{Y_i\}_{i=1}^n = b_1 \{y_i^1\}_{i=1}^n + b_2 \{y_i^2\}_{i=1}^n + \cdots + b_q \{y_i^q\}_{i=1}^n$

正準相関分析では、このように定義された2つのデータ列 $\{X_i\}_{i=1}^n, \{Y_i\}_{i=1}^n$ の相関係数が最大となるような係数列 $\{a_k\}_{k=1}^p, \{b_k\}_{k=1}^q$ を求めます。

そのうえで、 $\{X_i\}_{i=1}^n$ と $\{Y_i\}_{i=1}^n$ とから 相関係数を計算します。

ここでは 以下の例で説明します。

ある販売チェーンの広告費、販促費、新規入会者数、売上高のデータです。

店舗 No.	広告費(万円)	販促費(万円)	新規入会者数(人)	売上高(万円)
1	436	64	235	8,400
2	327	192	167	12,900
3	727	153	471	40,200
4	369	225	378	17,500
5	121	109	187	8,800
6	106	190	889	21,600
7	505	125	359	26,700
8	304	124	792	17,200

(広告費、販促費) と (新規入会者数、売上高) との 相関を考えます。まず、

- ・広告費 と 新規入会者数
- ・広告費 と 売上高
- ・販促費 と 新規入会者数
- ・販促費 と 売上高

との相関は高そうな気がしますが、相関行列は以下です。

変数名	広告費	販促費	新規入会者数	売上高
広告費(万円)	1.0	-0.1159	-0.2093	0.6807
販促費(万円)	-0.1159	1.0	0.2272	0.2548
新規入会者数(人)	-0.2093	0.2272	1.0	0.3541
売上高(万円)	0.6807	0.2548	0.3541	1.0

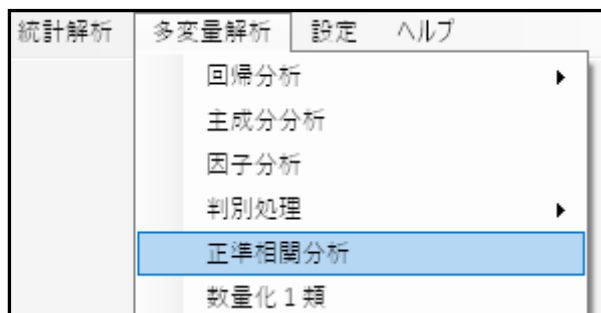
広告費と売上高との相関がせいぜい 0.68 で、それ以外は低いようです。

そこで、広告費 ($\{x_i^1\}_{i=1}^n$) と 販促費 ($\{x_i^2\}_{i=1}^n$) とで新たな変数 $\{X_i\}_{i=1}^n$ を、
新規者数 ($\{y_i^1\}_{i=1}^n$) と 売上高 ($\{y_i^2\}_{i=1}^n$) とで新たな変数 $\{Y_i\}_{i=1}^n$ を構成して、
相関係数を計算することにします。

2. 使用法

(1) メニューの選択

メニューの「多変量解析→正準相関分析」を選択します。



棄却域の確率を示します。
通常 5%を利用するので、
デフォルトで5が指定されて
いる。変更可能。

(2) パネルが表示されます。

計算結果が
表示される部分

(3) データの入力

パネルの下グリッド部分にデータを入力します。

☐ 表データを貼り付け
 ☐ 先頭行をラベルとして使用

	N0	ID	Value1	Value2
*				

データの入力は、表計算ソフトのデータをコピーしてグリッドに貼り付けます。
左の表のデータを右にコピーします。

店舗 No.	広告費 (万円)	販促費 (万円)	新規 入会 者数 (人)	売上高 (万円)
1	436	64	235	8,400
2	327	192	167	12,900
3	727	153	471	40,200
4	369	225	378	17,500
5	121	109	187	8,800
6	106	190	889	21,600
7	505	125	359	26,700
8	304	124	792	17,200



☐ 表データを貼り付け
 ☒ 先頭行をラベルとして使用

	N0	店舗No.	広告費(万円)	販促費(万円)	新規入会者数(人)	売上高(万円)
▶	1	1	436	64		
	2	2	327	192		
	3	3	727	153		
	4	4	369	225		
	5	5	121	109		

次に変数のグループ分けを定義します。

変数のグループ分け
 変数リスト
 第1群の変数リスト
 第2群の変数リスト

広告費(万円)
 販促費(万円)
 新規入会者数(人)
 売上高(万円)

< >

< >

広告費と販促費は 第1グループに、新規入会者数と売上高は第2グループに振り分けますが、左側の変数リストから項目を選択し、ボタン を利用して、定義します。

変数のグループ分け
 変数リスト
 第1群の変数リスト
 第2群の変数リスト

< >

< >

第1群と第2群には 共に 2つ以上の変数が定義される必要があります。
 どちらかが 1変数のみ という場合は認めていません。
 どちらかが1変数のみということは、回帰分析そのものになるからです。
 つまり その場合は、回帰分析を行ってください ということになります。

(4) 計算条件の指定

有意水準 α (%) : 5	計算実行
-----------------------	------

相関係数の検定の際に利用されます。
 “有意水準”には デフォルトで 5 が指定されています。変更できます。

(5) 計算実行

計算実行 ボタンを押すと計算されます。

(6) 計算結果

計算結果

データ数 : 8 第1正準相関係数 : 0.8657659 寄与率 : 0.7495506

T 値 : 4.893116 帰無仮説の採択域 : (-2.306727 , 2.306727)

P 値 (%) : 12.04 分布関数

結果 : 有意 : 帰無仮説(2つの要素は無相関)を棄却する

第1群の重み係数

変数名	重み
▶ 広告費(万円)	0.992312
販促費(万円)	0.2840115

第2群の重み係数

変数名	重み
▶ 新規入会者数...	-0.5387902
売上高(万円)	1.05454

正準スコア

店舗No.	<1>	<2>
▶ 1	-0.09575398	-0.730302
2	0.07304973	-0.103881
3	1.941888	2.17807
4	0.483576	-0.0579411
5	-1.484925	-0.585717
6	-1.095518	-0.701291

データ分布

計算結果が示されます。
 相関係数は 0.865 であり、寄与率は 0.750 です。
 ここでの寄与率は、相関係数の自乗であり、固有値でもありますが、
 正準変量が情報を要約している割合を表しています。

「T 値」、「帰無仮説の採択域」、「P 値」、「結果」は 相関係数の検定結果であり、
 この計算では、相関係数が 有意であること つまり「相関係数が0である」
 または「無相関である」ことを示す 帰無仮説が 棄却されたことを示します。

「第1群の重み係数」は、データ列 $\{X_i\}_{i=1}^n$ を定義する係数列 $\{a_1, a_2\}$
 で $\{0.992, 0.284\}$ となり、「第2群の重み係数」は、データ列 $\{Y_i\}_{i=1}^n$
 を定義する係数列 $\{b_1, b_2\}$ であり $\{-0.538, 1.054\}$ となりました。

右側の正準スコアは、 $\{a_1, a_2\}$ 、 $\{b_1, b_2\}$ を用いて計算される $i=1\sim 8$ についての

$$\cdot \{X_i\}_{i=1}^n = a_1\{x_i^1\}_{i=1}^n + a_2\{x_i^2\}_{i=1}^n$$

$$\cdot \{Y_i\}_{i=1}^n = b_1\{y_i^1\}_{i=1}^n + b_2\{y_i^2\}_{i=1}^n$$

を示します。

ただし、 $\{x_i^1\}_{i=1}^n$ 、 $\{x_i^2\}_{i=1}^n$ 、 $\{y_i^1\}_{i=1}^n$ 、 $\{y_i^2\}_{i=1}^n$ は入力値そのものではなく、それぞれが標準化された値を利用します。

この正準スコアを用いたデータの分布を表示すると、以下のようになります。

