システム工学　データ解析　資料

〈 例　題 〉

小売業の競争は激化している。デパート、スーパーマーケットなど、毎年前年同月比を落としているなか、これらの小売業では競争下での生き残りをかけてデータを分析し、売り上げを伸ばし、利益をあげようとしている。ここでは、回帰分析、主成分分析、クラスター分析を用いて小売店の店舗の売り上げ予測、店舗の特徴づけ、店舗の分類を行う例を解説する。小売店の店舗の売り上げ予測ができれば、新規店舗の出店に際して、精密な計画が立案できる。また、店舗の特徴づけができれば、その特徴にあわせた品揃えをすることによって売り上げを伸ばすことができる。

〈 解　説 〉

上記のルールに示したとおり，データ解析の目的は、データの裏に隠れている特徴を各種の統計的な手法によって見出していくことである。第 8 章までの OR のモデルには、必ずパラメータが存在していた。そして、これらのパラメータはすでに与えられたものとして解説されていた。しかし、実際の問題では、これらのパラメータは、与えられたものではなく、データから求めて行かなくてはならない場合が多い。そのようなときに頼りになるのは本章で述べるデータ解析の手法である。回帰分析は、ある変数 y（被説明変数と呼ばれる）が他のいくつかの量 x1, x2, …, xn（説明変数と呼ばれる）の線形関数で表されると仮定したときに説明変数の係数を求めて、y を x1, x2, …, xn の線形関数であらわす手法である。主成分分析は、いくつかの量 x1, x2, …, xn であらわされるデータを、より少ない x1, x2, …, xn の線形関数で特徴つける手法である。

9.1　回帰分析

本節では，回帰分析について説明する．まず一つの説明変数の線形式で、

被説明変数をあらわす単回帰分析について説明しよう。単回帰分析では被説明変数 y を説明変数 x の 1 次式、

$$y = ax + b \qquad (9.1)$$

であらわす。このような関係が厳密に成立するデータは実際にはなく、(x1, y1), …, (xn, yn)がデータだとすると、(9.1)式の右辺に誤差項 ε を加えた

$$y_i = ax_i + b + \varepsilon_i \qquad (9.2)$$

という関係が成り立つ。(9.1)式、(9.2)式で被説明変数 y を「よりよく」説明変数 x で説明できるように a, b を決定する。今、表 1 のようなデータが与えられているときに、単回帰分析を行ってみよう。「よりよく」説明するには、(9.2)式の ε i の 2 乗の和を最小にすればよい。

このような考え方で a、b を決定する方法を最小 2 乗法と呼ぶ。この 2 乗の和を G(a, b)とすると、

$$G(a,b) = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \qquad (9.3)$$

となる。G(a,b)を最小にする a, b をもとめるには、G(a,b)を a, b で偏微分して 0 とおいた連立方程式を解けばよい。

$$\frac{\partial}{\partial b} G(a,b) = -2\sum_{i=1}^{n} (y_i - ax_i - b) = 0 \qquad (9.4)$$

$$\frac{\partial}{\partial a} G(a,b) = -2\sum_{i=1}^{n} (y_i - ax_i - b)\, x_i = 0 \qquad (9.5)$$

(9.4)式、(9.5)式を整理すると、以下のように a, b の連立方程式になる。

$$(\sum_{i=1}^{n} x_i)a + nb = \sum_{i=1}^{n} y_i \qquad (9.6)$$

$$(\sum_{i=1}^{n} x_i^{\,2})\, a + (\sum_{i=1}^{n} x_i)\, b = \sum_{i=1}^{n} x_i y_i \qquad (9.7)$$

これらを解くと、

$$a=\frac{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}x_iy_i}{(\sum_{i=1}^{n}x_i)^2-n(\sum_{i=1}^{n}x_i{}^2)}\text{、}\quad b=\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)a=\bar{y}-a\bar{x}\qquad(9.8)$$

となる。

このようにaとbを決めたとき、説明変数で被説明変数をどのくらい良く説明できるかを以下の決定変数で示す。

$$R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\qquad(9.9)$$

ただし　$\hat{y}_i=ax_i+b,\quad \bar{\hat{y}}_i=\sum_{i=1}^{n}\hat{y}_i/n$　である。決定係数は被説明変数の観測値のうちの何パーセントが、説明変数によって説明されるかをあらわしている。

表 9.1 店舗の売り上げと床面積

店舗	売り上げ(万円)	売り場面積（平方メートル）
1	216	250
2	225	235
3	207	220
4	128	265
5	152	200
6	137	240
7	239	210
8	46	215

実用的には、80%程度であれば、十分に被説明変数は説明変数によって説明されていると解釈されている。

表 9.1 はあるチェーンストアの 8 店舗のある期間の売り上げと、売り場面積とをあらわしている。説明変数 x を売り場面積、被説明変数 y を被説明変

数として回帰分析を行ってみよう。

(9.8)式に従って a, b を計算すると、a=0.450, b=44.2 となる。また R2 は 0.503 となる。R2 の値は被説明変数の変動の 50 パーセント程度が説明変数によって説明されていることを示している。実際、この直線を x-y 平面上にあらわすと図 9.1 のようになる。これをどのように判断するかは実際の問題を解決する実務家の問題となる。一般には、50 パーセントという数値は、説明変数はあまりよく被説明変数を説明できていないとされる。そのような場合には、よりよく説明できる説明変数を新しく探すかモデル自体を見直すことになる。ここでは、あらたに説明変数を加える場合の説明をする。

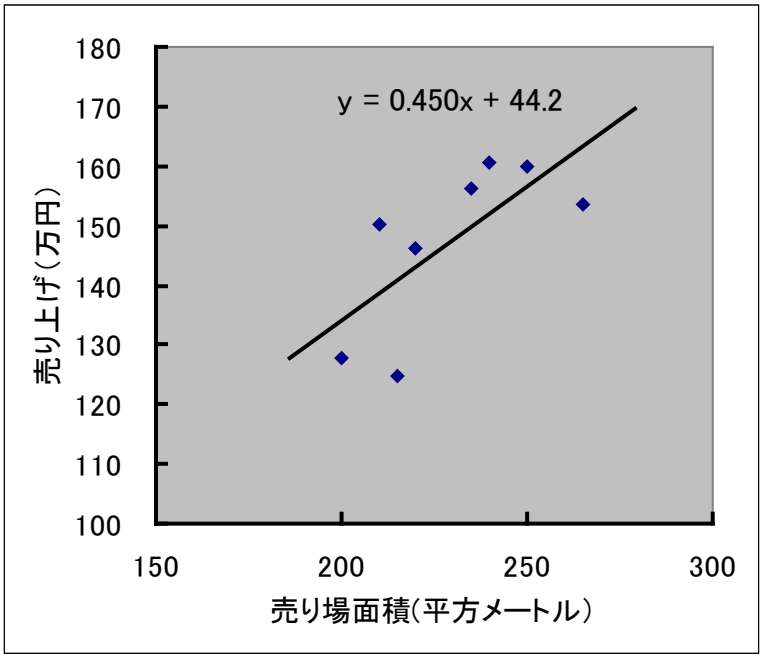

図 9.1

表 9.2 はあたらしい説明変数として、店舗に併設されている駐車場の面積を加えたデータである。この場合、説明変数を売り場面積 x1、駐車場面積 x2 として、被説明変数である売り上げ y を説明する回帰式を求めることになる。すなわち、

$$y = a_1 x_1 + a_2 x_2 + b \qquad (9.10)$$

という回帰式で a1、a2、b を説明変数が 1 つの場合と同じように求めればよ

い。(9.2)式と同様に誤差項を考えると、

$$y_i = a_1 x_{1i} + a_2 x_{2i} + b + \varepsilon_i \qquad (9.11)$$

という関係式が成り立つ。最小２乗法の考え方で、誤差項の２乗和を最小にするような a1、a2、b を求めるには、

表 9.2 店舗の売り上げと売り場面積、駐車場面積

店舗	売り上げ（万円）	売り場面積（平方メートル）	駐車場面積（平方メートル）
1	160	250	720
2	156	235	960
3	146	220	430
4	153	265	350
5	128	200	530
6	161	240	840
7	150	210	600
8	125	215	320

(9.3)式と同じように

$$G(a_1, a_2, b) = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - a_1 x_{1i} - a_2 x_{2i} - b)^2 \qquad (9.12)$$

を a1、a2、b で偏微分した式を 0 とおき、それらを a1、a2、b の連立方程式として解く。この連立方程式は、正規方程式と呼ばれている。この場合の正規方程式は、

$$(\sum_{i=1}^{n} x_{1i}) a_1 + (\sum_{i=1}^{n} x_{2i}) a_2 + nb = \sum_{i=1}^{n} y_i \qquad (9.13)$$

$$(\sum_{i=1}^{n} x_{1i}^{\,2}) a_1 + (\sum_{i=1}^{n} x_{1i} x_{2i}) a_2 + (\sum_{i=1}^{n} x_{1i}) b = \sum_{i=1}^{n} x_{1i} y_i \qquad (9.14)$$

$$(\sum_{i=1}^{n} x_{1i} x_{2i}) a_1 + (\sum_{i=1}^{n} x_{2i}^{\,2}) a_2 + (\sum_{i=1}^{n} x_{2i}) b = \sum_{i=1}^{n} x_{2i} y_i \qquad (9.15)$$

となる。a1、a2、b を数式で表現するのは煩瑣になるので省略する。

正規方程式は、式の数が多いときにはコンピュータによって計算しなくてはならない。実際には統計解析のソフトウェアによって回帰分析が行われる。現在、普及しているマイクロソフト社の表計算ソフトウェア、エクセルにも統計解析の機能があり、回帰分析も行える。

エクセルを用いて計算すると、a1、a2、b の値は、それぞれ、0.394、0.031、37.9 となる。決定係数 R2 は、0.778 となり、被説明変数の変動の約 80 パーセントが説明変数の変動で説明できることになる。

回帰分析はエクセルなどの表計算ソフトウェアや、さまざまな統計解析のソフトウェアによって、ごく手軽に実行できるようになっている。実際、回帰分析は OR のモデルのパラメータを決定するのに役立っている。例えば、ここの例で取り扱ったように、ある小売店の売り上げが売り場面積、駐車場面積、周辺の人口などであらわされれば、出店計画の OR モデルを作成するのに役立つ。

9.2　主成分分析

本節では主成分分析について説明する。表 9.1 では、ある小売店、8 店舗の売り上げと売り場面積が与えられている。これらをそれぞれ x1, x2 とする。いま、x1 と x2 であらわされているこの 8 店舗の特徴を、「よりよく」あらわす x1, x2 の 1 次結合を求めることを考える。このような 1 次結合は主成分と呼ばれている。「よりよく」の意味は、この 1 次結合の分布を考えると 8 店舗の分布が最もばらつくと考えることができる。すなわち、分散が最も大きくなるような 1 次結合を求めればよい。ここでは、このような 1 次結合を求めることが、相関行列の固有ベクトルを求めることに帰結し、分散が固有ベクトルになることを説明しよう。

では、店舗数を n として一般的な式を求めよう。x1, x2 を平均 0 、分散 1 の変数 u1, u2 に変換しておく。すなわち、i を各店舗をあらわす添え字として、

$$\bar{x}_1 = \frac{\sum_{i=1}^{n} x_{1i}}{n}、\quad s_1 = \sqrt{\frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}{n-1}}、\quad \bar{x}_2 = \frac{\sum_{i=1}^{n} x_{2i}}{n}、\quad s_2 = \sqrt{\frac{\sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2}{n-1}}$$

を計算し、

$$u_{1i} = \frac{x_{1i} - \bar{x}_1}{s_1} \qquad (9.16)$$

$$u_{2i} = \frac{x_{2i} - \bar{x}_2}{s_2} \qquad (9.17)$$

とおく。ここで、平均０、分散１であるから、

$$\frac{\sum_{i=1}^{n} u_{1i}}{n} = 0、\quad \frac{\sum_{i=1}^{n} u_{1i}^{\,2}}{n-1} = 1、\quad \frac{\sum_{i=1}^{n} u_{2i}}{n} = 0、\quad \frac{\sum_{i=1}^{n} u_{2i}^{\,2}}{n-1} = 1$$

である。また、u1 と u2 の相関係数 r12 は、

$$r_{12} = \frac{\sum_{i=1}^{n} u_{1i} u_{2i}}{n-1} \qquad (9.18)$$

である。このような前提のもとで、u1, u2 の 1 次結合

$$z_1 = a_1 u_1 + a_2 u_2 \qquad (9.19)$$

を考え、a1, a2 を z1 の分散が最大になるように定める。z1 の分散は、

$$V_1 = \frac{\sum_{i=1}^{n} z_{1i}^{\,2}}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (a_1 u_1 + a_2 u_2)^2 \qquad (9.20)$$
$$= a_1^{\,2} + a_2^{\,2} + 2\, r_{12} a_1 a_2$$

となる。a1, a2 が大きくなれば、V1 は限りなく大きくなってしまうので、

$$a_1^{\,2} + a_2^{\,2} = 1 \qquad (9.20)$$

の条件のもとで V1 を最大化する a1, a2 を求めよう。この問題は非線形計画法の問題となる。実際、a1, a2 を求めるには、非線形計画法のラグランジュ乗数法という方法を用いる。ラグランジュ関数は、

$$L_1 = a_1^{\,2} + a_2^{\,2} + 2 r_{12} a_1 a_2 - \lambda (a_1^{\,2} + a_2^{\,2} - 1) \qquad (9.21)$$

となる。L1 を a1, a2 で偏微分して 0 とおくと、

$$\frac{\partial L_1}{\partial a_1} = 2a_1 + 2r_{12}a_2 - 2\lambda a_1 = 0$$
$$\frac{\partial L_1}{\partial a_2} = 2a_2 + 2r_{12}a_1 - 2\lambda a_2 = 0 \qquad (9.22)$$

この連立方程式を行列とベクトルであらわすと、

$$\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \qquad (9.23)$$

となる。これを見れば、(a1, a2)'は相関行列の固有ベクトルであり、λは相関行列の固有値であることがわかる。すなわち、(a1, a2)'を求めることは、相関行列の最大固有値とその固有値に対する大きさ 1 の固有ベクトルを求めることに帰結する。(9.**)式をもう少し詳しく見てみよう。

$$\begin{pmatrix} a_1 & a_2 \end{pmatrix}\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = {a_1}^2 + {a_2}^2 + 2r_{12}a_1a_2 = V_1$$
$$\lambda\begin{pmatrix} a_1 & a_2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda({a_1}^2 + {a_2}^2) = \lambda$$

であるから、V1 の最大値はλとなることがわかる。

さて、このようにして求められた主成分は第 1 主成分と呼ばれる。主成分を求めることは、相関行列の固有値と固有ベクトルを求めることに帰結していた。第 1 主成分の固有値の次に大きな固有値とそれに対応する固有ベクトルを求めることができれば、第 1 主成分の次によりよくデータの特徴をあらわす 1 次結合を求めることができる。このような 1 次結合は第 2 主成分と呼ばれる。以下では、第 2 主成分を求める方法を説明しよう。方法は第 1 主成分を求めたのと同じである。ただし、第 2 主成分を与える固有ベクトルは、第 1 主成分を与える固有ベクトルに直交すると仮定しておく。第 2 主成分を

$$z_2 = b_1u_1 + b_2u_2 \qquad (9.24)$$

とおき、Z2 を最大にする b1, b2 を求める。Z2 の分散は、

$$V_2 = {b_1}^2 + {b_2}^2 + 2\,r_{12}b_1b_2 \qquad (9.25)$$

第 1 主成分を求めたときと同じように、以下の条件のもとで b1, b2 を求める。

$$b_1^2 + b_2^2 = 1 \qquad (9.26)$$

$$a_1 b_1 + a_2 b_2 = 0 \qquad (9.27)$$

2 番目の条件は、２つの主成分が直交する条件である。

ラグランジュ関数は、

$$L_2 = b_1^2 + b_2^2 + 2r_{12}b_1 b_2 - \lambda(b_1^2 + b_2^2 - 1) - \eta(a_1 b_1 + a_2 b_2) \qquad (9.28)$$

である。L2 を b1, b2 で偏微分して 0 とおくと、

$$\begin{aligned} \frac{\partial L_2}{\partial b_1} &= 2b_1 + 2r_{12}b_2 - 2\lambda b_1 - \eta a_1 = 0 \\ \frac{\partial L_2}{\partial b_2} &= 2b_2 + 2r_{12}b_1 - 2\lambda b_2 - \eta a_2 = 0 \end{aligned} \qquad (9.29)$$

第 1 主成分を求めたときと同じように行列とベクトルで表現すると、

$$\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \lambda\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \frac{\eta}{2}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \qquad (9.30)$$

となる。(a1, a2)’との内積をとると、

$$\begin{pmatrix} a_1 & a_2 \end{pmatrix}\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \left(\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}\right)'\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \lambda\begin{pmatrix} a_1 & a_2 \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \lambda(a_1 b_1 + a_2 b_2) = 0$$

$$\lambda\begin{pmatrix} a_1 & a_2 \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \frac{\eta}{2}\begin{pmatrix} a_1 & a_2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \frac{\eta}{2}(a_1^2 + a_2^2) = \frac{\eta}{2}$$

従って、η =0 となることがわかる。すなわち、第 2 主成分を求める問題は、相関行列の固有値とそれに対応する固有ベクトルを求める問題に帰着する。ただし、第 1 主成分と同じ a1, a2 では、第 1 主成分と直交する条件は満たされないので、第 2 主成分は、相関行列の 2 番目に大きい固有値に対応する大きさ 1 の固有ベクトルとなる。

第 1 主成分、第 2 主成分の分散は最大固有値と、次に大きい固有値である。これらを$\lambda 1$, $\lambda 2$ であらわす。このとき、それぞれの主成分がデータをどのくらい良く説明しているかは、寄与率と呼ばれる量であらわされる。第 1 主成分の寄与率は、$\lambda_1 / (\lambda_1 + \lambda_2)$、第 2 主成分の寄与率は、$\lambda_2 / (\lambda_1 + \lambda_2)$である。主成分の寄与率を、第 1 主成分、第 2 主成分の順に足したものを累積寄与率という。

また、第 1 主成分、第 2 主成分と、元の変数 x1, x2 との相関係数は因子負荷量と呼ばれる。因子負荷量は、第 1 主成分、第 2 主成分を解釈するのに役立つ。さらに、各サンプルに対して、第 1 主成分、第 2 主成分を計算したものを主成分得点と呼ぶ。各サンプルの第 1 主成分得点、第 2 主成分得点を軸として、散布図を描くと、各サンプルの特徴を考えることができる。また、その特徴に従って、各サンプルをいくつかのグループに分類することもできる。

表 9.1 の例で、第 1 主成分と第 2 種主成分を計算すると、表 9.2 のようになる。

表 9.2　ある小売店の売り上げと床面積の主成分分析

	第 1 主成分	第 2 主成分
売り上げ	0.9246	−0.3809
床面積	0.9246	0.3809
固有値	1.7097	0.2902
寄与率	85.489	14.511
累積寄与率	85.489	100

この結果を見ると、第 1 主成分は、売り上げも床面積も正なので、店の規模をあらわし、第 2 主成分は、売り上げが負、床面積が正なので、床面積に対しての売り上げの大きさをあらわしていると解釈できる。この主成分に対して主成分得点を計算すると、表 9.3 のようになる。

表 9.3　各店舗の主成分得点

店舗	第 1 主成分	第 2 主成分
1	1.3890	0.0317
2	0.6758	-0.2883
3	-0.3844	-0.2613
4	1.5540	0.8999
5	-2.0768	0.0533

6	1.0788	-0.3469
7	-0.5056	-0.8290
8	-1.7309	0.7407

主成分得点の散布図は図 9.2 のようになる。

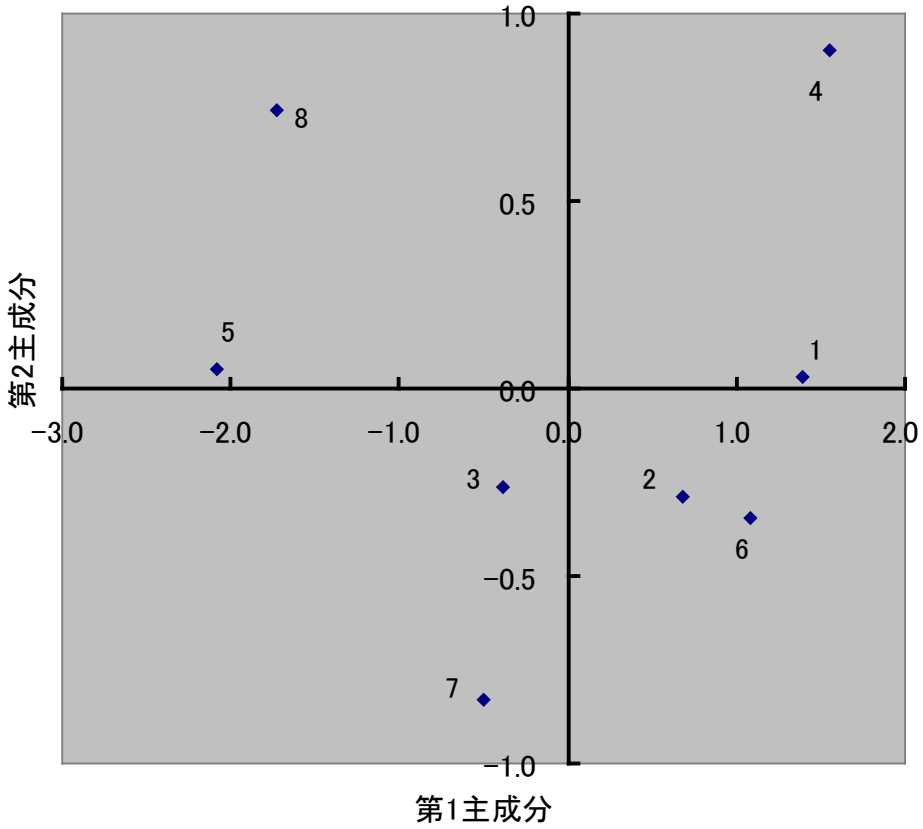

図 9.2

例えば、4 の店舗は、規模も大きいが、売り上げに比べて床面積も大きい、7 の店舗は、規模は平均的だが、売り上げに比べて床面積が小さい、８の店舗は、規模は小さく、売り上げに比べて床面積が大きいというような解釈ができる。

さて、この小売店の店舗で、商品グループごとの売り上げが表 9.4 のよう

であるとする。このデータから８つの店舗について、主成分分析を行ってみよう。

表 9.4　商品グループごとの売り上げ

店舗	グループ1	グループ2	グループ3	グループ4
1	34	18	21	87
2	18	57	63	18
3	24	20	52	50
4	25	19	25	85
5	30	1	14	84
6	33	23	59	46
7	7	29	32	82
8	15	23	44	43

主成分分析を行った結果は表 9.5 のようになる。

表 9.5　店舗のグループごとの売り上げに対する主成分分析

	第 1 主成分	第 2 主成分	第 3 主成分	第 4 主成分
グループ1	0.4741	0.8603	0.1869	-0.0038
グループ2	-0.8922	-0.1592	0.4225	-0.0026
グループ3	-0.9192	0.3051	-0.1514	0.1971
グループ4	0.9219	-0.2923	0.1617	0.1959
固有値	2.7159	0.9441	0.2625	0.0772
寄与率	67.899	23.604	6.5641	1.9318
累積寄与率	67.899	91.504	98.068	100

この結果を見ると、第 2 主成分までの累積寄与率が 91%を越えており、８つの店舗に対して、よりよく特徴づけができることが期待できる。第 1 主成分はグループ１、４の商品とグループ２，３の商品の対比をあらわし、第 2 主成分はグループ１，３の商品とグループ２，４の商品の対比をあらわす。例えば、グループ１が食品、グループ２が電気製品、グループ３が園芸用品、

グループ４が日用品をあらわすとすると、第 1 主成分はいわゆる買回り品の割合が高いかどうか。第 2 主成分は食品と園芸用品が売れているかどうかなので、女性客の割合が高いかどうかをあらわすと解釈できる。

主成分得点は表 9.6 のようになる。

表 9.6　主成分得点

店舗	第 1 主成分	第 2 主成分
1	1.7177	0.5352
2	-3.2001	0.0583
3	-0.5280	0.5330
4	1.2318	-0.2754
5	2.3550	0.1669
6	-0.6692	1.5449
7	-0.0448	-2.0572
8	-0.8623	-0.5060

主成分得点の散布図は図 9.3 のようになる。

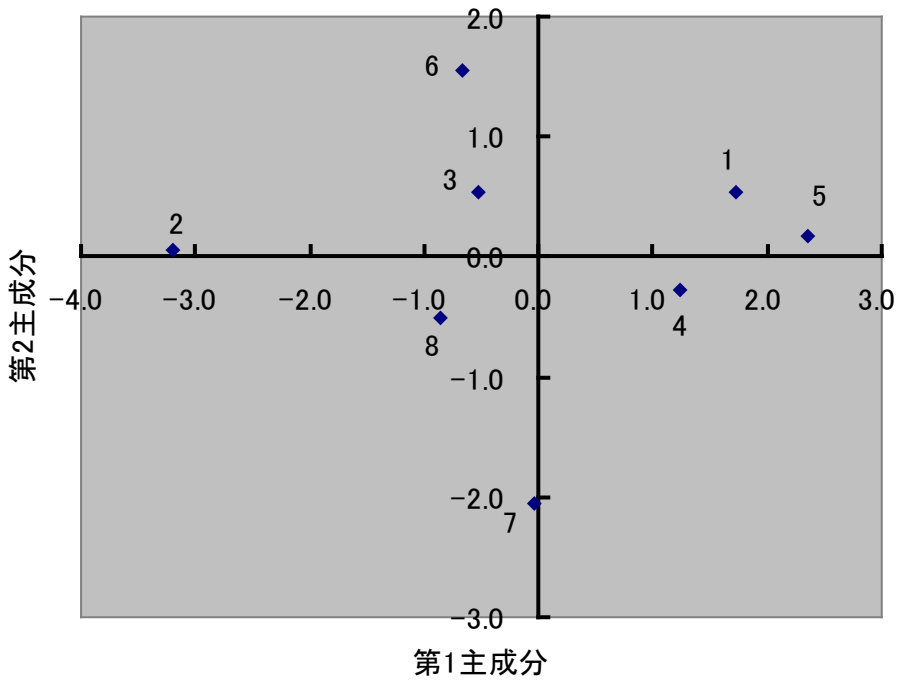

図 9.3

この図を見ると、店舗１、２、５，６，７が特徴のある店舗であると解釈できる。

このようにデータ解析により店舗の特徴づけが得られ、これをもとにオペレーションズ・リサーチのモデルを作成することができる。回帰分析、主成分分析の計算もエクセルで計算できるソフトウェアが数多く発売されており、これによって容易にデータ解析を行うことができる。

〈演習問題〉

- 表 9.1 の駐車場面積を説明変数、売り上げを被説明変数として回帰分析しなさい。
- 図 9.3 から店舗１，２，５，６，７がどのような特徴があるか解釈しなさい。

参考文献

1）P.G. ホーエル，浅井　晃、村上　正康訳、『初等統計学』，培風館，1981年.