相関係数のグラフ表現と、その重回帰分析 への応用について

石 村 貞 夫*, 玉 村 千 治**

A GRAPHICAL REPRESENTATION OF CORRELATION COEFFICIENTS AND ITS
APPLICATION TO THE MULTIPLE REGRESSION ANALYSIS

Sadao, ISHIMURA Chiharu TAMAMURA

It is the purpose of the data analysis to clarify the characteristic of the phenomena and to find the meaningful relations among them. From the point of view, a graphical description of the data is effective for the understanding of the structure of them and, in particular, is important for the multiple analysis.

In this paper, we consider the method of the graphical description of the correlation coefficient, the multi correlation coefficient and the partial correlation coefficient in the multiple analysis. It makes useful for the visualization of the complex relationship among the data. Moreover it applies to the choice of the variables and the multicollinearity.

1. 序

データ解析の目的は、現象の特性をより明確にし、意味のある因果関係を発見することにある。その点、データを数値そのもので見るよりも、一目でわかるグラフにすることは、データの構造を理解するのに、より効果的である。特に多変量解析においては、データのグラフ表現は重要であり、多変量データのグラフ表現に関する多くの試みが、脇本 (1979) や馬場 (1983) 等によってなされている。

この論文においては、以上の観点から、多変量解析における目的変数と説明変数の単相関係数、重相関係数、偏相関係数を、グラフにより平面上に表現する方法について述べる。このことにより、目的変数と説明変数の因果関係を視覚的にとらえることが出来る。またこの応用として、重回帰分析における変数選択や多重共線性の問題についても述べる。

2. 相関係数のグラフ表現

$p+1$ 変数の n 組のデータ $(y_i, x_{1i}, \dots, x_{pi})$ ($i=1, 2, \dots, n$) が与えられたとき、一つの目的変数 Y と p 個の説明変数 $X_1, X_2, \dots, X_p$ の間に定義される単相関係数、重相関係数、偏相関係数を平面上にグラフ化する。

2.1 単相関係数と重相関係数の幾何学的意味

二変数 Y, X_i の単相関係数 r_{Yi} は、ベクトル空間上の対応する二つのベクトル $\underline{x}_i = (x_{i1}, x_{i2}, \dots, x_{in})$ と $\underline{y} = (y_1, y_2, \dots, y_n)$ の間のなす角 φ の $\cos \varphi$ として表わされる。すなわち、分散を $s_y^2, s_{x_i}^2$ 、共分散を s_{Yi} とし、ベクトル $\underline{x}_i, \underline{y}$ における長さを $\|\underline{x}_i\|, \|\underline{y}\|$ 、内積を $(\underline{y}, \underline{x}_i)$ とすると、

$$r_{Yi} = \frac{s_{Yi}}{s_i s_Y} = \frac{(\underline{y}, \underline{x}_i)}{\|\underline{y}\| \cdot \|\underline{x}_i\|} = \cos \varphi$$

となる。

次に、 $p+1$ 変数 Y と $X_1, \dots, X_p$ の重相関係数 $R_{Y(1,2,\dots,p)}$ について考える。そこで $X = a_1 X_1 + a_2 X_2 + \dots + a_p X_p$ として、 X と Y の単相関係数を r_{XY} とすると、これは、ベクトル空間において対応するベクトル $\underline{x}_1, \dots, \underline{x}_p$ の張る部分空間に属するベクトル $\underline{x} = a_1 \underline{x}_1 + a_2 \underline{x}_2 + \dots + a_p \underline{x}_p$ と $\underline{y}$ のなす角の大きさ ϕ' の $\cos \phi'$ として表わされる。この値の最大となる場合、す

* 千葉工業大学非常勤講師 (Chiba Institute of Technology)

** アジア経済研究所 (Institute of Developing Economies)

なわち、ベクトル $\underline{y}-\underline{x}$ が、ベクトル $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p$ のすべてと直交している場合、 Y と $X_1 \dots X_p$ の重相関係数となる。これを $R_{Y(1,2,\dots,p)}$ と表わすことにする。ベクトル $\underline{y}$ を、ベクトル $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p$ の張る部分空間に正射影し、それと $\underline{y}$ とのなす角を ψ としたとき、

$$R_{Y(1,2,\dots,p)} = \cos \psi$$

である。

以上のことから分るように、単相関係数、重相関係数の概念を視覚的にとらえるためには、角度が適していると思われる。

単相関係数と重相関係数の性質

(1) ある二つの説明変数 X_i, X_j に対して

$$r_{ij} = 1$$

が成立するならば、対応するベクトル $\underline{x}_i$ と $\underline{x}_j$ は、同じ向きのベクトルとなるので、目的変数 Y に対応するベクトル $\underline{y}$ に対して $\underline{y}$ と $\underline{x}_i$ のなす角の大きさと、 $\underline{y}$ と $\underline{x}_j$ のなす角の大きさは等しくなる。よって

$$r_{Yi} = r_{Yj}$$

となる。

(2) 逆に、

$$r_{Yi} = r_{Yj} = R_{Y(1,2,\dots,p)}$$

が成立するならば、

$$r_{ij} = 1$$

となる。なぜならば、ベクトル $\underline{y}$ をベクトル $\underline{x}_1, \dots, \underline{x}_p$ の張る部分空間に正射影したとき、そのベクトルを $\underline{y}'$ とすれば、 $r_{Yi} = R_{Y(1,2,\dots,p)}$ より、 $\underline{y}$ と $\underline{y}'$ のなす角の大きさと、 $\underline{y}$ と $\underline{x}_i$ のなす角の大きさは等しくなる。同様に、 $r_{Yj} = R_{Y(1,2,\dots,p)}$ より、 $\underline{y}$ と $\underline{y}'$ のなす角の大きさと、 $\underline{y}$ と $\underline{x}_j$ のなす角の大きさも等しい。よって、 $\underline{y}'$ と $\underline{x}_i$ は同じ向きのベクトルとなり、 $\underline{y}'$ と $\underline{x}_j$ も同じ向きのベクトルとなる。

2.2 相関係数のグラフ表現法

目的変数 Y と、説明変数 $X_1, \dots, X_p$ の間で定義される単相関係数、重相関係数、偏相関係数を角の大きさや、線分の長さに対応させて、同一平面上にグラフ表現し、変数間の相関の度合を視覚により判定することを考える。

グラフ化する手順

手順1. 重相関係数を $R_{Y(1,2,\dots,p)} = \cos \psi$ とする。水平な線分 AB を引く。これを横軸と呼ぶ。横軸上的一点 O から時計回りに、横軸と ψ のなす角をもつ線分 OP をひく。 OP を斜軸と呼ぶ。斜軸は、横軸の長さよりも、少し短かくした方が、見易くなる。(図1参照)

手順2. 説明変数 X_i に対して、 Y との偏相関係数

を $r_{Yi,1,\dots,i-1,i+1,\dots,p}$ とする。斜軸 OP 上に点 Q_i を、点 O から長さ $\overline{OP} \cdot (1 - r_{Yi,1,\dots,i-1,i+1,\dots,p})/2$ のところにとる。(図2参照) 特に、 $r_{Yi,1,\dots,i-1,i+1,\dots,p} = 1$ ならば、点 Q_i は点 O に一致し、 $r_{Yi,1,\dots,i-1,i+1,\dots,p} = -1$ ならば、点 Q_i は点 P に一致する。

手順3. 単相関係数を $r_{Yi} = \cos \varphi_i$ とする。点 Q_i を中心にして、線分 PQ_i から時計まわりと反時計まわりに、それぞれ角の大きさ φ_i となるように、線分 $Q_i R_{i1}$, $Q_i R_{i2}$ をひく。線分 $Q_i R_{i1}$ と $Q_i R_{i2}$ の長さは等しくする。このようにして描いたものが、図3の矢印である。

この操作をすべての説明変数についておこない、図4のようなグラフを描く。

注意. 説明変数の数が多い場合には、一つのグループの変数の数を、4~5程度になるように、共通点を

図1. 重相関係数の角度による描き方

図2. 説明変数に対する偏相関係数の長さによる描き方

図3. 説明変数に対する単相関係数の角度による描き方

図4. 各説明変数に対するグラフ表現の描き方

もったいくつかのグループに分類する。次に、各グループに対して、手順1.2.3.の操作をくり返し、グラフ表現をする。このとき、目的変数と各説明変数グループの重相関係数の大きい順に、左からならべると、見易くなる。例3の図8を参照のこと。

2.3 相関係数のグラフ表現の見方と特徴

ここでは、具体例を用いて、相関係数のグラフ表現の見方と特徴について述べる。

例1.

例1のデータを用い、単相関係数と偏相関係数の見方を示す。図5は、表1をグラフ表現したものである。

(1) 図5から単相関係数について読み取れることは、 X_1 のように矢印が下向きの場合は、 X_1 と Y の正の相関が高く、反対に X_2 のように矢印が上向きの場合には、 X_2 と Y は負の相関となっている。また、 X_3 のように斜軸とほぼ垂直になっている場合は、 X_3 と Y は無相関であることを示す。

表1. 例1のデータ行列

	X_1	X_2	X_3	Y
1	4.000	3.000	7.000	1.000
2	3.000	7.000	3.000	2.000
3	3.000	8.000	2.000	3.000
4	4.000	3.000	8.000	4.000
5	6.000	2.000	2.000	5.000
6	9.000	3.000	6.000	6.000

表3. 例2のデータ行列

	X_1	X_2	X_3	Y
1	5.000	4.000	3.000	2.000
2	4.000	4.000	7.000	2.000
3	4.000	2.000	5.000	3.000
4	4.000	4.000	4.000	4.000
5	6.000	6.000	2.000	5.000
6	9.000	9.000	2.000	6.000

表4. 表3のデータ行列から得られる相関行列

	X_1	X_2	X_3	Y
X_1	1.000	0.945	-0.716	0.788
X_2	0.945	1.000	-0.650	0.799
X_3	-0.716	-0.650	1.000	-0.715
Y	0.788	0.799	-0.715	1.000

図5. 表1のグラフ表現

(2) Y との偏相関係数について読み取れることは、 X_1 のように矢印の先端が横軸に近いほど、 X_1 と Y の正の偏相関が高く、 X_3 のように矢印の先端が横軸から離れるほど、 X_3 と Y の負の偏相関が高いということである。また、 X_2 のように斜軸の中心近くに矢印の先端がある場合には、 X_2 と Y の偏相関が小さいことを示す。

(3) (1)と(2)を総合すると、説明変数 X_1, X_2, X_3 において X_1 は下向きの鋭い矢印をもち、その矢印の先端は横軸にもっとも近いので、 X_1 は Y と重要な相関関係にあると考えられる。

表2. 表1のデータ行列から得られた相関係数表

	Yとの単相関係数		Yとの偏相関係数		重相関係数
X_1	0.807	36.2°	0.755		
X_2	-0.427	115.3°	-0.039	0.830	33.9°
X_3	-0.040	92.3°	-0.299		

表 5. 表3のデータ行列から得られた相関係数表

	Yとの単相関係数		Yとの偏相関係数	重相関係数	
X ₁	0.788	37.9°	-0.005	0.839	32.9°
X ₂	0.799	36.9°	0.328		
X ₃	-0.715	135.6°	-0.398

図 6. 表3のグラフ表現

表 6. 昭和57年度総理府統計局『社会生活統計指標』をもとにして得られた、単相関係数、偏相関係数、重相関係数

U: 運動環境グループ $\rho(Y, U_1 \cdots U_4) = 0.494$ 角度 = 60°

	説 明 変 数	Yとの単相関係数		Yとの偏相関係数
U ₁	人口100万人当たりの運動広場数	-0.387	112°	-0.866
U ₂	人口100万人当たりのテニスコート数	-0.277	106°	0.607
U ₃	人口100万人当たりの体育館数	-0.472	118°	-0.806
U ₄	人口100万人当たりのプール数	-0.359	111°	-0.994

E: 教育環境グループ $\rho(Y, E_1 \cdots E_5) = 0.571$ 角度55°

	説 明 変 数	Yとの単相関係数		Yとの偏相関係数
E ₁	歳出決算総額による教育費割合	0.141	82°	0.067
E ₂	児童一人当たりの児童福祉費	-0.221	102°	0.019
E ₃	住民一人当たりの教育費	-0.536	122°	-0.842
E ₄	高等学校進学率	0.014	89°	0.627
E ₅	人口100万人当たりの青少年教育施設数	-0.175	100°	-0.074

K: 家庭環境グループ $\rho(Y, K_1 \cdots K_5) = 0.730$ 角度=43°

	説 明 変 数	Yとの単相関係数		Yとの偏相関係数
K ₁	一世帯当たり一ヶ月間の娯楽費割合	0.124	83°	0.863
K ₂	核家族世帯割合	0.485	60°	0.996
K ₃	最低居住水準以上の世帯割合	-0.557	124°	-0.995
K ₄	一世帯当たり一ヶ月間の食料費	0.498	60°	-0.138
K ₅	平均消費性向	0.315	72°	0.997

S: 自然環境グループ $\rho(Y, S_1 \cdots S_5) = 0.935$ 角度=20°

	説 明 変 数	Yとの単相関係数		Yとの偏相関係数
S ₁	可住地面積1km当たり人口	0.907	25°	-0.951
S ₂	自然公園面積割合	0.140	82°	-0.997
S ₃	工業専用地域面積割合	-0.071	94°	0.991
S ₄	人口10万人当たりの映画館数	-0.172	100°	0.587
S ₅	人口10万人当たりの派出所数	-0.607	127°	-0.304

例 2.

例2のデータを用いて、説明変数間の相関の度合を、図6から読み取ることが出来ることを示す。ただし、図6は、表5をグラフ表現したものである。

(4) X_1 と X_2 のように矢印が平行の場合には、 X_1 と X_2 の正の相関が高く、さらにそれらが横軸とも平行の場合には、単相関係数と重相関係数の性質(2)より、 X_1 と X_2 の単相関係数はほとんど1に近くなる。

このことは、表4からも確かめられる。

(5) X_2 と X_3 (又は、 X_1 と X_3)のように、矢印が互いに反対に向いているときには、 X_2 と X_3 (又は、 X_1 と X_3)の負の相関が高い。このことは、表4からも確かめられる。

(6) 散布図と比較することにより、相関係数のグラフ表現の特徴を明確にする。図7は、表5を散布図に表わしたものである。

図6のグラフ表現においては、単相関係数、重相関係数、偏相関係数を一度に平面上に表わすことが出来るが、図7の散布図においては、重相関係数をなんらかの方法で表わしておかねばならない。また、(4)や(5)で述べたように説明変数間の相関の度合を、この散布図から読みとることは難しい。

例3. 表6は、昭和57年度総理府統計局『社会生活統計指標』をもとにして、少年非行(年間少年刑法犯

図7. 表5の散布図

検挙人員) におよぼすいろいろな要因を調べたものである。目的変数 Y は、年間少年刑法犯検挙人員をとり、説明変数は、人口100万人当たりの運動広場等、19の変数からなっている。この例では、変数の数が多いので、社会生活統計指標の分類に従って、全体を運動環境グループ、教育環境グループ、家庭環境グループ、社会環境グループの四つに分類し、各グループごとに単相関係数、重相関係数、偏相関係数を求めた。その結果をグラフ表現したものが、図8である。

図8において、少年非行に影響を与えていると思われる要因と、非行防止に役立っていると思われる要因について考える。少年非行に影響を与えていると思われる要因は、変数の矢印が下向きで、横軸に近いものであるから、それにあてはまるものとしては、 K_2 と K_5 がある。逆に、非行防止に役立っていると思われる要因は、変数の矢印が上向きで横軸から離れているものである。ところで、 E_3 、 K_3 、 S_5 、 U_1 、 U_3 、 U_4 がある。ところで、 E_1 、 E_2 、 E_5 については、 E の斜軸とほとんど直交し、斜軸の中心近くにあるので、これらは、非行にあまり関係のない変数であろうと思われる。それぞれの環境グループで考えると、 S の斜軸が横軸の方へもっとも傾いているので、 S のグループは、非行と強い関係があるのではないかとと思われるが、これは、説明変数全体の中で、 S_1 の矢印が最も鋭いためである。このグラフ表現からは、 E_2 や E_5 (又は、 U_1 や U_4)のように、単相関係数のわずかの差までを、読み取ることは出来ない。しかしながら、変数の持つ意味や、グループごとの傾向を全体の中でとらえる場合には、相関係数の表6をみるよりも、図8のグラフ表現の方が適し

図8. 表6のグラフ表現

ているのではないと思われる。

3. 重回帰分析への応用

重回帰分析の考え方は、多変量解析など多くの統計解析の基礎を与えており、その適用も自然科学から社会科学の分野まで広範囲に渡っている。重回帰分析では、説明変数と目的変数との関係が評価され、その関係に基づく予測や制御が検討される。しかしながら、問題点はその説明変数をどのように決定するか、又、どのように選べば最も効率的であるかということにある。これまでの変数選択は、奥野(1971)によれば、最重要と認められた一つの変数を決め、そこから逐次変数を増加してゆく方法や、初めに全変数を決め、逆に逐次変数を減少してゆく方法などが、とられて来た。そこで、今まで述べてきた相関係数のグラフ表現が、変数選択の問題に、どのように応用できるかについて述べる。

重回帰分析において、変数の数が多ければ多いほど目的変数と説明変数の重相関係数は、1に近くなる。よって、形式的には初めから変数の数を限定せず、出来る限り選び出せばよいことになるが、変数の数が多くなると問題を複雑にして見通しのきかないものにしてしまう。そこで、2.2の注意でも述べたように、変数の数が多いときには、説明変数全体を共通点をもったいくつかのグループに分類し、各グループごとにグラフ表現をする。このことにより、説明変数と目的変数の関係を見やすいものにすることが出来る。そこで、各グループから、重要と思われる変数を求めればよい。その際、グラフ表現の見方と特徴の(4)で述べたように、変数間の多重共線性についても、充分注意を払わなければならない。

3.1 相関係数のグラフ表現による変数選択

変数選択については、多くの方法が開発されている。ここでは、

表 7. Haldによるデータ行列

	X_1	X_2	X_3	X_4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

表 8. 表7のHaldのデータをもとに、各変数の全組合せに対する R^{*2} 、AIC、Cp統計量

	変数の組合せ	R^{*2}	AIC	Cp
1	X_1, X_2, X_4	0.97645	13.37	7.19
2	X_1, X_2, X_3	0.97638	13.39	7.23
3	X_1, X_3, X_4	0.97504	13.70	7.92
4	X_1, X_2	0.97441	12.43	7.72
5	X_1, X_2, X_3, X_4	0.97356	15.36	9.16
6	X_1, X_4	0.96697	13.87	11.99
7	X_2, X_3, X_4	0.96376	15.80	13.76
8	X_3, X_4	0.92235	18.76	37.65
9	X_2, X_3	0.81643	23.56	98.58
10	X_4	0.64496	25.82	215.6
11	X_2	0.63592	25.96	221.3
12	X_2, X_4	0.61607	27.72	213.8
13	X_1	0.49160	27.86	312.5
14	X_1, X_3	0.45780	29.67	304.8
15	X_3	0.22090	30.26	407.6

- (i) コンピューターによる全変数の全組合せ
- (ii) AIC (赤池情報基準)
- (iii) Cp 統計量

の三つの方法と、グラフ表現による方法との比較をおこなう。

表7は、Hald (1952) のデータである。表8は、中村 (1983) によるもので、Hald (1952) のデータをもとにして、コンピューターによって全変数の全組合せに対して回帰式の重決定係数 R^{*2} を求め、その値を順にならべたものである。また、その変数の各組合せに対して AIC と、Cp 統計量をそれぞれ求めている。図9は、同じく、Hald のデータをもとにして、単相関係数、重相関係数、偏相関係数を求め、グラフ表現したものである。

図9. 表7のグラフ表現

図9のグラフ表現からは、次のような事を読み取ることが出来る。変数 X_1, X_2, X_3, X_4 の中で、 X_1 の矢印は、下向きで鋭くかつ横軸に最も近いので、 Y と重要な因果関係にあると思われる。次に変数 X_2 の矢印も下向きで鋭く、横軸に近いので、変数を二つ選択するのであれば、 (X_1, X_2) の組ということになる。この組合せは、AIC の方法において、 (X_1, X_2) の組が最小

値となっていることに一致している。また、変数 X_3 と X_4 を比較すると、それらの矢印の先端は、共に斜軸の中心近くにあるが、 X_4 の矢印の方が上向きに鋭いので、 X_3 よりも X_4 の方が、重要な変数と考えられる。よって、三変数を選択するのであれば、 (X_1, X_2, X_4) の組となる。この組合せについては、Cp 統計量の方法において、 (X_1, X_2, X_4) の組が最小値をとっていることに一致している。

謝 辞

この論文を書くにあたり、適切な御教示をいただいたレフェリーの方々に、深く感謝いたします。

参 考 文 献

- 馬場康雄, 脇本和昌(1983) ベクトル変換を用いた数量化法 統計数理研究所彙報第30巻第2号
- Chatterjee, S., Price, B. (1977) *Regression Analysis by Example* John Wiley & Sons.
- Hald, A. (1952) *Statistical Theory with Engineering Applications*, Wiley.
- 中村正一 (1983) 実践多変量解析法, 現代数学社.
- 奥野忠一, 久米 均, 芳賀敏郎, 吉澤 正(1971) 多変量解析法, 日科技連
- 坂元慶行, 石黒真木夫, 北川源四郎(1982) 情報量統計学, 共立出版
- 脇本和昌, 後藤昌司, 松原義弘(1979) 多変量グラフ解析法, 朝倉書店
- 昭和57年度社会生活統計指標 総理府統計局

—1983年12月受付, 1984年10月再受付—