

統計講座 (第9回)

多変量解析の基礎知識

高木 廣文

■ はじめに

実験, 調査, 検査などによって得られる個人のデータは, 1つの項目についてではなく, 年齢, 性, 職業, 血圧値, コレステロール値など多数あるのが普通である. このようなデータは「多変量データ multivariate data」と呼ばれることがある.

これまで解説してきた方法の多くは, 1つの変数についてのものであり, 相関とクロス表に関するものが2変数についての方法であった. より多くの変数(3変数以上)に関する多変量データを, 変数相互の関係を考慮して, 目的に応じて分析するための手法群は「多変量解析 multivariate analysis」と呼ばれている.

多変量解析の目的は, 以下のように大別して考えることができる.

- 1) 事象の簡潔な記述, 情報の圧縮
 - 2) 事象の背後にある潜在因子の探索
 - 3) 特定事象に対する他の要因の影響を検討
 - 4) 個体の判別や識別
- の4つである.

多変量解析とは, 上記の目的に応じた多くの方法の総称である. そのうち, 主要な方法について, 以下にその概略をまず説明しよう.

■ 情報圧縮の考え方と基準について

ある大学の入学試験は英語, 数学, 国語の3科目であるとしよう. この場合, 入学試験の結果を

どのように評価するのだろうか. よく用いる方法は,

$$\text{合計点} = \text{英語} + \text{数学} + \text{国語}$$

のように, 「合計点」として, 3科目の得点を1つにまとめたものを使用するという考え方があるだろう.

このように, 複数の科目の得点から1つの合計点にまとめるという考え方は, 一種の情報の圧縮を行っていることになる.

ところで, 合計点の計算をするのに, 各科目の得点を単純に足しただけでよいのだろうか. 例えば,

$$\text{合計点} = 2 \times \text{英語} + \text{数学} + 0.5 \times \text{国語}$$

ではいけないのだろうか. また,

$$\text{合計点} = \text{英語} + 2 \times \text{数学} + 0.2 \times \text{国語}$$

ではどうなのか.

単純に足し算すればよいのか, または各科目にかける数値(重み)を変えた方がよいのだろうか. これは, どのような学生を大学が入学させたいかによるだろう. 理科的能力の高い人を選びたいのか, 文科的能力の高い人を選びたいのか, 入学後の成績のよい人を選びたいのか等, どのような目的で選ぶのかにより, 各学科の重みも変わってくるだろう.

例えば, 入学後の成績の平均点と入学試験の関係が高くなるように重みを決めたいとしよう. 一般的に書くと,

$$\text{合成得点} = \text{重み}1 \times \text{英語} + \text{重み}2 \times \text{数学} + \text{重み}3 \times \text{国語}$$

のような計算式を考えればよい.

実際に各科目にかける重みを決める方法は、

- 1) 合成得点と入学後の教科の平均点の差を最小
 - 2) 合成得点と平均点の相関最大
- のような基準をもとにして、重みを求めればよいだろう。このような基準のもとに解析を行う方法は「重回帰分析multiple regression analysis」と呼ばれている。

それでは、評価の基準が大学での適応／不適応のような場合、どのように重みを決めればよいだろうか。

考え方として、適応群では重み付けした得点がプラスになり、不適応群では得点がマイナスになるように決めることができれば、その合成得点を使用することにより、簡単に適応者と不適応者をかき分けることができるようになるだろう。

このような考え方は、「2つの群での平均得点の距離を最大にする」ということであり、「判別分析discriminant analysis」と呼ばれる方法の基本的な考え方である。

このように、入学後の成績や大学への適応・不適応のように、式の左側にきて、重み付けの基準となるものは、「基準変数criteria variable」、「従属変数dependent variable」、「内生変数endogenous variable」、または「外的基準external criteria」などと呼ばれている。呼び方が数種類あるのは、多変量解析の各手法が多くの分野で発展してきたためである。各分野でその呼び方に違いがあるので注意が必要である。

同様に、英語、数学、国語の試験得点のように、式の右側にきて、重み付けした合計点を求めるために使われるものは、「説明変数exploratory variable」、「独立変数independent variable」、「外生変数exogenous variable」、「内的基準internal criteria」などと呼ばれている。

ところで、上記のような基準変数がない場合もある。例えば、英語、数学、国語、物理、化学、世界史などの試験成績で何がわかるのかを考えてみよう。このような学力試験は、一般にはある能力を測るために行われるのであるが、これは暗黙に学力に関する因子が存在することを仮定しているともいえよう。このように学力に関する潜在的な能力因子の探索により、試験結果によって「理

科的能力」や「文科的能力」を測定することが可能になるかもしれない。そのような目的のための方法として「因子分析factor analysis」や「主成分分析principal component analysis」と呼ばれる分析方法がある。

■ 多変量解析の理解のための基礎知識

多変量解析の具体的な各方法を理解するためには、ある程度の統計学的な知識が必要である。それらの正確な知識がなくても、実際のデータへの応用にとくに困るということはない。しかし、分析結果の正しい解釈にはかなり正確な知識がある方がよいし、そもそもどの方法を用いれば、自分の考えている結果を得ることができるのかを知るには、正しい知識は不可欠である。

統計学で基本的な統計量である平均値、分散は当然の知識であるが、多変量解析では2変数間の関係を表すために共分散と相関係数（ピアソンの積率相関係数）は必須の知識である。これらの統計量に関してはすでに説明を行っている。この他に、知っている多変量解析の理解が飛躍的に向上するものとして、ベクトルと行列に関する知識があげられる。とくに、行列の固有値と固有ベクトルに関する知識は、主成分分析と因子分析などの方法の正確な理解に欠かせないものである。また、行列の逆行列に関する知識は、重回帰分析や判別分析などの計算に用いられている。

統計学は数学とは異なるので、これらのベクトルや行列に関する正確な数理的な知識は必要ではないが、以下に簡単にそれらに関して解説しておこう。

1. 平均値と分散

絶対に必要な知識として、基本的な統計量である平均値mean、分散varianceがある。すでに説明している内容であるが、重要であるので再度以下に定義を述べよう。

標本数をN、データを x_1, x_2, \dots, x_N とすると平均値 x_M は、

$$\begin{aligned} x_M &= (x_1 + x_2 + \dots + x_N) / N \\ &= \frac{1}{N} \sum_i x_i \end{aligned}$$

により求められる。なお、 Σ_i はすべてのデータの合計を求めることを表す。

変数 X の分散 V_x は、分布のばらつきの指標（散布度）であり、各データと平均値 x_M の差（偏差）の 2 乗の平均値として定義される。すなわち、

$$V_x = [(x_1 - x_M)^2 + (x_2 - x_M)^2 + \dots + (x_N - x_M)^2] / N$$

$$= \frac{1}{N} \Sigma_i (x_i - x_M)^2$$

である。上式の N の代わりに $N - 1$ を用いると、母集団での分散の不偏推定量（期待値が母集団値と一致）である「不偏分散」となるが、ここでは後述する統計量でもすべて分母は N で統一する。

分散は計算式からも分かるように、データを 2 乗しているのだから、その単位も身長なら cm^2 、体重なら Kg^2 となる。このため、分布の散布度を直観的に把握するのが難しいため、その平方根を求め、単位をもとに戻したものを多く用いることが多い。これは標準偏差 standard deviation, SD と呼ばれている。

2. 共分散と相関係数

多変量解析では、基本的に分析に使用する各 2 変数間の関係をもとに解析を行う。したがって、2 変数間の関係の指標についての理解は極めて重要である。通常は、共分散 covariance と（ピアソンの積率）相関係数 correlation coefficient が用いられる。これらの統計量もすでに説明してあるが、重要なので以下に再度その定義を述べよう。

共分散 C_{xy} は 2 変数 X, Y の偏差の積の平均値として定義される。すなわち、

$$C_{xy} = [(x_1 - x_M)(y_1 - y_M) + (x_2 - x_M)(y_2 - y_M) + \dots + (x_N - x_M)(y_N - y_M)] / N$$

$$= \frac{1}{N} \Sigma_i (x_i - x_M)(y_i - y_M)$$

により求められる。ここで、 x_M, y_M は変数 X, Y の平均値である。

共分散の単位は、2 変数の単位の積となるため、変数間の関係の強さをその数値の大きさのみから測ることはできない。そのため、共分散を 2 変数の標準偏差で割り、単位をなくし、かつその取り得る数値の範囲を $-1 \sim 1$ にしたものが、相関係

数である。すなわち、相関係数 r_{xy} は、

$$r_{xy} = \frac{X \text{ と } Y \text{ の共分散}}{[X \text{ の標準偏差}] \times [Y \text{ の標準偏差}]}$$

$$= \frac{\Sigma_i (x_i - x_M)(y_i - y_M)}{\sqrt{[\Sigma_i (x_i - x_M)^2] [\Sigma_i (y_i - y_M)^2]}}$$

により求められる。

相関係数の 2 乗は決定係数 coefficient of determination と呼ばれており、一方の変数で他方の変数の変動をどれだけ説明できるかを示すための統計量として用いられている。

3. ベクトルの基礎知識

多変量解析では、多数の変数を扱うため、ベクトル vector と行列 matrix による数式表現の方法を用いる方が簡単であり、分析方法の理解も容易になる。

いま、3 人の身長が 170, 165, 180 (cm) であるとしよう。このような場合、身長の変数 X を、 $x = (170, 165, 180)$ のようにデータを人数分ならべ、括弧で閉じたものは行ベクトル、または横ベクトルと呼ばれる。ベクトルは、縦に並べることも多く、

$$y = \begin{pmatrix} 170 \\ 165 \\ 180 \end{pmatrix}$$

のように縦に並べたものは、列ベクトルまたは縦ベクトルと呼ばれている。

縦ベクトルを横ベクトルにすること（またはその逆も）を「転置」と呼び、変数名にダッシュをつけて表現する。すなわち、 $x' = y$ 、および、 $y' = x$ 、の関係がある。

多変量解析でよく用いるベクトルは、「変数ベクトル」と呼ばれるもので、身長ベクトル、体重ベクトルなどのように、ベクトルを構成するの要素が、各ケースのデータからなっているものである。

特別なベクトルとして、すべての要素がある変数の平均値からなる「平均（値）ベクトル」がある。平均ベクトルは、各データから平均値を引く目的で用いられることが多い。例えば、変数 X の変数ベクトルを $x = (x_1, x_2, \dots, x_N)$ とし、平

均値を x_M とした場合、その平均ベクトルは $x_M = (x_{M1}, x_{M2}, \dots, x_{MN})$ のようにかける。この場合、2つのベクトルの差のベクトルは「偏差ベクトル」と呼ばれており、

$$x_d = x - x_M = (x_1 - x_{M1}, x_2 - x_{M2}, \dots, x_N - x_{MN})$$

となる。偏差ベクトルは、各要素の合計が0になり、以後の計算が便利になるのでよく用いられている。

2つのベクトル同士の演算は、ベクトルの大きさ(含まれる要素の個数)が等しい場合にのみ定義される。

ベクトル同士の足し算と引き算は、それぞれのベクトルの対応する要素同士をそれぞれ足し算引き算すればよい。例えば、 $x = (x_1, x_2, \dots, x_N)$ 、 $y = (y_1, y_2, \dots, y_N)$ の場合、

$$x \pm y = (x_1 \pm y_1, x_2 \pm y_2, \dots, x_N \pm y_N)$$

となる。

2つのベクトルの掛け算は、掛け算をする左側にくるベクトルは、横ベクトルであり、右側にくるベクトルは縦ベクトルでなければならない。例えば $x = (x_1, x_2, \dots, x_N)$ 、 $y = (y_1, y_2, \dots, y_N)$ の場合、 xy では計算が行えず、形式的には、上述したベクトルの転置を行う必要がある。すなわち x と y の掛け算は、

$$\begin{aligned} xy' &= (x_1, x_2, \dots, x_N)(y_1, y_2, \dots, y_N)' \\ &= x_1y_1 + x_2y_2 + \dots + x_Ny_N \end{aligned}$$

と2つのベクトルの対応した要素同士の積の和となる。したがって、同一のベクトルの積は、

$$\begin{aligned} xx' &= x_1x_1 + x_2x_2 + \dots + x_Nx_N \\ &= x_1^2 + x_2^2 + \dots + x_N^2 \end{aligned}$$

と各要素の2乗和となる。これは、ベクトルの「ノルムnorm」と呼ばれており、 $xx' = \|x\|^2$ のように表すことも多い。

もしも、ベクトル x が偏差ベクトルであれば、分散 V_x は、

$$V_x = \|x\|^2 / N$$

のように、極めて簡潔に表現できる。

同様に2変数 X と Y の共分散 C_{xy} は、変数ベクトル x 、 y が偏差ベクトルであれば、

$$C_{xy} = xy' / N$$

と表せるので、相関係数 r_{xy} は、共分散 C_{xy} を2変数の標準偏差 $\|x\| / \sqrt{N}$ と $\|y\| / \sqrt{N}$ で割ることで求められるので、

$$r_{xy} = \frac{xy'}{\|x\| \cdot \|y\|}$$

と表すことができる。Σ記号を用いた表現に比べて、極めて簡潔でわかりやすい表現になっている。

4. 行列についての基礎知識

多変量解析でよく用いる行列は、データ行列、分散共分散行列、相関係数行列などである。実際には、計算に便利のように、データ行列の各成分(要素)である個々のデータは、各変数平均値を引いてある偏差データとして、最初から定義しておくことが多い。

いま、変数が4つで、 X_1 が身長、 X_2 が体重、 X_3 が最高血圧、そして X_4 が最低血圧である5人のデータ行列(粗データ行列) X_R を考えよう。そのような行列は、以下のような形式をしている。

$$X_R = \begin{pmatrix} X_1 & X_2 & X_3 & X_4 \\ 170 & 65 & 120 & 80 \\ 165 & 57 & 122 & 78 \\ 180 & 72 & 115 & 75 \\ 168 & 55 & 124 & 60 \\ 162 & 50 & 130 & 82 \end{pmatrix} \begin{array}{l} \text{1 番目のケース} \\ \text{2 番目のケース} \\ \text{3 番目のケース} \\ \text{4 番目のケース} \\ \text{5 番目のケース} \end{array}$$

この粗データ行列 X_R の各データから、それぞれの変数の平均値を引き算することで、偏差データ行列 X を求めることができる。各変数の平均値は身長169、体重59.8、最高血圧122.2、最低血圧75であるから、これらの値を引くと、

$$X = \begin{pmatrix} \text{身長} & \text{体重} & \text{最高血圧} & \text{最低血圧} \\ 170-169 & 65-59.8 & 120-122.2 & 80-75 \\ 165-169 & 57-59.8 & 122-122.2 & 78-75 \\ 180-169 & 72-59.8 & 115-122.2 & 75-75 \\ 168-169 & 55-59.8 & 124-122.2 & 60-75 \\ 162-169 & 50-59.8 & 130-122.2 & 82-75 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 5.2 & -2.2 & 5 \\ -4 & -2.8 & -0.2 & 3 \\ 11 & 12.2 & -7.2 & 0 \\ -1 & -4.8 & 1.8 & -15 \\ -7 & -9.8 & 7.8 & 7 \end{pmatrix}$$

となる。

分散共分散行列 C_{xx} は上記の偏差データ行列 X の積 $X'X$ を標本数 5 で割ることで求められる。すなわち、

$$C_{xx} = X'X / 5$$

$$= \begin{pmatrix} 1 & -4 & 11 & -1 & -7 \\ 5.2 & -2.8 & 12.2 & -4.8 & -9.8 \\ -2.2 & -0.2 & -7.2 & 1.8 & 7.8 \\ 5 & 3 & 0 & -15 & -7 \end{pmatrix} \begin{pmatrix} 1 & 5.2 & -2.2 & 5 \\ -4 & -2.8 & -0.2 & 3 \\ 11 & 12.2 & -7.2 & 0 \\ -1 & -4.8 & 1.8 & -15 \\ -7 & -9.8 & 7.8 & 7 \end{pmatrix} \times \frac{1}{5}$$

$$= \begin{pmatrix} X_1 & X_2 & X_3 & X_4 \\ 37.6 & 44.8 & -27.4 & -8.2 \\ 44.8 & 60.6 & -36.8 & 4.2 \\ -27.4 & -36.8 & 24.2 & 3.2 \\ -8.2 & 4.2 & 3.2 & 61.6 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix}$$

と求められる (計算結果は小数点以下第 2 位四捨五入している)。

上記の分散共分散行列の対角成分 (それぞれの変数の組み合わせが同一の場合) は、各変数の分散を示している。例えば、変数 X_2 (体重) の分散は 60.6 であることを示している。また、縦と横の変数同士の組み合わせは、それらの変数同士の共分散を示している。例えば、変数 X_2 (体重) と変数 X_3 (最高血圧) の共分散は -36.8 であることが分かる。なお、分散共分散行列をみて分かるように、対角成分を挟んで右上の三角部分の成分と左下の三角部分の成分は対称になっていることが分かるだろう。このような行列は「対称行列」と呼ばれている。

分散共分散行列 C_{xx} の各成分を、対応する 2 変数の標準偏差で割ることで相関係数が求められるので、相関係数行列 R_{xx} は簡単に求めることができ、以下ようになる。

$$R_{xx} = \begin{pmatrix} 1.000 & 0.939 & -0.909 & -0.170 \\ 1.000 & 1.000 & -0.961 & 0.068 \\ -0.909 & -0.961 & 1.000 & 0.082 \\ -0.170 & 0.068 & 0.082 & 1.000 \end{pmatrix}$$

である (小数点以下第 2 位四捨五入)。

多変量解析の手法の多くで共通に使用するベクトルや行列の基礎的な知識は以上である。しかし、分析結果を得るための過程を理解するには、さらに行列の逆行列と固有値・固有ベクトルの知識が不可欠である。

5. 行列の逆行列とは

まず、簡単な例として 5 の逆数について考えてみると、

$$5^{-1} = 1 / 5$$

が 5 の逆数である。5 と 5 の逆数である 5^{-1} の積は、

$$5 \times 5^{-1} = 5^{-1} \times 5 = 1$$

と 1 になる。もともと、逆数の定義が積が 1 になるような数を、その数の逆数と定義しているのだから、これは当然である。

この考えを行列にも拡張して、行列 A の逆行列 A^{-1} はその積が、

$$AA^{-1} = A^{-1}A = \begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ & & \cdot & \\ 0 & & & 1 \end{pmatrix} = I$$

のように対角成分のみ 1 で、他は 0 となるような行列 (単位行列 I) となるような行列を、行列 A の逆行列 A^{-1} と呼ぶことにする。

なお、実際にある行列の逆行列を求めるためには、かなり大変な計算が必要であり、コンピュータを使って計算するのが普通である。多変量解析では、計算は何事もコンピュータに依存し、自分で電卓をたたくなどということはあり得ない。そのために、具体的なイメージが湧きにくい、あまり頑なに 1 変量解析のときの学習法に拘泥すると、多変量解析の理解はおぼつかない、発想

の切り替えが必要である。

上述した相関係数行列 R_{xx} の逆行列 R_{xx}^{-1} を求めると (著者のホームページで計算できる),

$$R_{xx}^{-1} = \begin{pmatrix} 20.408 & -29.638 & -10.451 & 6.341 \\ -29.638 & 61.297 & 32.942 & -11.908 \\ -10.451 & 32.942 & 23.645 & -5.955 \\ 6.341 & -11.908 & -5.955 & 3.376 \end{pmatrix}$$

となる。実際に, R_{xx} と R_{xx}^{-1} の積の対角成分である 1 行 1 列 (1, 1) について計算してみると,

$$1 \times 20.408 + 0.939 \times (-29.638) - 0.909 \\ \times (-10.451) - 0.170 \times 6.341 = 1$$

となる (実際には, 計算誤差があるが)。また, 同様に 2 つの行列の積の 1 行 2 列の計算を行うと,

$$1 \times (-29.638) + 0.939 \times 61.297 - 0.909 \\ \times 32.942 - 0.170 \times (-11.908) = 0$$

となることがわかる。読者自身で, 同様な計算で確認するとよいだろう。なお, ここでは専門的になるので詳細は触れないが, 上記のようにどの行列にも逆行列が存在するわけではない点に注意がいる。この点については, 重回帰分析などの説明で再度触れたい。

6. 行列の固有値と固有ベクトルについて

主成分分析や因子分析では, 相関係数行列の固有値, 固有ベクトルを解析の過程で求める必要がある。ここで, 行列の固有値と固有ベクトルについて, 簡単に説明する。

ある行列 A が大きさ $p \times p$ (縦と横が p 個) の固有値 λ と固有ベクトル a とは,

$$A a = \lambda a$$

の方程式 (「正規方程式」と呼ばれる) を満たすようなベクトル a と適当な数値 (スカラー) λ を, それぞれ固有ベクトル eigen vector, 固有値 eigen value と呼んでいる。

これを, 行列 A とベクトル a について, それぞれ要素 (成分) を用いて書き出してみると,

$$\begin{pmatrix} A_{11} & A_{12} & A_{1p} \\ A_{21} & A_{22} & A_{2p} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ A_{p1} & A_{p2} & A_{pp} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_p \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_p \end{pmatrix}$$

のように書ける。さらに, これを個別の方程式に書き下すと,

$$\begin{aligned} A_{11}a_1 + A_{12}a_2 + \cdots + A_{1p}a_p &= \lambda a_1 \\ A_{21}a_1 + A_{22}a_2 + \cdots + A_{2p}a_p &= \lambda a_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ A_{p1}a_1 + A_{p2}a_2 + \cdots + A_{pp}a_p &= \lambda a_p \end{aligned}$$

のように, p 個の方程式となる。逆にいえば, $A a = \lambda a$, という簡単な式が, 実は p 個の方程式を表していたことになる。

結局, $p \times p$ の行列 A の固有値を, 固有ベクトルを求めることは, 上記の p 元連立方程式の解を求めることに帰着する。しかしながら, 求めるべき未知数の個数は, 固有値 λ とベクトル a の p 個の要素の $p + 1$ 個あるので, 上記の方程式からだけでは, 一意な解を求めることができない。このため, 通常は,

$$a_1^2 + a_2^2 + \cdots + a_p^2 = 1 \quad (\text{または } \lambda)$$

のような制約をもうけ, 未知数の個数と方程式の個数を一致させて, 解を求めるのが一般的である。なお, 上記の連立方程式を満たす解は 1 通りとは限らず, 最大で行列の大きさと同じ p 通り存在することが知られている。それらの解は, 固有値の大きさの順に, 第 1 固有値 / 第 1 固有ベクトル, 第 2 固有値 / 第 2 固有ベクトル, ..., のように呼ばれる。

当然であるが, 固有値と固有ベクトルの計算にはコンピュータは不可欠である。上述の例の相関係数行列の固有値と固有ベクトルを求めると次の表のようになる。

表 相関係数行列の固有ベクトルと固有値

ベクトルの要素	固有ベクトル			
	第1	第2	第3	第4
a_1	0.574	0.082	0.717	-0.387
a_2	0.579	-0.162	-0.014	0.799
a_3	0.576	0.015	0.693	0.433
a_4	0.056	-0.983	0.072	-0.157
固有値	2.879	1.024	0.087	0.010

上記のように固有値・固有ベクトルは4つ存在することが分かる。これらの数値の意味や解釈については、今後それぞれの方法の解説の場で述べていきたい。

■ おわりに

今回は、多変量解析の概略を紹介するとともに、今後の各手法の説明の理解のために、ベク

トルと行列について解説した。内容的にはそれほど難しいものではないが、これらの用語や表記に不慣れな読者は直観的な把握が困難かもしれない。今後の解説で必要に応じて今回の説明が必要になると思われるので、この号は無くしたりしないように注意がいる。

なお、行列やベクトルに関する計算は、平均値を電卓で求めたりすることさえまれなのであるから、絶対にあり得ないし、やっても大抵はどこかで計算間違いするのが普通である。

多変量解析では、データから相関係数行列、逆行列、固有値、固有ベクトルなどを求める計算は、すべてパソコンを用いて行う。

我々が行うことは、研究のためにデータを集めること、正しい分析方法の選択、そして得られた結果の正しい解釈ということになる。このために、次号以降、順に多変量解析の各手法を解説していくことにする。

