

相関係数、回帰直線

1 散布図

1 個体が複数の変数で表される（データがベクトルで与えられる）場合で、2 変数のデータ（平面上のデータ）であれば、

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

平面状にプロットした図を散布図という。また 3 変数のデータ（空間上のデータ）であれば、つぎのような形がデータである。

$$(x_1, y_1, z_1), (x_2, y_2, z_2) \dots, (x_n, y_n, z_n)$$

このようなデータを整理するには、いろいろな方法があるが、2 変数データについて、つぎのような概念が重要である。

[共分散]

2 変数の間の全体的なばらつき具合を考える。一つの変数の増加について、他方も増加であれば、正の値、逆の場合には、負の値をとる。増減がはっきりしないときには、ゼロに近い値となる。つぎの相関係数と対応する。

$$s_{XY} = \frac{1}{n} \sum_n \{(X_i - \bar{X})(Y_i - \bar{Y})\}$$

2 相関係数

相関係数

共分散を各変数の標準偏差で基準化したもの。

$$\rho = \frac{s_{XY}}{s_X s_Y} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{1}{n} \sum_n \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

$-1 \leq r \leq 1$ が成り立ち、絶対値が 1 に近いときには直線傾向が強く、絶対値がゼロに近いときには、直線傾向がみられないことを意味する。

3 曲線の当てはめ

曲線の当てはめ、最小 2 乗法

散布図からある程度の直線的な関係があるとき、これらの 2 変数データに直線をつぎの基準で当てはめる。

$$\sum_i (Y_i - a - bX_i)^2$$

を最小にするような a, b を求める。

[回帰直線] この方法を最小 2 乗法といい、求めた直線を Y の X への回帰直線という。最小 2 乗法で求めた直線の係数を回帰係数という。

$$b = \frac{s_{XY}}{s_X^2} = r \frac{s_Y}{s_X}, \quad a = \bar{Y} - r \frac{s_Y}{s_X} \bar{X}$$

問 3.1

相関係数は線形変換をしても同じ値になることを示しなさい。つまり $(X_1, Y_1), \dots, (X_n, Y_n)$ と $(aX_1 + b, cY_1 + d), \dots, (aX_n + b, cY_n + d)$ (ただし $ac > 0$) とは同じ相関係数をもつ。

問 3.2

与えられた 2 変量データ X, Y が順位データであるとき、すなわちそれぞれが $1, 2, 3, \dots, n$ までを順位づけられていて、並び替えたものである。ただし同順位はないとする。このときには相関係数が

$$r = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}, \quad (d_i = X_i - Y_i)$$

で与えられることを示しなさい。これをスピアマンの順位相関係数という。

4 多変量データの解析

4.1 回帰分析

いま 2 つの変数 (x, y) の間に一次の関数関係 $y = \alpha + \beta x$ があるものとする。 x の値を与えると、係数 α, β が既知ならば、 y の値を定めることができる。具体的な例として、測定誤差など \dots しかし、いまこの構造式の中に確率変動の要素が加わると考えよう。観測値 y のなかには確率変動が含まれるとする。いま n 個のデータ

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

について

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \dots$$

なる関係があり、この ϵ_i は、各 i について独立で同一の分布 $N(0, \sigma^2)$ にしたがうとする。すなわち各 x_i が指定されると、上の式で定まる y_i は ϵ_i によって確率変動し、平均 μ_i は、 $\alpha + \beta x_i$ 、分散 $\sigma_i^2 = \sigma^2$ (一定) の正規分布にしたがう。

このように 2 つの変数間が一次式で結ばれて変動するような関係を線形回帰モデル (Linear Regression Model) とよぶ。変数 x のことを説明変数 (explanatory variable) または独立変数 (independent variable) とよび、これに対する変数 y のことを従属変数 (dependent variable) という。一般に、説明変数は非確率的なものとする。係数 α, β を回帰係数 (regression coefficient) とよび、これは未知の母数である。 x を与えて観測された y の値をデータとして、この回帰係数を推定することが問題となる。

回帰係数の推定には、ガウスによる最小 2 乗の原理 (principle of minimum square) に基づく。いま n 個のデータ :

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

が与えられたとする。未知パラメータ α, β を

$$S(\alpha, \beta) = \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2$$

が最小となるように決定しようとする。このような方法を最小 2 乗法 (least square method) とよぶ。データは固定されていて、 α と β がどんな値であっても

$$S(\hat{\alpha}, \hat{\beta}) \leq S(\alpha, \beta)$$

を満たす最小値 $\hat{\alpha}, \hat{\beta}$ を求めようとするわけで、このようにして定まる係数の値を最小 2 乗推定値 とよぶ。 $y = \alpha + \beta x$ は直線の方程式であるから、係数を変化させることは、いろいろ直線を動かすことで、最小 2 乗法とは各点での残差 $y_i - \alpha - \beta x_i$ の 2 乗和が小さくなるよう直線をあてはめようとするのである。

実際、 $\hat{\alpha}, \hat{\beta}$ を決定するには、微分積分学の極値問題を使って、まず必要条件として $S(\alpha, \beta)$ に対する 2 変数の微分係数をゼロとした方程式を解き、最小値の候補を求める。すなわち、

$$\begin{aligned}\frac{\partial S}{\partial \alpha} &= -2 \sum_{i=1}^n (y_i - \alpha - \beta x_i) = 0 \\ \frac{\partial S}{\partial \beta} &= -2 \sum_{i=1}^n x_i (y_i - \alpha - \beta x_i) = 0\end{aligned}$$

これを整理すれば

$$\begin{cases} n\alpha + (\sum x_i)\beta = \sum y_i & \dots\dots(1) \\ (\sum x_i)\alpha + (\sum x_i^2)\beta = \sum x_i y_i & \dots\dots(2) \end{cases}$$

を得る。これを **正規方程式** (normal equation) とよぶ。 α と β に関する 2 元連立線形方程式である。(1) より、 $\alpha = \bar{y} - \beta \bar{x}$ (ただし \bar{x} は、 x_1, \dots, x_n の平均値、 \bar{y} は y_1, \dots, y_n の平均値) となるから、(2) に代入して、 β を得る。さらに α も得られる。こうして得た α, β の値は

$$\begin{aligned}\hat{\alpha} &= \bar{y} - \hat{\beta} \bar{x} = \bar{y} - \frac{\sigma_{xy} \bar{x}}{\sigma_{xx}} \\ \hat{\beta} &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma_{xy}}{\sigma_{xx}}\end{aligned}$$

ここで σ_{xx} と σ_{xy} はそれぞれ x, y についての分散、共分散を表すとする。さらに、いま求めた値に対して極値問題の十分条件 $S''_{\alpha\alpha} - S''_{\alpha\beta} S''_{\beta\beta} < 0$ を調べ、これを満たしているからことがわかる。以上により求める直線の方程式は

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_{xx}} (x - \bar{x})$$

与えられ、この回帰直線は点 (\mathbf{X}, \bar{y}) を通り、傾きが $\frac{\sigma_{xy}}{\sigma_{xx}}$ である。

つぎに最小 2 乗法の幾何的意味を考えよう。与えられた構造を表すモデルの式は

$$y_i = \alpha + \beta x_i + e_i, \quad i = 1, \dots, n$$

と書けていたから、まずこれをベクトルや行列を用いて表す。

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

つまり、行列とベクトルを使って

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

が線形回帰モデルの表現である。ただし

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \beta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

行列の転置 (transpose) を表すため、左上に t をつけるとしよう。

$${}^t\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix}$$

これを用いると、正規方程式は

$$[{}^t\mathbf{X}\mathbf{X}]\beta = {}^t\mathbf{X}\mathbf{y}$$

と書ける。2次の正方行列

$${}^t\mathbf{X}\mathbf{X} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

を考えると、その行列式 (determinant) は $n \sum_i x_i^2 - (\sum x_i)^2$ で、これは n^2 で割ると分散 σ_{xx} であるから、0でないとしてよい。したがって行列は正則である。逆行列が存在するから $\beta = [{}^t\mathbf{X}\mathbf{X}]^{-1} {}^t\mathbf{X}\mathbf{y}$ 回帰直線はしたがって $\mathbf{y} = \mathbf{X}[{}^t\mathbf{X}\mathbf{X}]^{-1} {}^t\mathbf{X}\mathbf{y} + \mathbf{e}$ となる。この式において右辺の \mathbf{y} の係数行列を P とおくと、

$$P = \mathbf{X} [{}^t\mathbf{X}\mathbf{X}]^{-1} {}^t\mathbf{X}$$

この P はつぎを満たしている。

- 1) $P^2 = P$ (ベキ等という)
- 2) ${}^tP = P$ (対称という)

一般にこのような行列を係数として定まる線形写像を射影 (projection) という。 $\mathbf{y} = [y_1 y_2 \cdots y_n]$ は n 次元ベクトル空間の点で、 $P\mathbf{y}$ は2次元ベクトル空間 (平面) の点で、 \mathbf{y} を射影したベクトルとなる。

図

$\mathbf{y} = P\mathbf{y} + \mathbf{e}$ より $\mathbf{e} = (\mathbf{I} - P)\mathbf{y}$ は単位行列。

残差ベクトル \mathbf{e} は n 次元ベクトルの点 \mathbf{y} より2次元ベクトル空間へおろして垂線を表す。もし

$$\beta = [{}^t\mathbf{X}\mathbf{X}]^{-1} {}^t\mathbf{X}\mathbf{y}$$

でない一般の β を用いると、 $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ であり、このときの残差ベクトルは $\mathbf{e} = \mathbf{y} - \mathbf{X}\beta$ は $\mathbf{e} = (\mathbf{I} - P)\mathbf{y}$ と比べれば、ベクトルの大きさは長くなっている。これが最小2乗法の幾何的意味である。

さて α, β のそれぞれの推定値 $\hat{\alpha}, \hat{\beta}$ を $S(\alpha, \beta)$ に代入した最小値は $\sum_i (y_i - \hat{y}_i)^2$ となり、これを残差平方和 (sum of square of residue) といい、 $n - 2$ で割ったものを残差分散 (variance of residue) とよぶ。

$$S_e = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

この式はつぎのように変形できる。

$$\begin{aligned} \sum_i (y_i - \hat{y}_i)^2 &= \sum (y_i - \bar{y})^2 - 2\hat{\beta} \sum (y_i - \bar{y})(x_i - \bar{x}) + \hat{\beta}^2 \sum (x_i - \bar{x})^2 \\ &= \sum (y_i - \bar{y})^2 - \hat{\beta}^2 \sum (x_i - \bar{x})^2 \\ &= \sum (y_i - \bar{y})^2 - \sum (\hat{y}_i - \bar{y})^2 \end{aligned}$$

ゆえに

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \bar{y})^2$$

が成り立つ。この式は y の全変動 $\sum_i (y_i - \bar{y})^2$ が回帰による変動 $\sum_i (\hat{y}_i - \bar{y})^2$ と残差変動 $\sum_i (y_i - \bar{y})^2$ に分解されることを示している。変量 y の全変動のうち、回帰による変動の占める割合、

$$\rho = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

を回帰の決定係数 (coefficient of determination) という。この ρ の値が1に近ければ、 y の変動の大部分が x に対する回帰によって説明されることを意味し、もし、0に近ければ y の変動の説明に x はほとんど無力であることになる。

また $\frac{1}{n} \sum_i (\hat{y}_i - \bar{y})^2 = \frac{2\hat{\beta}}{n} \sum_i (y_i - \bar{y})(x_i - \bar{x}) - \frac{\hat{\beta}^2}{n} \sum_i (x_i - \bar{x})^2 = \frac{\sigma_{xy}^2}{\sigma_{xx}}$ より

$$\rho = \sigma_{xy}^2 / \sigma_{yy} = \frac{\sigma_{xy}^2}{\sigma_{xx}\sigma_{yy}}$$

したがって $\sigma_x^2 = \sigma_{xx}$, $\sigma_y^2 = \sigma_{yy}$ であるから

$$\sqrt{\rho} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

すなわち回帰の決定係数 ρ の平方根は x と y との相関係数となる。

4.2 主成分分析

主成分分析とは多変量データがもつ情報をなるべく損なうことなく、縮約するための統計的手法である。たとえば、学生の数学試験を前期と後期の2回おこなった成績データが与えられたとする。この相関係数はたかく、一方の成績が良ければ他方もよいという傾向が認められるであろう。単に算術平均をするのではなく、2つの成績の数値を1つの代表値にまとめることを考える。また英語の成績と数学のそれとをまとめる場合も同じ様に考えられる。いま n 個の2次元データ

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

が与えられたとする。これらに

$$z_i = ax_i + by_i, \quad i = 1, 2, \dots, n$$

という線形変換を施して、2次元データを1次元に縮約する。2次元データの情報がなるべく損なわれないう縮約するには、どのような変換を行えばよいだろうか？まずこの線形変換が幾何的にどういう関係を表しているかを考えよう。原点と点 (a, b) を結んでできる直線を延長して新しい軸を設ける。与えられた2次元の点 (x, y) からこの軸に垂直におろした垂線の足を z とおけば、線形変換は2次元データの点 (x_i, y_i) を対応する新しい軸上の点 z_i に対応、詳しくいうと、平面の点をこの軸のうえに射影させている。

図

この線形変換によって2次元のデータとして散らばり、広がっていたものが、1次元の新しい軸上の点として並び換えられたことになる。したがって問題は、散布していたデータを、なるべくその特徴をよく代表するよう、適切な軸を選ぶことになる。データの特徴を代表するという考えかたを、分散の最大化、すなわち個体の識別効果を大きくすることと意味づけてみる。また係数 a, b の役割は直線の方角を決めている。

点 (a, b) を単位円周上の点、いいかえると、ベクトル $\begin{bmatrix} a \\ b \end{bmatrix}$ の長さが1、 $a^2 + b^2 = 1$ を満たすとしてよい。

この立場に立てば、問題は [条件: $a^2 + b^2 = 1$ の下で、 $z = ax + by$ の分散 $\sigma_z^2 = a^2\sigma_x^2 + 2ab\sigma_{xy} + b^2\sigma_y^2$ を最大にせよ。] という形に帰着された。さてこの条件つき最大化問題を解かなければならない。線形代数における対称行列の対角化を用いよう。ベクトルと行列をもちいて、目的関数を表現すると、分散と共分散を要素とする対称行列 (これを分散共分散行列という) $\sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$ とおいて

$$\sigma_z^2 = a^2\sigma_x^2 + 2ab\sigma_{xy} + b^2\sigma_y^2 = (a, b)\sigma \begin{bmatrix} a \\ b \end{bmatrix}$$

これはベクトル $\begin{bmatrix} a \\ b \end{bmatrix}$ に関する2次形式と呼ばれる。係数行列は対称であるから、適当な直交行列 P を用いて、

$$P^{-1}\sigma P = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \Lambda$$

と対角化可能である。(詳しくは線形代数のテキストを参考にされたい)。ここで λ_1, λ_2 は行列 σ の固有値である。実際、係数行列 $\sigma - \lambda \mathbf{I}$ の行列式が 0 という 2 次方程式：

$$\lambda^2 - (\sigma_x^2 + \sigma_y^2)\lambda + \sigma_x^2\sigma_y^2 - \sigma_{xy}^2 = 0$$

から、2 解 λ_1, λ_2 が固有値であった。方程式はとくに固有方程式とよばれている。上の式における σ を $\sigma = PAP^{-1}$ として目的関数の式に代入すると、 $\sigma_z^2 = \lambda_1 \tilde{a}^2 + \lambda_2 \tilde{b}^2$ に等しい。ただしベクトル $\begin{bmatrix} \tilde{a} \\ \tilde{b} \end{bmatrix} = P^{-1} \begin{bmatrix} a \\ b \end{bmatrix}$ 。この固有値の性質をみると、固有方程式の判別式より 2 つの解は実数解で、 $\sigma_x^2 + \sigma_y^2 \geq 0$, $\sigma_x^2\sigma_y^2 - \sigma_{xy}^2 \geq 0$ より 2 解とも非負でとなっている。したがって、条件を用いて $\lambda_M = \max(\lambda_1, \lambda_2)$, $\lambda_m = \min(\lambda_1, \lambda_2)$ とおくと、 $\lambda_M = \lambda_M \tilde{a}^2 + \lambda_M \tilde{b}^2 \geq \lambda_1 \tilde{a}^2 + \lambda_2 \tilde{b}^2 \geq \lambda_m \tilde{a}^2 + \lambda_m \tilde{b}^2 = \lambda_m$ から、結局、 z の分散 σ_z^2 は最大値が行列 σ の最大固有値となり、最小値は最小固有値となる。このようにして問題の最大値は分散共分散行列の最大固有値であることが分かる。また $\sigma = PAP^{-1}$ よりトレースの値は変わらない、 $\text{tr}(\Lambda) = \text{tr}(PAP^{-1}) = \text{tr}(\Lambda)$ 。すなわち $\sigma_x^2 + \sigma_y^2 = \lambda_1 + \lambda_2$ という関係が得られる。この式から、もとのデータと線形変換されたデータの 2 つには、分散の合計が同じという意味で等しい情報をもつととらえることができる。

特に大切なポイントとしては、得られた固有値の大きさが問題である。いま対角化するための直交行列をつくるベクトルの順序を換えて、固有値が大きいものから小さいものへとならぶようにする。 $\lambda_1 \geq \lambda_2$ のようにすると最大固有値に対する第一成分が分散の和 $\lambda_1 + \lambda_2$ のうち、最も大きな情報を含む。 $\frac{\lambda_1}{\lambda_1 + \lambda_2}$ この値が大きければ大きいほど、第 1 成分だけで多次元のデータを縮約している、第 1 成分で 2 次元データを説明できていることになる。すべての固有値の総和の中で 1 つの固有値の占める割合を寄与率とよび、この値の最大値に対する成分が第 1 主成分、つぎのものが第 2 主成分とよばれる。最大固有値にたいする固有ベクトルを $\begin{bmatrix} a^* \\ b^* \end{bmatrix}$ とすると、線形変換 $z = a^*x + b^*y$ が、変数 z の分散を最もおおくなるような変換であり、2 次元データ (x, y) を 1 次元データ z に寄与率 $\frac{\lambda_1}{\lambda_1 + \lambda_2}$ で縮小した第 1 主成分である。

変数が (x, y, z) の 3 個であっても、このデータを $w = ax + by + cz$ と 1 次元に縮約することが考えられる。3 次元空間に原点から延びる直線を引き、それを新しい軸として、空間の点をこの軸の上に射影して並べる。軸上に並んだ点の分散がなるべく大きくなるよう方向を決める。このときも、3 次の対称行列は直交行列で対角化され、分散共分散行列に対する最大固有値がその最大分散で、対応する固有ベクトルがその方向を決定するのである。さらに大きさの順に固有値をならべ、第 1 主成分、第 2 主成分としていく。このように、与えられた高次元のデータをなるべく少ない個数の主成分で説明しようということが主成分分析である。

例題.

変数 x, y の平均、分散、共分散が $\bar{x} = k, \bar{y} = l, \sigma_x^2 = A, \sigma_y^2 = B, \sigma_{xy} = C$ と与えられたとき、 $z = ax + by$ の分散が大きくなるよう a, b を定めよ。解.

分散共分散行列は $\begin{bmatrix} A & C \\ C & B \end{bmatrix}$ 。固有方程式は $\lambda^2 - (A + B)\lambda + AB - C^2 = 0$ であるから、解を $\lambda_1 \geq \lambda_2$ として a と b にかんする連立方程式

$$\begin{cases} (A - \lambda_1)a + Cb = 0 \\ Ca + (B - \lambda_1)b = 0 \end{cases}$$

を求め、 $a^* = \frac{a}{\sqrt{a^2 + b^2}}, b^* = \frac{b}{\sqrt{a^2 + b^2}}$ とおき、 $z = a^*x + b^*y$ が分散最大となる代表値である。