鹿児島大学　渡邉研究室
wata104@eee

相関を表す直線の方程式

1. はじめに

対になった 2 組のデータ x_i, y_i（$i = 1, 2, \cdots, n$）の間に線形の相関があるとき、通常は最小 2 乗法により、y_i の残差の 2 乗和が最小になるような回帰直線を求めて解析を行う。これは、変数 x の値が与えられたときに、最も確からしい y の値を推定するという考え方である。これに対して、2 次元の散布図上に n 個の点 (x_i, y_i) が与えられたときに、それらの近傍を通る最も確からしい直線を引くことを考えると、各点 (x_i, y_i) から直線までの距離の 2 乗和が最小になる直線を求めることになる。このようにして定めた直線は、最小 2 乗法で求めた回帰直線とは一般に傾きが異なる。本稿では、まず、通常の最小 2 乗法による回帰直線の求め方を述べた後、与えられた n 個の点 (x_i, y_i) から直線までの距離の 2 乗和が最小になる直線を求める方法を示す。続いて、その直線の方向が、主成分分析における共分散行列の固有ベクトルの向きと一致することを示す。

2. 通常の最小 2 乗法

対になった 2 組のデータ x_i, y_i（$i = 1, 2, \cdots, n$）を回帰直線

$$y = ax + b$$

で表すとき、y_i の残差の 2 乗和

$$V = \sum_{i=1}^{n} \big(y_i - (ax_i + b)\big)^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 = \sum_{i=1}^{n} (ax_i - y_i + b)^2$$

を最小にする a, b は

$$\frac{\partial V}{\partial a} = 2\sum_{i=1}^{n} x_i(ax_i - y_i + b) = 0$$

$$\frac{\partial V}{\partial b} = 2\sum_{i=1}^{n} (ax_i - y_i + b) = 0$$

より、連立方程式

$$a\sum_{i=1}^{n} {x_i}^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} 1 = \sum_{i=1}^{n} y_i$$

の解である。ここで、

$$\sum_{i=1}^{n} 1 = n$$

であり、また

$$\Sigma_{x^2} = \sum_{i=1}^{n} {x_i}^2, \qquad \Sigma_{xy} = \sum_{i=1}^{n} x_i y_i, \qquad \Sigma_x = \sum_{i=1}^{n} x_i, \qquad \Sigma_y = \sum_{i=1}^{n} y_i$$

と表すと

$$a\Sigma_{x^2} + b\Sigma_x = \Sigma_{xy}$$

$$a\Sigma_x + bn = \Sigma_y$$

となる。これを解いて

$$a = \frac{n\Sigma_{xy} - \Sigma_x\Sigma_y}{n\Sigma_{x^2} - {\Sigma_x}^2}$$

$$b = \frac{\Sigma_{x^2}\Sigma_y - \Sigma_{xy}\Sigma_x}{n\Sigma_{x^2} - {\Sigma_x}^2}$$

を得る。この解は、偏差の 2 乗和

$$S_{xy} = \Sigma_{xy} - \frac{\Sigma_x\Sigma_y}{n}, \qquad S_{xx} = \Sigma_{x^2} - \frac{{\Sigma_x}^2}{n}$$

および平均

$$\bar{x} = \frac{\Sigma_x}{n}, \qquad \bar{y} = \frac{\Sigma_y}{n}$$

を用いて

$$a = \frac{n\Sigma_{xy} - \Sigma_x\Sigma_y}{n\Sigma_{x^2} - {\Sigma_x}^2} = \frac{\Sigma_{xy} - \frac{\Sigma_x\Sigma_y}{n}}{\Sigma_{x^2} - \frac{{\Sigma_x}^2}{n}} = \frac{S_{xy}}{S_{xx}}$$

$$b = \frac{\Sigma_{x^2}\Sigma_y - \Sigma_{xy}\Sigma_x}{n\Sigma_{x^2} - {\Sigma_x}^2} = \frac{\Sigma_{x^2}\frac{\Sigma_y}{n} - \Sigma_{xy}\frac{\Sigma_x}{n}}{\Sigma_{x^2} - \frac{{\Sigma_x}^2}{n}} = \frac{\Sigma_{x^2}\frac{\Sigma_y}{n} - \left(a\left(\Sigma_{x^2} - \frac{{\Sigma_x}^2}{n}\right) + \frac{\Sigma_x\Sigma_y}{n}\right)\frac{\Sigma_x}{n}}{\Sigma_{x^2} - \frac{{\Sigma_x}^2}{n}}$$

$$= \frac{\frac{\Sigma_y}{n}\left(\Sigma_{x^2} - \frac{{\Sigma_x}^2}{n}\right) - a\left(\Sigma_{x^2} - \frac{{\Sigma_x}^2}{n}\right)\frac{\Sigma_x}{n}}{\Sigma_{x^2} - \frac{{\Sigma_x}^2}{n}} = \frac{\Sigma_y}{n} - a\frac{\Sigma_x}{n} = \bar{y} - a\bar{x}$$

とも表せる。したがって、y_i の残差の 2 乗和 V が最小となる回帰直線を表す式は

$$y = ax + b = ax + (\bar{y} - a\bar{x}) = a(x - \bar{x}) + \bar{y}$$

となる。これは、傾きが $a = S_{xy}/S_{xx}$ で、平均の点 $(\bar{x}, \bar{y})$ を通る直線である。

3. 点と直線との距離についての最小 2 乗法

点 (x_i, y_i) と直線

$$y = ax + b$$

との距離は

$$d_i = \frac{|ax_i - y_i + b|}{\sqrt{a^2 + 1}}$$

である。したがって、その 2 乗和

$$D=\sum_{i=1}^{n}{d_i}^2=\frac{1}{a^2+1}\sum_{i=1}^{n}(ax_i-y_i+b)^2$$

を最小にする $a,\ b$ は

$$\frac{\partial D}{\partial a}=\frac{-2a}{(a^2+1)^2}\sum_{i=1}^{n}(ax_i-y_i+b)^2+\frac{1}{a^2+1}\sum_{i=1}^{n}2x_i(ax_i-y_i+b)=0$$

$$\frac{\partial D}{\partial b}=\frac{1}{\sqrt{a^2+1}}\sum_{i=1}^{n}2(ax_i-y_i+b)=0$$

より求められる。

まず、第 1 式より

$$-a\sum_{i=1}^{n}(a^2{x_i}^2+{y_i}^2-2ax_iy_i+2abx_i-2by_i+b^2)+(a^2+1)\sum_{i=1}^{n}(a{x_i}^2-x_iy_i+bx_i)=0$$

$$\left(-a^3+a(a^2+1)\right)\sum_{i=1}^{n}{x_i}^2-a\sum_{i=1}^{n}{y_i}^2+\left(2a^2-(a^2+1)\right)\sum_{i=1}^{n}x_iy_i$$

$$+\left(-2a^2b+b(a^2+1)\right)\sum_{i=1}^{n}x_i+2ab\sum_{i=1}^{n}y_i-ab^2\sum_{i=1}^{n}1=0$$

$$a\sum_{i=1}^{n}{x_i}^2-a\sum_{i=1}^{n}{y_i}^2+(a^2-1)\sum_{i=1}^{n}x_iy_i+(-a^2b+b)\sum_{i=1}^{n}x_i+2ab\sum_{i=1}^{n}y_i-ab^2n=0$$

$$a\sum_{i=1}^{n}{x_i}^2-a\sum_{i=1}^{n}{y_i}^2+(a^2-1)\sum_{i=1}^{n}x_iy_i-b(a^2-1)\sum_{i=1}^{n}x_i+2ab\sum_{i=1}^{n}y_i-ab^2n=0$$

となる。ここで、

$$\Sigma_{x^2}=\sum_{i=1}^{n}{x_i}^2,\qquad \Sigma_{y^2}=\sum_{i=1}^{n}{y_i}^2,\qquad \Sigma_{xy}=\sum_{i=1}^{n}x_iy_i,\qquad \Sigma_x=\sum_{i=1}^{n}x_i,\qquad \Sigma_y=\sum_{i=1}^{n}y_i$$

と表すと

$$a\Sigma_{x^2}-a\Sigma_{y^2}+(a^2-1)\Sigma_{xy}-b(a^2-1)\Sigma_x+2ab\Sigma_y-ab^2n=0$$

$$a\left(\Sigma_{x^2}-\Sigma_{y^2}\right)+(a^2-1)\Sigma_{xy}-b\left((a^2-1)\Sigma_x-2a\Sigma_y\right)-ab^2n=0$$

である。また、第 2 式より

$$a\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}y_i+bn=0$$

$$b=\frac{1}{n}\left(\sum_{i=1}^{n}y_i-a\sum_{i=1}^{n}x_i\right)=\frac{\Sigma_y-a\Sigma_x}{n}$$

であるから、これを代入して b を消去すると

$$a\left(\Sigma_{x^2}-\Sigma_{y^2}\right)+(a^2-1)\Sigma_{xy}-\frac{\Sigma_y-a\Sigma_x}{n}\left((a^2-1)\Sigma_x-2a\Sigma_y\right)-a\left(\frac{\Sigma_y-a\Sigma_x}{n}\right)^2n=0$$

$$a\left(\Sigma_{x^2}-\Sigma_{y^2}\right)+(a^2-1)\Sigma_{xy}-\frac{\Sigma_y-a\Sigma_x}{n}\left(\left((a^2-1)\Sigma_x-2a\Sigma_y\right)+a\left(\Sigma_y-a\Sigma_x\right)\right)=0$$

$$a\left(\Sigma_{x^2}-\Sigma_{y^2}\right)+(a^2-1)\Sigma_{xy}-\frac{\Sigma_y-a\Sigma_x}{n}\left(-\Sigma_x-a\Sigma_y\right)=0$$

$$a\left(\Sigma_{x^2}-\Sigma_{y^2}\right)+(a^2-1)\Sigma_{xy}-\frac{a\Sigma_x-\Sigma_y}{n}\left(a\Sigma_y+\Sigma_x\right)=0$$

$$a\left(\Sigma_{x^2}-\Sigma_{y^2}\right)+(a^2-1)\Sigma_{xy}-\frac{1}{n}\left((a^2-1)\Sigma_x\Sigma_y+a{\Sigma_x}^2-a{\Sigma_y}^2\right)=0$$

$$(a^2-1)\left(\Sigma_{xy}-\frac{\Sigma_x\Sigma_y}{n}\right)+a\left(\Sigma_{x^2}-\Sigma_{y^2}-\frac{{\Sigma_x}^2-{\Sigma_y}^2}{n}\right)=0$$

$$(a^2-1)\left(\Sigma_{xy}-\frac{\Sigma_x\Sigma_y}{n}\right)+a\left(\Sigma_{x^2}-\frac{{\Sigma_x}^2}{n}-\Sigma_{y^2}+\frac{{\Sigma_y}^2}{n}\right)=0$$

となり、a の 2 次方程式

$$\left(\Sigma_{xy}-\frac{\Sigma_x\Sigma_y}{n}\right)a^2+\left(\Sigma_{x^2}-\frac{{\Sigma_x}^2}{n}-\Sigma_{y^2}+\frac{{\Sigma_y}^2}{n}\right)a-\left(\Sigma_{xy}-\frac{\Sigma_x\Sigma_y}{n}\right)=0$$

が得られる。ここで、偏差の 2 乗和

$$S_{xy}=\Sigma_{xy}-\frac{\Sigma_x\Sigma_y}{n},\qquad S_{xx}=\Sigma_{x^2}-\frac{{\Sigma_x}^2}{n},\qquad S_{yy}=\Sigma_{y^2}-\frac{{\Sigma_y}^2}{n}$$

を用いると

$$S_{xy}a^2+\left(S_{xx}-S_{yy}\right)a-S_{xy}=0$$

と表され、これを解くと

$$a=\frac{-\left(S_{xx}-S_{yy}\right)\pm\sqrt{\left(S_{xx}-S_{yy}\right)^2+4{S_{xy}}^2}}{2S_{xy}}$$

$$b=\frac{\Sigma_y-a\Sigma_x}{n}=\bar{y}-a\bar{x}$$

を得る。ただし、平均

$$\bar{x}=\frac{\Sigma_x}{n},\qquad \bar{y}=\frac{\Sigma_y}{n}$$

を用いた。したがって、この直線も、平均の点 $(\bar{x},\bar{y})$ を通る直線になる。直線の傾き a については 2 つの解 a_+, a_- があるが、解と係数の関係より $a_+\cdot a_-=-1$ であるから、2 つの解に対応する直線は互いに直交している。点から直線までの距離の 2 乗和 D が最小になる直線の傾きは、複号が正の解

$$a_+=\frac{-\left(S_{xx}-S_{yy}\right)+\sqrt{\left(S_{xx}-S_{yy}\right)^2+4{S_{xy}}^2}}{2S_{xy}}$$

であり、複号が負の解

$$a_{-} = \frac{-(S_{xx} - S_{yy}) - \sqrt{(S_{xx} - S_{yy})^2 + 4{S_{xy}}^2}}{2S_{xy}}$$

は、D が最大になる直線の傾きである。

ここで、D が最小になる直線の傾き a_{+} は、$S_{xx} \gg S_{yy}$ のとき、

$$a_{+} \cong \frac{-S_{xx} + \sqrt{{S_{xx}}^2 + 4{S_{xy}}^2}}{2S_{xy}}$$

となり、さらに $S_{xx} \gg S_{xy}$ のとき、

$$a_{+} \cong \frac{S_{xx}}{2S_{xy}}\left(-1 + \sqrt{1 + \frac{4{S_{xy}}^2}{{S_{xx}}^2}}\right) \cong \frac{S_{xx}}{2S_{xy}}\left(-1 + \left(1 + \frac{2{S_{xy}}^2}{{S_{xx}}^2}\right)\right) = \frac{S_{xy}}{S_{xx}}$$

となる。これは、通常の最小 2 乗法で得られる回帰直線（y_i の残差の 2 乗和 V が最小となる回帰直線）の傾きに一致する。

4. 主成分分析

主成分分析における共分散行列の固有値方程式は、固有値を λ、固有ベクトルを (u_x, u_y) とすると、

$$\begin{pmatrix} S_{xx} & S_{xy} \\ S_{xy} & S_{yy} \end{pmatrix}\begin{pmatrix} u_x \\ u_y \end{pmatrix} = \lambda \begin{pmatrix} u_x \\ u_y \end{pmatrix}$$

より

$$\begin{pmatrix} S_{xx} - \lambda & S_{xy} \\ S_{xy} & S_{yy} - \lambda \end{pmatrix}\begin{pmatrix} u_x \\ u_y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

となる。これが $u_x = u_y = 0$ 以外の解をもつための条件は

$$(S_{xx} - \lambda)(S_{yy} - \lambda) - {S_{xy}}^2 = 0$$

$$\lambda^2 - (S_{xx} + S_{yy})\lambda + S_{xx}S_{yy} - {S_{xy}}^2 = 0$$

である。これを解いて

$$\lambda = \frac{(S_{xx} + S_{yy}) \pm \sqrt{(S_{xx} + S_{yy})^2 - 4(S_{xx}S_{yy} - {S_{xy}}^2)}}{2}$$

$$= \frac{(S_{xx} + S_{yy}) \pm \sqrt{(S_{xx} - S_{yy})^2 + 4{S_{xy}}^2}}{2}$$

を得る。これより

$$(S_{xx} - \lambda)u_x + S_{xy}u_y = 0$$

$$\left(S_{xx}-\frac{\left(S_{xx}+S_{yy}\right)\pm\sqrt{\left(S_{xx}-S_{yy}\right)^2+4{S_{xy}}^2}}{2}\right)u_x+S_{xy}u_y=0$$

$$\left(\frac{\left(S_{xx}-S_{yy}\right)\mp\sqrt{\left(S_{xx}-S_{yy}\right)^2+4{S_{xy}}^2}}{2}\right)u_x+S_{xy}u_y=0$$

であるから

$$\frac{u_y}{u_x}=-\frac{2S_{xy}}{\left(S_{xx}-S_{yy}\right)\mp\sqrt{\left(S_{xx}-S_{yy}\right)^2+4{S_{xy}}^2}}$$

$$=-\frac{2S_{xy}}{\left(S_{xx}-S_{yy}\right)^2-\left(\left(S_{xx}-S_{yy}\right)^2+4{S_{xy}}^2\right)}\left(\left(S_{xx}-S_{yy}\right)\pm\sqrt{\left(S_{xx}-S_{yy}\right)^2+4{S_{xy}}^2}\right)$$

$$=\frac{\left(S_{xx}-S_{yy}\right)\pm\sqrt{\left(S_{xx}-S_{yy}\right)^2+4{S_{xy}}^2}}{2S_{xy}}$$

となる。これは、3 項で求めた直線の傾き a_+, a_- と一致している。すなわち、主成分分析における共分散行列の固有値ベクトルの向きは、平均の点 $(\bar{x},\bar{y})$ を通り、各点 (x_i,y_i) から直線までの距離の 2 乗和が最小または最大になる直線の方向である。

5. まとめ

対になった 2 組のデータ x_i, $y_i(i=1,2,\cdots,n)$の相関を表す直線として、n 個の点 (x_i,y_i) から直線までの距離の 2 乗和が最小になる直線を求める方法を示した。この直線の傾きは、通常の最小 2 乗法で求められる回帰直線の傾きと一般には異なる。さらに、点から直線までの距離の 2 乗和が最小と最大になる直線の方向が、それぞれ、主成分分析における共分散行列の固有ベクトルの向きと一致することを示した。

参考文献

・金谷 健一「これなら分かる応用数学教室－最小二乗法からウェーブレットまで－」共立出版、2003
・永田 一清、飯尾 勝矩、宮田 保教「基礎物理実験」東京教学社、1989
・永島 弘文、大舘 孝幸、荒井 太紀雄「最小距離 2 乗法による回帰直線の求め方」体外循環技術、vol. 12、no. 1、pp. 51-54、1986