平成 16 年 10 月 18 日

相関係数に関する一考察

新潟工科大学 情報電子工学科 竹野茂治

1 はじめに

以前、確率・統計の講義を行なったときに、回帰直線と相関係数の話で疑問に思った箇所があった。それを計算したときのノートを元に、ここにまとめておくことにする。

2 通常の相関係数の定義

まず、通常の相関係数の話を簡単に述べる。

2 次元のデータ (x_j, y_j) $(j = 1, 2, \ldots, n)$ があるとき、これを xy 平面上に表示したときに (散布図)、その点がある直線に近い、すなわち x と y にほぼ一次的な関係があるときに相関があると言い、そういう直線的な相関の見られないデータを相関がない、と言う。

その相関を計る指標として相関係数がある。それは以下のように定義される。まず、x_j の標本平均 $\overline{x}$、 y_j の標本平均 $\overline{y}$ を

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{x_1 + x_2 + \cdots + x_n}{n}, \quad \overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

と定め、x の平方和 S_{xx}, y の平方和 S_{yy}, および x と y の積和 S_{xy} を

$$S_{xx} = \sum_{j=1}^{n}(x_j - \overline{x})^2, \quad S_{yy} = \sum_{j=1}^{n}(y_j - \overline{y})^2, \quad S_{xy} = \sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y})$$

と定める。このとき、相関係数 r は

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \tag{1}$$

で定義される。そして、$|r| \leq 1$ であり、$r = 1$ に近ければ正の相関 (傾きが正の直線による相関)、$r = -1$ に近ければ負の相関 (傾きが負の直線による相関)、$r = 0$ に近ければ相関がない、とするのである。

この $|r| \leq 1$ であること、そして $r = \pm 1$ のときにデータが本当に一直線上にのるかを以下に説明する。

n 次元ベクトル $\vec{x}, \vec{y}$ を

$$\vec{x} = (x_1 - \overline{x}\,, x_2 - \overline{x}\,, \ldots, x_n - \overline{x}\,), \quad \vec{y} = (y_1 - \overline{y}\,, y_2 - \overline{y}\,, \ldots, y_n - \overline{y}\,)$$

とすると、

$$S_{xy} = \vec{x} \cdot \vec{y}, \quad S_{xx} = |\vec{x}|^2, \quad S_{yy} = |\vec{y}|^2$$

なので

$$r = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|}$$

となる。厳密には、シュワルツの不等式から、

$$-|\vec{x}||\vec{y}| \leq \vec{x} \cdot \vec{y} \leq |\vec{x}||\vec{y}|$$

で、かつ等号成立は $\vec{x} // \vec{y}$ となることが導かれ、よって $-1 \leq r \leq 1$ で、

$$r = -1 \Rightarrow \vec{y} = -\alpha\vec{x}\ (\alpha > 0), \quad r = 1 \Rightarrow \vec{y} = \alpha\vec{x}\ (\alpha > 0)$$

となることが言えるのであるが、多少図形的なイメージで説明すると、高校の内積の定義にあるように

$$\vec{x} \cdot \vec{y} = |\vec{x}||\vec{y}| \cos\theta \quad (\theta \text{ は } \vec{x} \text{ と } \vec{y} \text{ のなす角}, \ 0 \leq \theta \leq \pi)$$

なので $r = \vec{x} \cdot \vec{y} / |\vec{x}||\vec{y}| = \cos\theta$ となり、よってまず $|r| \leq 1$ がいえる。

$r = -1$ となるのは $\theta = \pi$ のときなので $\vec{x}$ と $\vec{y}$ が丁度逆向きのベクトルのとき、すなわち $\vec{y} = -\alpha\vec{x}\ (\alpha > 0)$ となるが、それを成分で見ると

$$y_j - \overline{y}\ = -\alpha(x_j - \overline{x}\,) \quad (j = 1, 2, \ldots, n)$$

となり、これは (x_j, y_j) が一つの直線 $y - \overline{y}\ = -\alpha(x - \overline{x}\,)$ 上にあることを意味することになる。

$r = 1$ の場合も同様で、この場合は $\theta = 0$ となるので $\vec{x}$ と $\vec{y}$ が同じ向きのベクトルになり、後は上の $-\alpha$ を α に変えれば良い。

これにより、$-1 \leq r \leq 1$ で、$r = 1$ と $r = -1$ は確かに直線相関なので、そこから離れて 0 に近くなると確かに相関が小さいような気がする。しかし、例えば $r = 0$ の場合に本当に相関がない、と言えるのだろうか。上の式によれば $r = 0$ の場合は $\vec{x} \cdot \vec{y} = 0$、すなわち $\vec{x} \perp \vec{y}$ を意味するが、それが「相関がない状態を意味している」と見なせるだろうか。

これが私が感じた最初の疑問である。

問題 1 $r = 0$ の状態は本当に (直線的な) 相関がない、といえるのだろうか

そして、r に含まれる式をみていてぼんやり思ったのは以下の疑問である。

問題 2 r の値は、(x_j, y_j) 全体を原点の周りに θ だけ回転しても変わらないだろうか

本来 r が「直線相関」を計る指標である、というからにはそのような不変性も同時に備えている必要があると思うが、果して r にはそのような性質があるだろうか。これは後 (4 節) で検証する。

3 通常の回帰直線の定義

次に通常の回帰直線の話を簡単に述べる。

回帰直線とは、データに直線的な相関がある場合に、それを表す、最も妥当だと思われる直線を回帰直線という。実際にはそれは以下のようにして求める。

まず、求める直線を $y = ax + b$ とすると、データ (x_j, y_j) に直線相関がある場合は $y_j \approx ax_j + b$ となるので、その誤差 $(= y_j - (ax_j + b))$ の平方和を最も小さくする a,b を取る、という最小自乗法を用いる。すなわち、

$$f(a,b) = \sum_{j=1}^{n} \{y_j - (ax_j + b)\}^2$$

として、この 2 変数関数 $f(a,b)$ の最小値を与える a, b を求める。通常の教科書では、偏微分を用いてこの 2 変数関数 $f(a,b)$ の最小値を求めるものが多いように思うが、$f(a,b)$ は 2 次式なので、ここではより素朴な方法、すなわち

b に関して最小になるところの中で、a に関して最小になるところを求める

によって求めることにする。そのために次の性質を利用する。

$$\begin{aligned} S_{xx} &= \sum_j (x_j - \overline{x}\,)^2 = \sum_j (x_j^2 - 2x_j\overline{x} + \overline{x}^{\,2}) \\ &= \sum_j x_j^2 - 2\overline{x} \sum_j x_j + n\overline{x}^{\,2} = n\overline{x^2} - 2n\overline{x}^{\,2} + n\overline{x}^{\,2} \quad (\sum_j x_j = n\overline{x}\,) \\ &= n(\overline{x^2} - \overline{x}^{\,2}), \end{aligned}$$

$$
\begin{aligned}
S_{yy} &= n(\overline{y^2} - \overline{y}^{\,2}),\\
S_{xy} &= \sum_j (x_j - \overline{x}\,)(y_j - \overline{y}\,) = \sum_j (x_j y_j - \overline{x}\,y_j - \overline{y}\,x_j + \overline{x}\,\overline{y}\,)\\
&= \sum_j x_j y_j - \overline{x}\sum_j y_j - \overline{y}\sum_j x_j + n\overline{x}\,\overline{y} = n\overline{xy} - n\overline{x}\,\overline{y} - n\overline{y}\,\overline{x} + n\overline{x}\,\overline{y}\\
&= n(\overline{xy} - \overline{x}\,\overline{y}\,)
\end{aligned}
$$

ここで、

$$
\overline{x^2} = \frac{1}{n}\sum_{j=1}^{n} x_j^2, \quad \overline{y^2} = \frac{1}{n}\sum_{j=1}^{n} y_j^2, \quad \overline{xy} = \frac{1}{n}\sum_{j=1}^{n} x_j y_j
$$

などとした。

これらを用いると、$f(a,b)$ は以下のように展開される。

$$
\begin{aligned}
f(a,b) &= \sum_j \{y_j - (ax_j + b)\}^2 = \sum_j \{y_j^2 - 2y_j(ax_j + b) + (ax_j + b)^2\}\\
&= \sum_j (y_j^2 - 2ax_j y_j - 2by_j + a^2 x_j^2 + 2abx_j + b^2)\\
&= n(\overline{y^2} - 2a\overline{xy} - 2b\overline{y} + a^2\overline{x^2} + 2ab\overline{x} + b^2)
\end{aligned}
$$

これを b に関する 2 次式と見て、b について整理する。

$$
\begin{aligned}
\frac{1}{n}f(a,b) &= b^2 + 2(a\overline{x} - \overline{y}\,)b + a^2\overline{x^2} - 2a\overline{xy} + \overline{y^2}\\
&= (b + a\overline{x} - \overline{y}\,)^2 - (a\overline{x} - \overline{y}\,)^2 + a^2\overline{x^2} - 2a\overline{xy} + \overline{y^2}\\
&= (b + a\overline{x} - \overline{y}\,)^2 + a^2(\overline{x^2} - \overline{x}^{\,2}) - 2a(\overline{xy} - \overline{x}\,\overline{y}\,) + \overline{y^2} - \overline{y}^{\,2}\\
&= (b + a\overline{x} - \overline{y}\,)^2 + \frac{a^2}{n}S_{xx} - \frac{2a}{n}S_{xy} + \frac{1}{n}S_{yy}
\end{aligned}
$$

よって、$f(a,b)$ は、b に関しては $b = \overline{y} - a\overline{x}$ のときに最小になり、その最小値は

$$
f_1(a) = f(a, \overline{y} - a\overline{x}\,) = a^2 S_{xx} - 2aS_{xy} + S_{yy}
$$

である。これは a に関する 2 次式であるから、これを今度は a について整理すれば、

$$
\begin{aligned}
f_1(a) &= a^2 S_{xx} - 2aS_{xy} + S_{yy} = S_{xx}\left(a^2 - 2a\frac{S_{xy}}{S_{xx}}\right) + S_{yy}\\
&= S_{xx}\left(a - \frac{S_{xy}}{S_{xx}}\right)^2 - S_{xx}\frac{S_{xy}^2}{S_{xx}^2} + S_{yy}\\
&= S_{xx}\left(a - \frac{S_{xy}}{S_{xx}}\right)^2 + \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}}
\end{aligned}
$$

となる。S_{xx} は定義より 0 以上で、これが 0 ではないとすれば (通常はそう)、$f_1(a)$ は $a = S_{xy}/S_{xx}$ のときに最小となり、その最小値は

$$f_m = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}} (= S_{yy}(1 - r^2))$$

となる。よって、

$$b = \overline{y} - a\overline{x}\,, \quad a = \frac{S_{xy}}{S_{xx}}$$

のときに回帰直線となり、よってそれは

$$y - \overline{y} = a(x - \overline{x}\,) = \frac{S_{xy}}{S_{xx}}(x - \overline{x}\,)$$

となる。これが通常教科書に書かれている結果である。

しかし、この式は明らかに x,y に関して対称ではない。すなわち、「元のデータを (x_j, y_j) とみて回帰直線を求めたもの」と、「元のデータを (y_j, x_j) とみて回帰直線を求めたもの」は、$y = x$ に関して対称にはならない。

例えば「x = 身長、y = 体重」のようなデータの場合、どちらを横軸に取ってどちらを縦軸に取るか、ということに余り意味はなさそうであるが、上の非対称性は、そのどちらを横軸に取るかで回帰直線が本質的に変わってしまう、ということを意味している。

前の、相関係数の回転不変性に対する疑問と同様に、これも直線相関を意味するものとして適当なのだろうかと疑問に思う。

そのような非対称性が起こるのは、もちろん、

$$f(a, b) = \sum_{j=1}^{n} \{y_j - (ax_j + b)\}^2$$

の定義に問題がある。つまり、この式は「データと直線の距離の平方和」を意味しているのではなく、「データと、それと同じ x 座標を持つ直線上の点との距離の平方和」を取っていて、すなわち y 軸に平行に距離を計っているためにそのような対称性が崩れてしまう。その定義からもすぐに分かるが、回帰直線にも回転不変性はない。

回帰直線に回転不変性や、x,y の入れ替えに対する不変性を持たせるためには、「データと直線の距離の平方和」を最小にする直線を考えれば良い。これらの疑問をまとめると以下のようになる。

問題 3　データ点と直線の距離の平方和を最小にする直線はどのような式になるか、また、なぜ通常それを用いないのか

4 回転不変性について

これまでにあげた疑問を考えていくが、まずは問題 2 としてあげた相関係数や回帰直線の回転不変性について考える。

データ (x_j, y_j) を、原点の周りに θ 回転したデータを (x'_j, y'_j) とする。すなわち

$$\begin{bmatrix} x'_j \\ y'_j \end{bmatrix} = A(\theta) \begin{bmatrix} x_j \\ y_j \end{bmatrix}, \quad A(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

とすれば、

$$\begin{bmatrix} \overline{x'} \\ \overline{y'} \end{bmatrix} = A(\theta) \begin{bmatrix} \overline{x} \\ \overline{y} \end{bmatrix}, \quad \begin{bmatrix} x'_j - \overline{x'} \\ y'_j - \overline{y'} \end{bmatrix} = A(\theta) \begin{bmatrix} x_j - \overline{x} \\ y_j - \overline{y} \end{bmatrix}$$

なので、

$$\begin{aligned}
S_{x'x'} &= \sum_j (x'_j - \overline{x'})^2 = \sum_j \{(x_j - \overline{x})\cos\theta - (y_j - \overline{y})\sin\theta\}^2 \\
&= S_{xx}\cos^2\theta - 2S_{xy}\cos\theta\sin\theta + S_{yy}\sin^2\theta, \\
S_{x'y'} &= \sum_j (x'_j - \overline{x'})(y'_j - \overline{y'}) \\
&= \sum_j \{(x_j - \overline{x})\cos\theta - (y_j - \overline{y})\sin\theta\}\{(x_j - \overline{x})\sin\theta + (y_j - \overline{y})\cos\theta\} \\
&= S_{xx}\cos\theta\sin\theta + S_{xy}(\cos^2\theta - \sin^2\theta) - S_{yy}\cos\theta\sin\theta, \\
S_{y'y'} &= \sum_j (y'_j - \overline{y'})^2 = \sum_j \{(x_j - \overline{x})\sin\theta + (y_j - \overline{y})\cos\theta\}^2 \\
&= S_{xx}\sin^2\theta + 2S_{xy}\cos\theta\sin\theta + S_{yy}\cos^2\theta
\end{aligned}$$

となる。この式から、r が θ に関して不変でないことはすぐに分かる。

しかし、回転不変な式もいくつか容易に見つかる。例えば

$$S_{x'x'} + S_{y'y'} = S_{xx} + S_{yy} \tag{2}$$

であるし、また、

$$\begin{aligned}
S_{x'x'} - S_{y'y'} &= (S_{xx} - S_{yy})(\cos^2\theta - \sin^2\theta) - 4S_{xy}\cos\theta\sin\theta \\
&= (S_{xx} - S_{yy})\cos 2\theta - 2S_{xy}\sin 2\theta, \\
2S_{x'y'} &= (S_{xx} - S_{yy})\sin 2\theta + 2S_{xy}\cos 2\theta
\end{aligned}$$

より、

$$(S_{x'x'} - S_{y'y'})^2 + 4S_{x'y'}^2 = (S_{xx} - S_{yy})^2 + 4S_{xy}^2 \tag{3}$$

のような不変量も得られるし、これら 2 つを組み合わせて $(((3) - (2)^2)/4)$、

$$S_{x'y'}^2 - S_{x'x'}S_{y'y'} = S_{xy}^2 - S_{xx}S_{yy} \tag{4}$$

のような不変量も得られる。

また、この回転されたデータに対する回帰直線は

$$y' - \overline{y'} = a'(x' - \overline{x'}) = \frac{S_{x'y'}}{S_{x'x'}}(x' - \overline{x'})$$

であり、これは

$$\begin{bmatrix} x' - \overline{x'} \\ y' - \overline{y'} \end{bmatrix} = t \begin{bmatrix} 1 \\ a' \end{bmatrix}$$

とパラメータ表示される。よって、この両辺に $A(-\theta)$ をかけてこの直線を原点の周りに $(-\theta)$ 回転すると

$$\begin{aligned} & A(-\theta) \begin{bmatrix} x' - \overline{x'} \\ y' - \overline{y'} \end{bmatrix} = \begin{bmatrix} x - \overline{x} \\ y - \overline{y} \end{bmatrix} \\ = \quad & tA(-\theta) \begin{bmatrix} 1 \\ a' \end{bmatrix} = t \begin{bmatrix} \cos\theta + a'\sin\theta \\ -\sin\theta + a'\cos\theta \end{bmatrix} = \frac{t}{S_{x'x'}} \begin{bmatrix} S_{x'x'}\cos\theta + S_{x'y'}\sin\theta \\ -S_{x'x'}\sin\theta + S_{x'y'}\cos\theta \end{bmatrix} \end{aligned}$$

となる。ここで、

$$\begin{aligned} & S_{x'x'}\cos\theta + S_{x'y'}\sin\theta \\ = \quad & (S_{xx}\cos^2\theta - 2S_{xy}\cos\theta\sin\theta + S_{yy}\sin^2\theta)\cos\theta \\ & +\{S_{xx}\cos\theta\sin\theta + S_{xy}(\cos^2\theta - \sin^2\theta) - S_{yy}\cos\theta\sin\theta\}\sin\theta \\ = \quad & S_{xx}\cos\theta - S_{xy}\sin\theta, \\ & -S_{x'x'}\sin\theta + S_{x'y'}\cos\theta \\ = \quad & -(S_{xx}\cos^2\theta - 2S_{xy}\cos\theta\sin\theta + S_{yy}\sin^2\theta)\sin\theta \\ & +\{S_{xx}\cos\theta\sin\theta + S_{xy}(\cos^2\theta - \sin^2\theta) - S_{yy}\cos\theta\sin\theta\}\cos\theta \\ = \quad & S_{xy}\cos\theta - S_{yy}\sin\theta \end{aligned}$$

となるので、この直線の傾き $a''(\theta)$ は

$$a''(\theta) = \frac{-S_{x'x'}\sin\theta + S_{x'y'}\cos\theta}{S_{x'x'}\cos\theta + S_{x'y'}\sin\theta} = \frac{S_{xy}\cos\theta - S_{yy}\sin\theta}{S_{xx}\cos\theta - S_{xy}\sin\theta}$$

となる。これは $\theta \neq 0$ であればもちろん通常の回帰直線の傾き $a = S_{xy}/S_{xx}$ とは一致しない。つまり、回帰直線も回転不変性を持たないことがわかる。

なお、この $a''(\theta)$ の、$\theta = 90^\circ$ のときの値は、

(x_j, y_j) を 90° 回したデータに対する回帰直線を -90° 回した直線の傾き

を意味するが、y 軸に関して折り返して考えれば容易に分かるが、その傾きは、3 節でも言及した、

(y_j, x_j) に対する回帰直線を、$y = x$ に関して対称に折り返したものの傾き

に等しい。つまり、そのような直線を表す式は

$$y - \overline{y} = \tilde{a}(x - \overline{x}) = \frac{S_{yy}}{S_{xy}}(x - \overline{x})$$

であることがわかる。この直線も元の回帰直線とは一致しない。

5 点と直線の距離を用いた回帰直線

この節では、通常の回帰直線とは違い、データ点と直線の距離の平方和を最小にする直線を求めることにする。

直線を $y = ax + b$ として、3 節と同様に行なう。ただし、この場合は $f(a, b)$ の代わりに

$$g(a, b) = \sum_{j=1}^{n} d_j^2 \quad (d_j = (x_j, y_j) \text{ と } y = ax + b \text{ との距離})$$

を考えることになる。

ところで、d_j と $|y_j - (ax_j + b)|$ を比較すると、直線の傾きが a なので、

$$d_j : |y_j - (ax_j + b)| = 1 : \sqrt{a^2 + 1}$$

となり、よって

$$g(a,b) = \frac{1}{a^2+1} f(a,b)$$

であることがわかる。よって、b に関する最小値は 3 節の計算と同じで、$b = \overline{y} - a\overline{x}$ のときにとる。その最小値 $g_1(a)$ は

$$g_1(a) = \frac{1}{a^2+1} f_1(a) = \frac{a^2 S_{xx} - 2aS_{xy} + S_{yy}}{a^2+1}$$

となる。この分数関数の最小値を求めれば良い。微分すると、

$$\begin{aligned} \frac{d}{da} g_1(a) &= \frac{(2aS_{xx} - 2S_{xy})(a^2+1) - 2a(a^2 S_{xx} - 2aS_{xy} + S_{yy})}{(a^2+1)^2} \\ &= \frac{2\{a^2 S_{xy} + a(S_{xx} - S_{yy}) - S_{xy}\}}{(a^2+1)^2} \end{aligned}$$

となる。この分子の a に関する 2 次式は、判別式が

$$D = (S_{xx} - S_{yy})^2 + 4S_{xy}^2 \geq 0$$

となるので、

$$S_{xy}(a - \lambda_1)(a - \lambda_2)$$

と書ける。ここで、λ_1, λ_2 は

$$\lambda_1 = \frac{S_{yy} - S_{xx} - \sqrt{D}}{2S_{xy}}, \quad \lambda_2 = \frac{S_{yy} - S_{xx} + \sqrt{D}}{2S_{xy}}$$

であり、これにより、$g_1(a)$ の微分は

$$\frac{d}{da} g_1(a) = \frac{S_{xy}(a - \lambda_1)(a - \lambda_2)}{(a^2+1)^2}$$

となる。

1. $S_{xy} > 0$ のとき
   このとき、$\lambda_1 < \lambda_2$ であり、よって最小値は $a = -\infty$ か $a = \lambda_2$ で取る。

2. $S_{xy} < 0$ のとき
   このときは、$\lambda_1 > \lambda_2$ であるが、$g_1'(a)$ には S_{xy} がかかっているので、最小値は $a = \infty$ か $a = \lambda_2$ で取る。

3. $S_{xy} = 0$ のとき
   このときは、

$$g_1(a) = \frac{a^2 S_{xx} + S_{yy}}{a^2 + 1} = S_{xx} + \frac{S_{yy} - S_{xx}}{a^2 + 1}$$

   より、$S_{yy} > S_{xx}$ ならば $|a| = \infty$ のときに最小値 S_{xx} を、$S_{yy} < S_{xx}$ ならば $a = 0$ のときに最小値 S_{yy} を、$S_{xx} = S_{yy}$ ならばつねに S_{xx} に等しい値を取る。

$g_1(\pm\infty) = S_{xx}$ であるから、次は $S_{xy} \neq 0$ のときに、これと $g_1(\lambda_2)$ とを比較する。

$a = \lambda_2$ は

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

の解なので

$$S_{yy} - S_{xx} = \frac{a^2 - 1}{a} S_{xy}$$

となる。これにより、$a = \lambda_2$ に対し、

$$g_1(a) = S_{xx} + \frac{S_{yy} - S_{xx} - 2aS_{xy}}{a^2 + 1} = S_{xx} + \frac{(a^2 - 1)S_{xy} - 2a^2 S_{xy}}{a(a^2 + 1)} = S_{xx} - \frac{S_{xy}}{a}$$

となるが、

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

より、

$$\begin{aligned} \frac{S_{xy}}{a} &= aS_{xy} + S_{xx} - S_{yy} = \lambda_2 S_{xy} + S_{xx} - S_{yy} \\ &= \frac{S_{yy} - S_{xx} + \sqrt{D}}{2} + S_{xx} - S_{yy} \\ &= \frac{S_{xx} - S_{yy} + \sqrt{D}}{2} (= -\lambda_1 S_{xy}) \end{aligned}$$

となる。ここで、D の定義より、$\sqrt{D} \geq |S_{xx} - S_{yy}|$ (等号は $S_{xy} = 0$) であるので、

$$g_1(a) = S_{xx} - \frac{S_{xx} - S_{yy} + \sqrt{D}}{2}$$

は、$S_{xy} \neq 0$ のとき、確かに S_{xx} より小さく、よってこれが最小値となる。

結局、$g_1(a)$ の最小値は以下のようになる。

- $S_{xy} = 0$ のときは、$S_{xx} < S_{yy}$ ならば $|a| = \infty$ のときに最小値 S_{xx}、$S_{xx} > S_{yy}$ ならば $a = 0$ のときに最小値 S_{yy}、$S_{xx} = S_{yy}$ ならば全ての a に対し $g_1(a) = S_{xx}(= S_{yy})$ となる。
- $S_{xy} \neq 0$ のときは、$a = \lambda_2$ のときに最小値

  $$g_1(a) = \frac{S_{xx} + S_{yy} - \sqrt{D}}{2}$$

  を取る。

この $S_{xy} \neq 0$ のときの最小値 $g_1(\lambda_2)$ は、以下のように書き換えることができる。

$$\begin{aligned} g_1(\lambda_2) &= \frac{S_{xx} + S_{yy} - \sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy}^2}}{2} \\ &= \frac{S_{xx} + S_{yy} - \sqrt{(S_{xx} + S_{yy})^2 + 4(S_{xy}^2 - S_{xx}S_{yy})}}{2} \\ &= \frac{S_{xx} + S_{yy}}{2}\left\{1 - \sqrt{1 - 4\frac{S_{xx}S_{yy} - S_{xy}^2}{(S_{xx} + S_{yy})^2}}\right\} \end{aligned}$$

ここで、

$$\hat{r} = \sqrt{1 - 4\frac{S_{xx}S_{yy} - S_{xy}^2}{(S_{xx} + S_{yy})^2}} \left(= \frac{\sqrt{D}}{S_{xx} + S_{yy}}\right) \tag{5}$$

とすると、最小値 $g_1(\lambda_2)$ は

$$g_1(\lambda_2) = \frac{S_{xx} + S_{yy}}{2}(1 - \hat{r}) \tag{6}$$

と書ける。

なお、$S_{xy}=0$ の場合、$\hat{r}$ は

$$\hat{r}=\frac{|S_{xx}-S_{yy}|}{S_{xx}+S_{yy}}$$

となるので、

$$\frac{S_{xx}+S_{yy}}{2}(1-\hat{r})=\frac{S_{xx}+S_{yy}-|S_{xx}-S_{yy}|}{2}=\min\{S_{xx},S_{yy}\}$$

となり、式 (6) は $S_{xy}=0$ の場合も最小値を与えていることになる。

この $\hat{r}$ は、以下に述べるような色々な性質を持っている。

- 回転不変性
  4 節で見た回転不変量で表現されるので、回転不変性を持つ。
- 散布図の広がりに関わらない
  $S_{xx}+S_{yy}$ は

  $$S_{xx}+S_{yy}=\sum_{j=1}^{n}\{(x_j-\overline{x}\,)^2+(y_j-\overline{y}\,)^2\}$$

  であり、これは回転不変で、かつ散布図の広がり (2 次元的な分散) を表しているが、一方 $\hat{r}$ は、

  $$\frac{1-\hat{r}}{2}=\frac{g(a,b)\text{ の最小値}}{S_{xx}+S_{yy}}$$

  と書けるので、この右辺は散布図の広がり (スケール) には関わらない量になっているので、$\hat{r}$ も散布図の広がりには影響を受けない値となる。
- 直線相関をあらわす
  $(g(a,b)$ の最小値$)/(S_{xx}+S_{yy})$ は、もちろんそれが小さい程直線相関が強く、それが大きければ直線相関が弱くなる。2 節で見たように、$S_{xy}^2\le S_{xx}S_{yy}$ なので $\hat{r}$ は $0\le\hat{r}\le1$ の値を取り、$\hat{r}=1$ ならば $S_{xy}^2=S_{xx}S_{yy}$ となり、r の場合と同様、確かに完全な直線相関となる。
- $\hat{r}=0$ の状態が説明できる (直線相関がない)
  $\hat{r}=0$ のときは、$D=0$、すなわち

  $$S_{xx}=S_{yy}\quad\text{かつ}\quad S_{xy}=0$$

  となり、この場合は常に $g_1(a)=S_{xx}$ となる。つまり、直線が $(\overline{x}\,,\overline{y}\,)$ を通れば $(b=\overline{y}\,-a\overline{x}\,)$ $g(a,b)$ の値はその直線の傾き a にはよらない。これは $(\overline{x}\,,\overline{y}\,)$ を

通る、どのような方向の直線に対しても、データからの距離の平方和は一定である、ということを意味している。「どのような方向にもデータからの誤差が一定」ということは「どのような方向にも相関性はない」ということを意味しているように思える。

通常の相関係数は、$r = 0$ のときには $S_{xy} = 0$ しか得られないが、$\hat{r}$ の場合はそれに加えて $S_{xx} = S_{yy}$ も得られるので、$r = 0$ よりもやや強いことが言えるのである。

- $\hat{r} \geq |r|$
  相乗平均と相加平均の関係より、

$$4S_{xx}S_{yy} \leq (S_{xx} + S_{yy})^2$$

なので、$S_{xx}S_{yy} - S_{xy}^2 = S_{xx}S_{yy}(1 - r^2)$ より、

$$\begin{aligned} \hat{r} &= \frac{\sqrt{(S_{xx} + S_{yy})^2 - 4S_{xx}S_{yy}(1 - r^2)}}{S_{xx} + S_{yy}} \\ &\geq \frac{\sqrt{(S_{xx} + S_{yy})^2 - (S_{xx} + S_{yy})^2(1 - r^2)}}{S_{xx} + S_{yy}} = |r| \end{aligned}$$

以上のことから、ある意味ではむしろ r よりも優れている性質を持つ、あらたな「相関係数」$\hat{r}$ が得られたことになる。相関係数として $\hat{r}$ を使えば、問題 1 もある意味で解決する。

また、上で得られた「回帰直線」の傾き $\hat{a} = \lambda_2$ も、もちろん回転不変性 (すなわちデータの回転に合わせて直線も同じだけ回転) を持ち、x, y の入れ替えにも対応することが、その定義からすぐに分かる。さらに次も言える。

命題 1

$\hat{a} = \lambda_2$, $a = S_{xy}/S_{xx}$, およびデータの x,y を入れ替えて作った回帰直線を $y = x$ に関して対称に折り返した直線の傾き $\tilde{a} = S_{yy}/S_{xy}$ (cf. 4 節) に対して次が成り立つ。

$$\begin{cases} S_{xy} > 0 \;\Rightarrow\; \tilde{a} \geq \hat{a} \geq a > 0 \\ S_{xy} < 0 \;\Rightarrow\; \tilde{a} \leq \hat{a} \leq a < 0 \end{cases}$$

なお、4 つの不等号の等号成立は、いずれも完全な直線相関のとき ($|r| = 1$)。

証明

$\hat{a}$ は、

$$\hat{a} = \lambda_2 = \frac{S_{yy} - S_{xx} + \sqrt{D}}{2S_{xy}}$$

なので、

$$\begin{aligned} \tilde{a} - \hat{a} &= \frac{S_{yy}}{S_{xy}} - \frac{S_{yy} - S_{xx} + \sqrt{D}}{2S_{xy}} = \frac{S_{xx} + S_{yy} - \sqrt{D}}{2S_{xy}} \\ &= \frac{S_{xx} + S_{yy} - \sqrt{(S_{xx} + S_{yy})^2 - 4(S_{xx}S_{yy} - S_{xy}^2)}}{2S_{xy}} \end{aligned}$$

で、$S_{xx}S_{yy} \geq S_{xy}^2$ より $\tilde{a}$ と $\hat{a}$ の大小関係が得られる。そして等号成立は $S_{xx}S_{yy} = S_{xy}^2$、すなわち $|r| = 1$ のときであることもわかる。

また、

$$\begin{aligned} \hat{a} - a &= \frac{S_{yy} - S_{xx} + \sqrt{D}}{2S_{xy}} - \frac{S_{xy}}{S_{xx}} = \frac{(S_{xx} - S_{yy})^2 - D}{2S_{xy}(S_{yy} - S_{xx} - \sqrt{D})} - \frac{S_{xy}}{S_{xx}} \\ &= \frac{-2S_{xy}}{S_{yy} - S_{xx} - \sqrt{D}} - \frac{S_{xy}}{S_{xx}} = S_{xy}\frac{S_{xx} + S_{yy} - \sqrt{D}}{S_{xx}(\sqrt{D} + S_{xx} - S_{yy})} \end{aligned}$$

であり、$\sqrt{D} + S_{xx} - S_{yy} > 0$ より $\hat{a}$ と a の大小関係が得られる。等号成立はこちらも $S_{xx}S_{yy} = S_{xy}^2$ の場合となる。■

なお、$S_{xy} \to 0$ のときは、$a \to 0$, $\tilde{a}$ は

$$\lim_{S_{xy} \to \pm 0} \tilde{a} = \pm\infty$$

であるが、$\hat{a}$ は、$S_{yy} > S_{xx}$ のときは

$$\lim_{S_{xy} \to \pm 0} (S_{yy} - S_{xx} + \sqrt{D}) = 2(S_{yy} - S_{xx}) > 0$$

なので

$$\lim_{S_{xy} \to \pm 0} \hat{a} = \pm\infty$$

であり、$S_{yy} < S_{xx}$ のときは

$$\lim_{S_{xy}\to\pm 0} \hat{a} = \lim_{S_{xy}\to\pm 0} \frac{2S_{xy}}{S_{xx} - S_{yy} + \sqrt{D}} = \lim_{S_{xy}\to\pm 0} \frac{2S_{xy}}{2(S_{xx} - S_{yy})} = 0$$

となる。$S_{xx} = S_{yy}$ のときは、$\hat{a} = |S_{xy}|/S_{xy} = \mathrm{sgn} S_{xy}$ より、

$$\lim_{S_{xy}\to\pm 0} \hat{a} = \pm 1$$

となる。

以上が問題 3 の前半部分に対する答えとなる。

6 スケール変換に対する不変性

データの指標としては、スケール変換に対する不変性も重要な性質である。$r, \hat{r}, a, \hat{a}$ 等について、これも調べてみる。

$x'_j = Ax_j$, $y'_j = By_j$ $(j = 1, 2, \ldots, n,\ A, B$ は正の定数$)$ とすると、

$$\overline{x'} = A\overline{x}\,, \quad \overline{y'} = B\overline{y}\,, \quad S_{x'x'} = A^2 S_{xx}, \quad S_{x'y'} = ABS_{xy}, \quad S_{y'y'} = B^2 S_{yy}$$

となることが容易に分かる。よって、$r(x', y') = (x', y'$ に対する r の値$)$ 等とすると、

$$r(x', y') = \frac{S_{x'y'}}{\sqrt{S_{x'x'} S_{y'y'}}} = \frac{ABS_{xy}}{\sqrt{A^2 S_{xx} B^2 S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = r(x, y)$$

となり、通常の相関係数はこのスケール変換に対しては不変であることが分かる。

一方、新たに作った $\hat{r}$ の方は、

$$\begin{aligned} \hat{r}(x', y') &= \frac{\sqrt{(S_{x'x'} - S_{y'y'})^2 + 4S_{x'y'}^2}}{S_{x'x'} + S_{y'y'}} = \frac{\sqrt{(A^2 S_{xx} - B^2 S_{yy})^2 + 4A^2 B^2 S_{xy}^2}}{A^2 S_{xx} + B^2 S_{yy}} \\ &= \frac{\sqrt{(S_{xx} - \delta^2 S_{yy})^2 + 4\delta^2 S_{xy}^2}}{S_{xx} + \delta^2 S_{yy}} \quad \left(\delta = \frac{B}{A}\right) \end{aligned}$$

となり、δ が 1 以外のときは明らかに $\hat{r}(x, y)$ とは等しくならない。つまり $\hat{r}$ はこのスケール変換に関しては不変ではないことが分かる。

同様に回帰直線についても同じスケール変換を考えてみると、通常の回帰直線は (x', y') については

$$y' - \overline{y'} = a(x', y')(x' - \overline{x'}) = \frac{S_{x'y'}}{S_{x'x'}}(x' - \overline{x'})$$

であるが、これは

$$B(y - \overline{y}) = \frac{ABS_{xy}}{A^2 S_{xx}} A(x - \overline{x})$$

となるので、(x, y) 座標系では

$$y - \overline{y} = \frac{S_{xy}}{S_{xx}}(x - \overline{x}) = a(x, y)(x - \overline{x})$$

となり (x, y) での回帰直線に一致する。つまり、一見

$$a(x', y') = \frac{S_{x'y'}}{S_{x'x'}} = \frac{B}{A} a(x, y)$$

となり、スケール変換で変わってしまうようにも見えるが、実際は本質的にスケール変換不変であることが分かる。

ところが、新たに考えた $\hat{a}$ を用いた回帰直線の方は、

$$\begin{aligned}
\hat{a}(x', y') &= \frac{S_{y'y'} - S_{x'x'} + \sqrt{(S_{x'x'} - S_{y'y'})^2 + 4S_{x'y'}^2}}{2S_{x'y'}} \\
&= \frac{B^2 S_{yy} - A^2 S_{xx} + \sqrt{(A^2 S_{xx} - B^2 S_{yy})^2 + 4A^2 B^2 S_{xy}^2}}{2ABS_{xy}} \\
&= \frac{\delta^2 S_{yy} - S_{xx} + \sqrt{(S_{xx} - \delta^2 S_{yy})^2 + 4\delta^2 S_{xy}^2}}{2\delta S_{xy}}
\end{aligned}$$

となり、これもやはり $\delta\hat{a}(x, y) = B\hat{a}(x, y)/A$ には一致せず、本質的にこのスケール変換で変わってしまうことになる。

7 最後に

4, 6 節等で調べた不変性と $r, \hat{r}, a, \hat{a}$ との関係を表にまとめると表 1 のようになる。

	r	$\hat{r}$	a	$\hat{a}$
x,y の入れ替え	不変	不変	本質的に変化	本質的に不変
回転	変化	不変	本質的に変化	本質的に不変
スケール変換	不変	変化	本質的に不変	本質的に変化

表 1: データのスケール変換や回転等に関する不変性

例えば x と y が身長と体重のように全く異なる種類のデータの場合、各軸の単位の取り方は任意であるため、各軸毎のスケール変換に関する不変性は、指標としては必須の条件となる。$\hat{a}, \hat{r}$ がその性質を満たさないということは、これらは異種のデータには弱い、あるいは全く使えない、ということを意味している。

元々回帰直線は、y 方向に誤差を計るということからもわかるように、通常の回帰直線は x を変数とみて、y をそれによる関数とみる、という関係を強く意識していて、よってそれぞれが同種のデータである必要はない。そういう場合には通常の回帰直線、通常の相関係数を使うべきであろうし、それで普段は通常の回帰直線が用いられているのだろうと思う。これが問題 3 の後半部分の回答になると思う。

ただし、単位が同じ同種のデータの直線相関性を調べる場合は、新たに提案した相関係数、回帰直線も 5 節で述べたようにそれなりの性質を持つ。それぞれの優位性を知り、うまく使い分けると良いのではないかと思う。

なお、5 節で提案した新たな回帰直線は、多変量解析で主成分分析と呼ばれるものに対応しているようである。主成分分析については、また機会があればまとめたいと思うが、詳しくは多変量解析の専門書を参照されたい。