2019 年 06 月 04 日

回帰直線との差の分布について

新潟工科大学 基礎教育・教養系 竹野茂治

1 はじめに

1 回目と 2 回目のテストの点数のような 2 次元データ (x_j, y_j) $(j = 1, 2, \ldots, N)$ の回帰直線は、学生の成績の伸びを見るのに使われたりする。

その際、1 回目の点数と、成績の伸びの関係を見るためにその散布図を書いてみると、かなり相関のなさそうな図ができる。本稿では、その相関について考えてみる。

2 計算式

各データに対する計算式をまずあげておく。

x_j の平均 $\bar{x}$, y_j の平均 $\bar{y}$ は

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j, \quad \bar{y} = \frac{1}{N}\sum_{j=1}^{N} y_j \tag{1}$$

平方和 S_{xx}, S_{yy} と積和 S_{xy} は以下の通り。

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

これらは、展開によって以下のようにも書ける。

$$S_{xx} = \sum_{j=1}^{N} x_j^2 - 2\bar{x}\sum_{j=1}^{N} x_j + N(\bar{x})^2 = \sum_{j=1}^{N} x_j^2 - N(\bar{x})^2, \tag{3}$$

$$S_{yy} = \sum_{j=1}^{N} y_j^2 - N(\bar{y})^2, \tag{4}$$

$$S_{xy} = \sum_{j=1}^{N} x_j y_j - \bar{x}\sum_{j=1}^{N} y_j - \bar{y}\sum_{j=1}^{N} x_j + N\bar{x}\bar{y} = \sum_{j=1}^{N} x_j y_j - N\bar{x}\bar{y} \tag{5}$$

x と y の回帰直線は、

$$y = \alpha_{xy}(x - \bar{x}) + \bar{y}, \quad \alpha_{xy} = \frac{S_{xy}}{S_{xx}} \tag{6}$$

で得られる y 方向の最小自乗直線で、相関係数

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \tag{7}$$

が ± 1 に近いほど回帰直線の近くに分布することが知られている。

3 伸び

x_j, y_j が 1 回目と 2 回目のテストの点数のように、同種のデータである場合は、その差 $z_j = y_j - x_j$ を値の「伸び」として考えることができる。これが x_j の値と相関があるのか、すなわち x_j が大きいほど伸びは大きくなるのか、または x_j がむしろ小さい方が伸びは大きくなるのか、などを調べたくなることもまた自然であろう。

伸びとしては、$z_j = y_j - x_j$ 以外に、y_j と x_j での回帰直線の値との差

$$w_j = y_j - (\alpha_{xy}(x_j - \bar{x}) + \bar{y}) \tag{8}$$

を考えることもできる。回帰直線の値は、x_j に対する平均的な y の値、期待される y の値を意味し、w_j はそれとの差であり、よって全体のデータから決まる相対的な伸びを意味することになる。

x_j と y_j の単位が違う場合には z_j のような差よりもむしろ w_j の方が伸びとしては適切だろうし、また w_j はスケール変換にも強い。例えば、$x'_j = px_j, y'_j = qy_j$ とすると、

$$z'_j = y'_j - x'_j = qy_j - px_j$$

より z_j の分布とはかなり変わってしまう可能性があるが、w'_j の方は、

$$\bar{x'} = p\bar{x},\ \ \bar{y'} = q\bar{y},\ \ S_{x'x'} = p^2 S_{xx},\ \ S_{y'y'} = q^2 S_{yy},\ \ S_{x'y'} = pqS_{xy}$$

より、

$$\alpha_{x'y'} = \frac{S_{x'y'}}{S_{x'x'}} = \frac{pqS_{xy}}{p^2 S_{xx}} = \frac{q}{p}\alpha_{xy}$$

となり、x', y' の回帰直線は、

$$y' = qy = \alpha_{x'y'}(x' - \bar{x'}) + \bar{y'} = \frac{q}{p}\alpha_{xy}(px - p\bar{x}) + q\bar{y} = q(\alpha_{xy}(x - \bar{x}) + \bar{y})$$

となり、実質的に (6) と同じものになり、

$$w'_j = y'_j - (\alpha_{x'y'}(x'_j - \bar{x'}) + \bar{y'}) = qy_j - \left(\frac{q}{p}\alpha_{xy}(px_j - p\bar{x}) + q\bar{y}\right) = qw_j$$

となって、w_j の分布を q 倍しただけなので、実質的に分布は変わらず、スケール変換に影響されないことがわかる。

4 伸びとの相関

本節で、x と伸びとの相関を調べてみる。まずは x と z から。

$$\bar{z} = \frac{1}{N}\sum_{j=1}^{N} z_j = \frac{1}{N}\sum_{j=1}^{N}(y_j - x_j) = \bar{y} - \bar{x}$$

より、

$$\begin{aligned} S_{xz} &= \sum_{j=1}^{N} x_j z_j - N\bar{x}\bar{z} = \sum_{j=1}^{N} x_j y_j - \sum_{j=1}^{N} x_j^2 - N\bar{x}\bar{y} + N(\bar{x})^2 \\ &= S_{xy} - S_{xx}, \end{aligned} \tag{9}$$

$$\begin{aligned} S_{zz} &= \sum_{j=1}^{N} z_j^2 - N(\bar{z})^2 = \sum_{j=1}^{N} (y_j - x_j)^2 - N(\bar{y} - \bar{x})^2 \\ &= \sum_{j=1}^{N} y_j^2 - 2\sum_{j=1}^{N} y_j x_j + \sum_{j=1}^{N} x_j^2 - N(\bar{y})^2 + 2N\bar{y}\bar{x} - N(\bar{x})^2 \\ &= S_{yy} - 2S_{xy} + S_{xx} \end{aligned} \tag{10}$$

となる。よって、x と z の相関係数 r_{xz} は

$$r_{xz} = \frac{S_{xz}}{\sqrt{S_{xx}S_{zz}}} = \frac{S_{xy} - S_{xx}}{\sqrt{S_{xx}(S_{xx} - 2S_{xy} + S_{yy})}} \tag{11}$$

となるので、x と z の相関は必ずしも 0 になるわけではなく、相関が 0 になるのは $S_{xy} = S_{xx}$ のとき、すなわち x と y の回帰直線 (6) の傾き α_{xy} が 1 のとき、となる。元々の回帰直線の傾きが 1 に近ければ x と z との相関は小さくなるが、一般にはそうとも限らない。

次は x と w の相関を考える。

$$w_j = (y_j - \bar{y}) - \alpha_{xy}(x_j - \bar{x}) \tag{12}$$

より、

$$\bar{w} = \sum_{j=1}^{N} (y_j - \bar{y}) - \alpha_{xy} \sum_{j=1}^{N} (x_j - \bar{x}) = 0$$

すなわち w の平均は 0 となる。よって、

$$\begin{aligned} S_{xw} &= \sum_{j=1}^{N} x_j w_j - N\bar{x}\bar{w} = \sum_{j=1}^{N} \{x_j(y_j - \bar{y}) - \alpha_{xy} x_j (x_j - \bar{x})\} \\ &= \sum_{j=1}^{N} x_j y_j - \bar{y}\sum_{j=1}^{N} x_j - \alpha_{xy}\left(\sum_{j=1}^{N} x_j^2 - \bar{x}\sum_{j=1}^{N} x_j\right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{N} x_j y_j - N\bar{x}\bar{y} - \alpha_{xy}\left(\sum_{j=1}^{N} x_j^2 - N(\bar{x})^2\right) \;=\; S_{xy} - \alpha_{xy} S_{xx} \\
&= S_{xy} - \frac{S_{xy}}{S_{xx}} S_{xx} \;=\; 0
\end{aligned}
$$

となり、x, w の積和は 0 となる。一応 S_{ww} も計算してみると、

$$
\begin{aligned}
S_{ww} &= \sum_{j=1}^{N} (w_j - \bar{w})^2 \;=\; \sum_{j=1}^{N} \{(y_j - \bar{y}) - \alpha_{xy}(x_j - \bar{x})\}^2 \\
&= \sum_{j=1}^{N} (y_j - \bar{y})^2 - 2\alpha_{xy} \sum_{j=1}^{N} (y_j - \bar{y})(x_j - \bar{x}) + \alpha_{xy}^2 \sum_{j=1}^{N} (x_j - \bar{x})^2 \\
&= S_{yy} - 2\alpha_{xy} S_{xy} + \alpha_{xy}^2 S_{xx} \;=\; S_{yy} - 2\frac{S_{xy}^2}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}} \;=\; S_{yy} - \frac{S_{xy}^2}{S_{xx}} \\
&= S_{yy}(1 - r_{xy}^2)
\end{aligned}
$$

となるので、x と y が完全に直線相関 ($r_{xy} = \pm 1$) でなければ $S_{ww} > 0$ であり、x, w の相関は

$$
r_{xw} = \frac{S_{xw}}{\sqrt{S_{xx} S_{ww}}} = 0
$$

すなわち、相関は常に 0 であることがわかる。

5 最後に

ふと、記録の伸びなどの相関を調べてみて気がついたことをまとめたが、易しい計算で導かれるものなので、多分良く知られていることだと思う。

なお、x と w は相関が完全に 0 であることが示されたが、これは、完全に 0 の相関を持つような分布の例を作る方法として使えるかもしれない。