2変数統計データの線形近似

Theorem. n 個の 2 次元データ $(x_1, y_1), \dots, (x_n, y_n)$ について, 最小二乗法を用いて $y = ax+b$ $(a, b \in \mathbb{R})$ で近似したとき

$$a = \frac{s_{xy}}{s_x^2} = r_{xy} \cdot \frac{s_y}{s_x}, b = \overline{y} - a\overline{x}$$

をみたす. ただし, $\overline{x}$ は x の平均, s_x は x の標準偏差, s_{xy} は x, y の共分散, r_{xy} は x, y の相関係数を表すとする.

Proof. 2 乗誤差関数 E を

$$E(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

とし, E を最小にする a, b を x, y を用いて表せばよい. E を a, b で偏微分し, $= 0$ として連立すると

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

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

となる. (2) 式の両辺を $-2n$ で割ることにより $\overline{y} = a\overline{x} + b$ が得られる. (1) 式も両辺 $-2n$ で割ることにより

$$\frac{1}{n} \sum_{i=1}^{n} x_i y_i - a \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2 - b\overline{x} = 0$$

が得られる. ここで, $s_x^2 = \overline{x^2} - \overline{x}^2, \overline{y} = a\overline{x} + b$ を代入すると

$$\frac{1}{n} \sum_{i=1}^{n} x_i y_i - a(s_x^2 + \overline{x}^2) - b\overline{x} = \frac{1}{n} \sum_{i=1} x_i y_i - a s_x^2 - \overline{x}(a\overline{x} + b) = \frac{1}{n} \sum_{i=1} x_i y_i - a s_x^2 - \overline{x} \cdot \overline{y} = 0$$

となるから, $a = \frac{1}{s_x^2} \left(\frac{1}{n} \sum_{i=1}^{n} x_i y_i - \overline{x} \cdot \overline{y} \right)$ となる。ここで, もう少し式変形を進めると

$$\begin{aligned}
a &= \frac{1}{s_x^2} \left(\frac{1}{n} \sum_{i=1}^{n} x_i y_i - \overline{x} \cdot \overline{y} \right) \\
&= \frac{1}{s_x^2} \left(\frac{1}{n} \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} \overline{x} y_i \right) \\
&= \frac{1}{s_x^2} \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}) y_i \right) \\
&= \frac{1}{s_x^2} \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \right) + \frac{1}{s_x^2} \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}) \overline{y} \right) \\
&= \frac{1}{s_x^2} \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \right) \\
&= \frac{s_{xy}}{s_x^2}
\end{aligned}$$

となる. ここで, $r_{xy} = \frac{s_{xy}}{s_x s_y}$ より, $s_{xy} = r_{xy} s_x s_y$ を代入すると, $a = r_{xy} \cdot \frac{s_y}{s_x}$ となる. ■