

§1.5 回帰直線を表す方程式

変量 X の値と変量 Y の値との順序対を要素とする 2 次元のデータについて、相関係数の絶対値が 1 に近いとき、散布図の点は一つ直線つまり 1 次関数のグラフの近くに分布します。つまり、データの各要素の Y の値が X の値の 1 次関数の値に近いこととなります。その 1 次関数のグラフである直線を Y の X への回帰直線といいます。回帰直線を表す方程式を求める公式は次のようになります。

公式 変量 X の値と変量 Y の値との順序対を要素とする 2 次元のデータについて、 X の平均を \bar{x} とおき、 Y の平均を \bar{y} とおき、 X の分散を v_X とおき、 X と Y との共分散を c_{XY} とおくと、 Y の X への回帰直線は xy 座標平面において方程式 $y = \frac{c_{XY}}{v_X}(x - \bar{x}) + \bar{y}$ で表される。

この公式を導きます。正の自然数 n に対して、変量 X の値と変量 Y の値との n 個の順序対

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), \dots, (x_n, y_n)$$

を要素とする 2 次元のデータがあるとします。各要素の Y の値が X の値の 1 次関数の値に近いとします。定数 a, b に対して、変数 x の 1 次関数 $ax + b$ を考えます。 $k = 1, 2, 3, \dots, n$ に対して、 Y の値 y_k が X の値 x_k の 1 次関数の値 $ax_k + b$ に近くなるようにします。そこで、 $ax_k + b$ と y_k との差の 2 乗の総和 $\sum_{k=1}^n (ax_k + b - y_k)^2$ の値が最小になるように a 及び b の値を定めます。 X の平均 \bar{x} 及び Y の平均 \bar{y} は、

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k, \quad \bar{y} = \frac{1}{n} \sum_{k=1}^n y_k,$$

これより、

$$\sum_{k=1}^n x_k = n\bar{x}, \quad \sum_{k=1}^n y_k = n\bar{y}.$$

X の分散 v_X 及び Y の分散 v_Y は、定理 1.3 より、

$$v_X = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 = \frac{1}{n} \sum_{k=1}^n x_k^2 - \bar{x}^2, \quad v_Y = \frac{1}{n} \sum_{k=1}^n (y_k - \bar{y})^2 = \frac{1}{n} \sum_{k=1}^n y_k^2 - \bar{y}^2,$$

これより

$$\sum_{k=1}^n x_k^2 = n(v_X + \bar{x}^2), \quad \sum_{k=1}^n y_k^2 = n(v_Y + \bar{y}^2).$$

X と Y との共分散 c_{XY} は、定理 1.4.1 より、

$$c_{XY} = \frac{1}{n} \sum_{k=1}^n \{(x_k - \bar{x})(y_k - \bar{y})\} = \frac{1}{n} \sum_{k=1}^n (x_k y_k) - \bar{x}\bar{y},$$

これより、

$$\sum_{k=1}^n (x_k y_k) = n(c_{XY} + \bar{x}\bar{y}).$$

これらの等式より、

$$\begin{aligned} & \sum_{k=1}^n (ax_k + b - y_k)^2 \\ &= \sum_{k=1}^n (a^2 x_k^2 + b^2 + y_k^2 + 2abx_k - 2ax_k y_k - 2by_k) \\ &= a^2 \sum_{k=1}^n x_k^2 + \sum_{k=1}^n b^2 + \sum_{k=1}^n y_k^2 + 2ab \sum_{k=1}^n x_k - 2a \sum_{k=1}^n (x_k y_k) - 2b \sum_{k=1}^n y_k \\ &= a^2 n(v_X + \bar{x}^2) + nb^2 + n(v_Y + \bar{y}^2) + 2abn\bar{x} - 2an(c_{XY} + \bar{x}\bar{y}) - 2bn\bar{y}. \end{aligned}$$

まず b について平方完成します：

$$\begin{aligned} & a^2 n(v_X + \bar{x}^2) + nb^2 + n(v_Y + \bar{y}^2) + 2abn\bar{x} - 2an(c_{XY} + \bar{x}\bar{y}) - 2bn\bar{y} \\ &= nb^2 + 2nb(a\bar{x} - \bar{y}) + na^2(v_X + \bar{x}^2) - 2na(c_{XY} + \bar{x}\bar{y}) + n(v_Y + \bar{y}^2) \\ &= n\{b^2 + 2b(a\bar{x} - \bar{y}) + (a\bar{x} - \bar{y})^2 - (a\bar{x} - \bar{y})^2\} + \\ & \quad na^2(v_X + \bar{x}^2) - 2na(c_{XY} + \bar{x}\bar{y}) + n(v_Y + \bar{y}^2) \\ &= n(b + a\bar{x} - \bar{y})^2 - n(a\bar{x} - \bar{y})^2 + na^2(v_X + \bar{x}^2) - 2na(c_{XY} + \bar{x}\bar{y}) + n(v_Y + \bar{y}^2) \\ &= n(b + a\bar{x} - \bar{y})^2 - n(a^2\bar{x}^2 - 2a\bar{x}\bar{y} + \bar{y}^2) \\ & \quad + na^2(v_X + \bar{x}^2) - 2na(c_{XY} + \bar{x}\bar{y}) + n(v_Y + \bar{y}^2) \\ &= n(b + a\bar{x} - \bar{y})^2 + na^2 v_X - 2nac_{XY} + nv_Y. \end{aligned}$$

$na^2 v_X - 2nac_{XY}$ の部分を a について平方完成します：

$$\begin{aligned} & n(b + a\bar{x} - \bar{y})^2 + na^2 v_X - 2nac_{XY} + nv_Y \\ &= n(b + a\bar{x} - \bar{y})^2 + nv_X \left\{ a^2 - 2a \frac{c_{XY}}{v_X} + \left(\frac{c_{XY}}{v_X} \right)^2 - \frac{c_{XY}^2}{v_X^2} \right\} + nv_Y \\ &= n(b + a\bar{x} - \bar{y})^2 + nv_X \left(a - \frac{c_{XY}}{v_X} \right)^2 - \frac{nc_{XY}^2}{v_X} + nv_Y. \end{aligned}$$

これらの等式より、

$$\sum_{k=1}^n (ax_k + b - y_k)^2 = n(b + a\bar{x} - \bar{y})^2 + nv_X \left(a - \frac{c_{XY}}{v_X} \right)^2 - \frac{nc_{XY}^2}{v_X} + nv_Y.$$

この式の値が最小になるのは、 $b + a\bar{x} - \bar{y} = 0$ かつ $a - \frac{c_{XY}}{v_X} = 0$ のとき、つまり $a = \frac{c_{XY}}{v_X}$ かつ $b = \bar{y} - a\bar{x}$ のときです；このとき、

$$ax + b = ax + \bar{y} - a\bar{x} = a(x - \bar{x}) + \bar{y} = \frac{c_{XY}}{v_X}(x - \bar{x}) + \bar{y}.$$

Y の X への回帰直線は、 xy 座標平面において方程式 $y = \frac{c_{XY}}{v_X}(x - \bar{x}) + \bar{y}$ で表されます。

例題 1.5 前節の例題 1.4 のように、4 人の学生 A 君、B 君、C 君、D 君のある科目の中間試験の点数 X と期末試験の点数 Y とについて以下のようになった：

学生	A	B	C	D
中間試験の点数 X	75	67	81	73
期末試験の点数 Y	67	71	83	71

xy 座標平面において、 Y の X への回帰直線を表す方程式を求める。

X の平均は $\bar{x} = 74$ であり、 X の分散は $v_X = 25$ であり、 Y の平均は $\bar{y} = 73$ であり、 X と Y との共分散は $c_{XY} = 20$ である。 Y の X への回帰直線を表す方程式は、 $y = \frac{c_{XY}}{v_X}(x - \bar{x}) + \bar{y}$ なので、 $y = \frac{20}{25}(x - 74) + 73$ 、整理すると $y = \frac{4x + 69}{5}$ 。 □

問題 1.5 4 人の学生 A 君、B 君、C 君、D 君の身長（単位は cm） X と体重（単位は kg） Y とについて調べると次のようになりました。

学生	A	B	C	D
身長（単位は cm） X	158	162	174	162
体重（単位は kg） Y	61	53	71	55

xy 座標平面において、 Y の X への回帰直線を表す方程式を求めなさい。