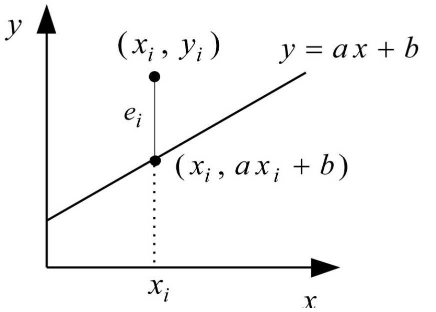
● **回帰直線の定義**

変量 x と y から得られた 2 次元データ (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n) に対して，a，b の関数

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

が最小になるように a，b の値を定める。

このとき，直線 $y = ax + b$ を「y の x への回帰直線」，傾き a を「y の x への回帰係数」という。



y の x への回帰直線

$e_i = y_i - (ax_i + b)$

距離の2乗の和が $\sum_i e_i^2$ が最小

方針　$e_i = y_i - (ax_i + b)$ として a , b の関数 $f(a,b) = \sum_{i=1}^{n} e_i^2$ が最小になるような a , b の値を求めます。

$e_i = y_i - (ax_i + b)$　の右辺に $-\overline{y} + \overline{y}$ と $-a\overline{x} + a\overline{x}$ を加えます。

$$= y_i - ax_i - b - \overline{y} + \overline{y} - a\overline{x} + a\overline{x}$$

$$= (y_i - \overline{y}) - a(x_i - \overline{x}) + \underbrace{\overline{y} - a\overline{x} - b}_{c\text{と置きます}}$$

技巧的な変形を私は
『しばいてさする』
と表現することがあります。

3項の２乗の展開公式を使います

$$e_i^2 = (y_i - \overline{y})^2 + a^2(x_i - \overline{x})^2 + c^2 - 2a(x_i - \overline{x})(y_i - \overline{y}) - 2ac(x_i - \overline{x}) + 2c(y_i - \overline{y})$$

i を 1 から n まで動かして和を考えます。（Σは簡略表現します。）

$$\sum e_i^2 = \sum (y_i - \overline{y})^2 + a^2 \sum (x_i - \overline{x})^2 + nc^2 - 2a \sum (x_i - \overline{x})(y_i - \overline{y}) \underbrace{- 2ac \sum (x_i - \overline{x}) + 2c \sum (y_i - \overline{y})}_{\text{この2項は0です。偏差の合計は0になるからです。}}$$

分散，共分散の定義に従うと

$$= n \cdot s_y^2 + na^2 \cdot s_x^2 + nc^2 - 2an \cdot s_{xy}$$

$r = \dfrac{s_{xy}}{s_x \cdot s_y}$ より

$s_{xy} = rs_x \cdot s_y$ と置き換えると

$$\frac{1}{n} \sum e_i^2 = a^2 \cdot s_x^2 - 2ars_x \cdot s_y + s_y^2 + c^2$$

$$= (as_x - rs_y)^2 - r^2 s_y^2 + s_y^2 + c^2$$

$$= (as_x - rs_y)^2 + (1 - r^2) s_y^2 + c^2$$

平方完成
「しばいてさする」

これにより，残差の２乗の平均は $\begin{cases} as_x - rs_y = 0 \\ c = \overline{y} - a\overline{x} - b = 0 \end{cases}$ のとき最小となります。

つまり $\begin{cases} a = \dfrac{rs_y}{s_x} = \dfrac{s_y}{s_x} \cdot \dfrac{s_{xy}}{s_x \cdot s_y} = \dfrac{s_{xy}}{s_x^2} \\ \overline{y} = a\overline{x} + b \end{cases}$ となります。　終