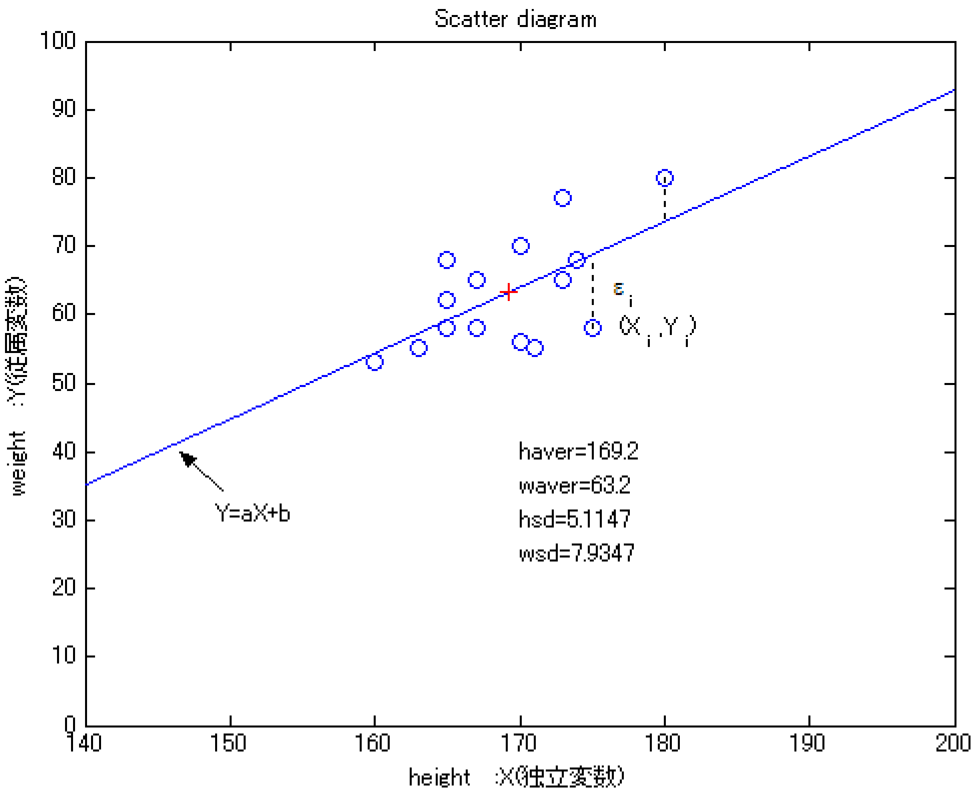
回帰直線式の誘導

横軸 X を独立変数, 縦軸 Y を従属変数とし, X に誤差は無く Y に誤差があると仮定する.

→　このとき, データ (X_i, Y_i) とそれに対応する回帰直線上の値 $(X_i, \hat{Y}_i)$ との差：即ち,

残差について, その平方和 $\sum_{i=1}^{n} \varepsilon_i{}^2 = \sum (Y_i - \hat{Y}_i)^2$ を最小にする様に回帰直線を決定する.

この様に回帰直線式を決める方法を最小二乗法または最小自乗法と呼ぶ.



回帰直線を $\hat{Y} = aX + b$ と置く．データ $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ は, Y_i に誤差 ε_i があるとすると, 次式が得られる.

$$Y_i = aX_i + b + \varepsilon_i \qquad \therefore \quad \varepsilon_i{}^2 = (Y_i - \hat{Y}_i)^2 = (Y_i - b - aX_i)^2$$

ここで, $S = \sum_{i=1}^{n} \varepsilon_i{}^2 = \sum_{i=1}^{n} (Y_i - b - aX_i)^2$ と置く.

残差平方和を最小にするための条件は,

$$\frac{\partial S}{\partial b} = \sum_{i=1}^{n} \left\{ -2(Y_i - b - aX_i) \right\} = 0$$

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{n} \left\{ -2X_i(Y_i - b - aX_i) \right\} = 0$$

上式を整理すると,

$$nb + a\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \qquad \sim(1)$$

$$b\sum_{i=1}^{n} X_i + a\sum_{i=1}^{n} {X_i}^2 = \sum_{i=1}^{n} X_i Y_i \qquad \sim(2)$$

(1)式より， $nb + an\overline{X} = n\overline{Y} \quad \therefore \quad b = \overline{Y} - a\overline{X}$ 〜(3)

(3)式を(2)式に代入すると， $(\overline{Y} - a\overline{X})\sum_{i=1}^{n} X_i + a\sum_{i=1}^{n} {X_i}^2 = \sum_{i=1}^{n} X_i Y_i$

$$\therefore \quad \left(\sum_{i=1}^{n} {X_i}^2 - \overline{X}\sum_{i=1}^{n} X_i\right)a = \sum_{i=1}^{n} X_i Y_i - \overline{Y}\sum_{i=1}^{n} X_i$$

$$\therefore \quad a = \frac{\sum_{i=1}^{n} X_i Y_i - \overline{Y}\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} {X_i}^2 - \overline{X}\sum_{i=1}^{n} X_i} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{Y}\cdot\overline{X}}{\frac{1}{n}\sum_{i=1}^{n} {X_i}^2 - \overline{X}\cdot\overline{X}} = \frac{Cov}{{S_x}^2} \qquad \sim(4)$$

ただし， Cov：共分散 covariance, S_x：x の標準偏差値 standard deviation

∵ 共分散の定義式から

$$Cov = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}(X_i Y_i - X_i\overline{Y} - \overline{X}Y_i + \overline{X}\overline{Y})$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\sum_{i=1}^{n} X_i \cdot \overline{Y} - \overline{X}\frac{1}{n}\sum_{i=1}^{n} Y_i + \overline{X}\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\overline{Y} - \overline{X}\overline{Y} + \overline{X}\overline{Y}$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\overline{Y}$$

標準偏差値の定義式から

$${S_x}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}({X_i}^2 - 2X_i\overline{X} + \overline{X}^2) = \frac{1}{n}\sum_{i=1}^{n} {X_i}^2 - 2\frac{1}{n}\sum_{i=1}^{n} X_i \cdot \overline{X} + \frac{1}{n}\cdot n \cdot \overline{X}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} {X_i}^2 - 2\overline{X}^2 + \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n} {X_i}^2 - \overline{X}^2$$

(4)式を更に相関係数(correlation coefficient)：r の定義式 $r = \frac{Cov}{S_x S_y}$ を用いて変形すると，

$$a = \frac{Cov}{S_x S_y}\cdot\frac{S_y}{S_x} = r\cdot\frac{S_y}{S_x} \qquad \sim(5)$$

なお，(3)式を回帰直線式：$\hat{Y} = aX + b$ に代入すると，

$$\hat{Y} = aX + (\overline{Y} - a\overline{X}) \quad \therefore \quad (\hat{Y} - \overline{Y}) = a(X - \overline{X}) \qquad \sim(6)$$

となり，回帰直線式は平均値 $(\overline{X}, \overline{Y})$ を通る傾き a の直線となることが分かる．
また次の式展開より，相関係数 r のとり得る範囲は，－１〜１となることが分かる．

$$\frac{S}{n} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \left\{\overline{Y} + r\frac{S_y}{S_x}(X_i - \overline{X})\right\}\right]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 - 2r\frac{S_y}{S_x}\cdot\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) + r^2\frac{{S_y}^2}{{S_x}^2}\cdot\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$= {S_y}^2 - 2r\frac{S_y}{S_x}\cdot r\cdot S_x S_y + r^2\frac{{S_y}^2}{{S_x}^2}\cdot {S_x}^2 = {S_y}^2(1-r^2)$$

$$\therefore\quad \frac{S}{n} = {S_y}^2(1-r^2)$$

ここで，$S > 0$，$n > 0$，${S_y}^2 > 0$ なので，$(1-r^2) \geq 0$ であらねばならない.

$$\therefore\quad -1 \leq r \leq 1 \qquad \sim(7)$$