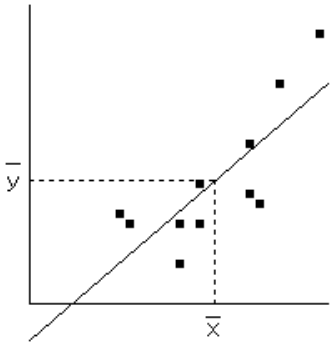
回帰直線

2 つのデータ $X = \{x_1, x_2, \ldots, x_n\}$，$Y = \{y_1, y_2, \ldots, y_n\}$ を考えます．xy 平面に点 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ を描いた図を**散布図**といいます．2 つのデータの相関が強いとき，散布図を見ると，ある直線の近くに点が集まっています．このような直線を**回帰直線**と言います．



それぞれのデータの，平均を $\overline{x}\,, \overline{y}$，標準偏差を σ_x, σ_y，2 つのデータの，共分散を σ_{xy}，相関係数を r とします．これらは左の式で定義され，右の性質を満たします．

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \sum_{k=1}^{n}(x_k - \overline{x}\,) = 0$$

$$\sigma_x^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x}\,)^2 \qquad \sigma_x^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \overline{x}^{\,2}$$

$$\sigma_{xy} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x}\,)(y_k - \overline{y}\,) \qquad \sigma_{xy} = \frac{1}{n}\sum_{k=1}^{n} x_k y_k - \overline{x}\,\overline{y}$$

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad \sigma_{xy} = \sigma_x \sigma_y\, r$$

上の性質の証明

$$\sum_{k=1}^{n}(x_k - \overline{x}\,) = \sum_{k=1}^{n} x_k - n\overline{x} = n\overline{x} - n\overline{x} = 0$$

$$\sigma_x^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x}\,)^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k^2 - 2\overline{x}\,x_k + \overline{x}^{\,2}) = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \frac{2}{n}\overline{x}\sum_{k=1}^{n} x_k + \overline{x}^{\,2}$$
$$= \frac{1}{n}\sum_{k=1}^{n} x_k^2 - 2\overline{x}^{\,2} + \overline{x}^{\,2} = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \overline{x}^{\,2}$$

$$\sigma_{xy} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x}\,)(y_k - \overline{y}\,) = \frac{1}{n}\sum_{k=1}^{n}(x_k y_k - x_k\overline{y} - \overline{x}\,y_k + \overline{x}\;\overline{y}\,)$$
$$= \frac{1}{n}\sum_{k=1}^{n} x_k y_k - \frac{1}{n}\sum_{k=1}^{n} x_k\overline{y} - \frac{1}{n}\overline{x}\sum_{k=1}^{n} y_k + \overline{x}\;\overline{y} = \frac{1}{n}\sum_{k=1}^{n} x_k y_k - \overline{x}\;\overline{y} - \overline{x}\;\overline{y} + \overline{x}\;\overline{y}$$
$$= \frac{1}{n}\sum_{k=1}^{n} x_k y_k - \overline{x}\;\overline{y}$$

最小２乗直線

直線 $y = f(x) = mx + n$ について，各 x_k に対する誤差（$f(x_k)$ と実際の値 y_k との差）$f(x_k) - y_k$ の自乗の総和 $E = \sum_{k=1}^{n} (f(x_k) - y_k)^2$ を考えます．これが最小になるような直線 $y = f(x)$ が回帰直線の 1 つで，**最小２乗直線**といいます．

回帰直線は，中心 $(\overline{x}\,, \overline{y}\,)$ を通る，すなわち，

$$y = f(x) = A(x - \overline{x}\,) + \overline{y}$$

と表されると推測されますが，推測が正しいとは限らないので

$$y = f(x) = A(x - \overline{x}\,) + \overline{y}\; + B$$

とおいて，A, B を求めることにします．

$$\begin{aligned}
E &= \sum_{k=1}^{n} (f(x_k) - y_k)^2 \\
&= \sum_{k=1}^{n} (A(x_k - \overline{x}\,) - (y_k - \overline{y}\,) + B)^2 \\
&= \sum_{k=1}^{n} (A^2(x_k - \overline{x}\,)^2 - 2A(x_k - \overline{x}\,)(y_k - \overline{y}\,) + (y_k - \overline{y}\,)^2 + 2AB(x_k - \overline{x}\,) - 2B(y_k - \overline{y}\,) + B^2) \\
&= n\sigma_x^2 A^2 - 2n\sigma_{xy} A + n\sigma_y^2 + 0 - 0 + B^2 \\
&= n\sigma_x^2 A^2 - 2n\sigma_x \sigma_y\; rA + n\sigma_y^2 + B^2 \\
&= n(\sigma_x A - \sigma_y\; r)^2 + n\sigma_y^2 (1 - r^2) + B^2
\end{aligned}$$

ゆえに，E は，$A = \dfrac{\sigma_y}{\sigma_x} r$，$B = 0$ のときに最小になります．したがって，最小 2 乗直線は

$$y = \frac{\sigma_y}{\sigma_x}\; r(x - \overline{x}\,) + \overline{y} \tag{1}$$

です．相関係数 r が ± 1 のときは，$E = 0$ すなわち，すべての点 (x_k, y_k) がこの直線上にあります．

回帰直線は，これだけではありません．今求めた直線は，x を独立変数，y を従属変数としていますが，逆に y を独立変数とすると，

$$\begin{aligned}
x &= \frac{\sigma_x}{\sigma_y}\; r(y - \overline{y}\,) + \overline{x} \\
y &= \frac{\sigma_y}{\sigma_x} \frac{1}{r} (x - \overline{x}\,) + \overline{y}
\end{aligned} \tag{2}$$

となります．(1) と (2) の傾きの相乗平均をとると

$$\begin{aligned}
y &= \frac{\sigma_y}{\sigma_x} (x - \overline{x}\,) + \overline{y} \quad （r > 0 \text{ すなわち，正の相関があるとき}） \\
y &= -\frac{\sigma_y}{\sigma_x} (x - \overline{x}\,) + \overline{y} \quad （r < 0 \text{ すなわち，負の相関があるとき}）
\end{aligned}$$

となります．