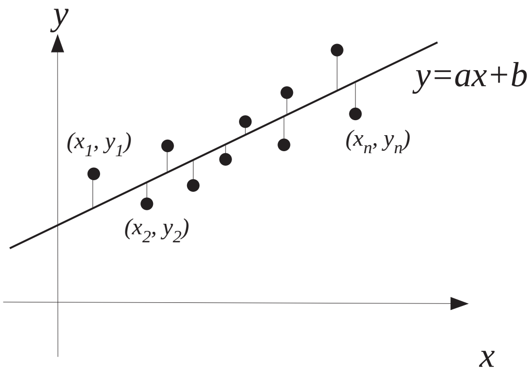
回帰分析と相関係数

1　回帰分析

下図のように (x, y) のデータをプロットしたとき，x, y に比例関係が成り立ちそうな場合がある．このデータにぴったり当てはまる直線の式 $y = ax + b$ を決定するのに最小二乗法が適用できる．このような，二つの変数の間の関係式を求める分析を**回帰分析**と呼んでいる．



求める直線の式を $y = ax + b$ とし，n 個のデータを $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$ とすると，各計測値の残差は $ax_i + b - y_i$ と表すことが出来る．最小二乗法を適用するには，この残差の二乗和が最小となる a, b を求めることである．残差の二乗和の関数 Φ は，以下の式で表すことが出来る．

$$\Phi = \sum_{i=1}^{n} (ax_i + b - y_i)^2 \tag{1}$$

Φ の最小値を求めるが，このとき変数が X, Y の二個あるので偏微分により最小値を求める．すなわち，Φ を X と Y とでそれぞれ偏微分し，それが 0 となる X, Y を計算する．

$$\begin{cases} \dfrac{\partial \Phi}{\partial a} = 2\displaystyle\sum_{i=1}^{n} \{x_i(ax_i + b - y_i)\} & = 0 \\ \dfrac{\partial \Phi}{\partial b} = 2\displaystyle\sum_{i=1}^{n} (ax_i + b - y_i) & = 0 \end{cases} \tag{2}$$

これを行列を用いて整理すると，以下の式を得る．なお総和記号 $\sum$ は，ガウスの総和記号 [] で代用した．

$$\begin{pmatrix} [x_i^2] & [x_i] \\ [x_i] & n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} [x_i y_i] \\ [y_i] \end{pmatrix} \tag{3}$$

この連立方程式を解けば，最確値 a, b が求まる．

2　相関係数

回帰分析によって求まった式が，どの程度の正確かを表すのに**相関係数**が一般に利用されている．相関係数は，**分散**と**共分散**を用いて計算できる．x の平均 $(\bar{x})$ 分散 (v_x)，y の平均 $(\bar{y})$ 分散 (v_y)，x, y の共分散 (v_{xy}) は，以下の式で表すことが出来る．

$$v_x = \sum_{i=1}^{n} (\bar{x} - x_i)^2 / n \tag{4}$$

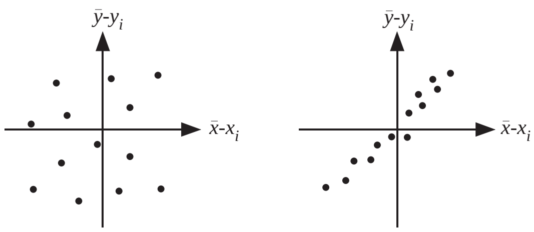
$$v_y = \sum_{i=1}^{n} (\bar{y} - y_i)^2 / n \tag{5}$$

$$v_{xy} = \sum_{i=1}^{n} (\bar{x} - x_i)(\bar{y} - y_i) / n \tag{6}$$

相関係数 (r) は，以下の式で表される．

$$r = \frac{v_{xy}}{\sqrt{v_x}\sqrt{v_y}} \tag{7}$$

相関係数は，$-1 \leqq r \leqq 1$ の範囲で，1 に近いほど正の相関が高く，0 に近いほど無相関，-1 に近いほど負の相関が高いといえる．下図においては，右側のグラフは相関が高く，左のグラフは相関が低い．



共分散が相関係数の値を左右しており，ランダムな点の集まりであれば，$(\bar{x} - x_i)$ の項と $(\bar{y} - y_i)$ の項の符号はランダムに出現し，最終的に共分散の値は小さくなる．逆に共分散の値が大きくなる場合は，$(\bar{x} - x_i)$ の項と $(\bar{y} - y_i)$ の項の符号プラス同士かマイナス同士で同じ場合である．

また共分散の意味は，ベクトルの内積と同じである．$(\bar{x} - x_i)$ の項を a_i，$(\bar{y} - y_i)$ の項を b_i とおき，全てのデータを n 次元のベクトル $\boldsymbol{a}(a_1, a_2, a_3, \cdots, a_n)$ とベクトル $\boldsymbol{b}(b_1, b_2, b_3, \cdots, b_n)$ とみなす．するとそれらの内積は，$a_1b_1 + a_2b_2 + a_3b_3 + \cdots + a_nb_n$ となり，それを n で割ったものが共分散である．一方，分散はベクトルの大きさと意味が同じである．つまり，$\boldsymbol{a}^2 = (a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2)$ となるからである．したがって相関係数 r は，2つのベクトルのなす角度 θ の余弦といえ，その値は $-1 \leqq r \leqq 1$ の範囲となる．

$$r = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}||\boldsymbol{b}|} = \cos\theta \tag{8}$$

したがって，相関係数が高いほど，二つのベクトルの方向が同じであるといえる．