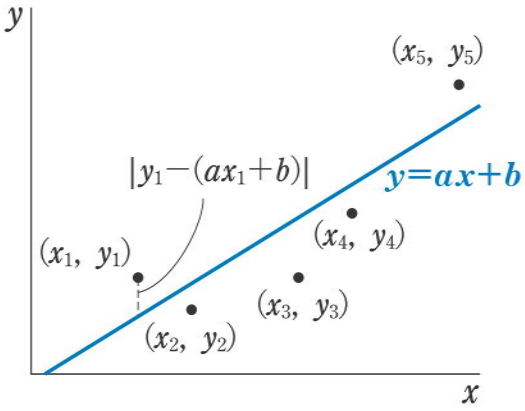
B 最小2乗法

回帰直線は，どのようにして求めることができるのだろうか。簡単のため，大きさ5のデータで考えることにする。

2つの変量 x，y のデータが，次のように与えられているとする。

$$(x_1,\ y_1),\ (x_2,\ y_2),\ (x_3,\ y_3),\ (x_4,\ y_4),\ (x_5,\ y_5)$$

x と y に直線的な相関関係があるとき，散布図の点は回帰直線の近くに分布する。



各点 $(x_k,\ y_k)$ が，$y=ax+b$ で表される直線上にあるとすると $y_k=ax_k+b$ であるが，実際のデータでは，ほとんどの場合 $y_k \neq ax_k+b$ である。そこで，y_k と ax_k+b の差の2乗の和

$$\{y_1-(ax_1+b)\}^2+\{y_2-(ax_2+b)\}^2+\cdots\cdots+\{y_5-(ax_5+b)\}^2 \quad ①$$

が最小となるように a，b を定め，直線 $y=ax+b$ を x，y の関係を近似する直線と考える。この直線 $y=ax+b$ が回帰直線の1つである。

このような回帰直線の求め方を，**最小2乗法** という。

このように，変量 x，y の間の関係をデータから統計的に推測する方法を **回帰分析** という。

2つの変量 x，y のデータの値の組が n 個与えられたとき，最小2乗法による回帰直線 $y=ax+b$ の a，b の値は，次のように求めることができる。

x，y のデータの平均値を $\overline{x}$，$\overline{y}$，分散を ${s_x}^2$，${s_y}^2$，x と y の相関係数を r とするとき　　$a=\dfrac{s_y}{s_x}r$，　$b=\overline{y}-a\overline{x}$　…… ②

練習 23

右の表は，同じ種類の 5 本の木の太さ x (cm) と高さ y (m) を測定した結果である。

木の番号	1	2	3	4	5
x	22	27	29	19	33
y	13	15	18	14	20

(1)　2 つの変量 x，y の散布図をかけ。

(2)　2 つの変量 x，y の回帰直線を表す 1 次関数を求めよ。また，その回帰直線を (1) の散布図に重ねてかけ。

■ 参考

2 つの変量 x，y のデータの値の組が n 個与えられたとき，回帰直線 $y=ax+b$ の a，b の値が上の ② で求められることを確かめてみよう。

回帰直線は点 $(\overline{x}, \overline{y})$ を通る，すなわち $y=a'(x-\overline{x})+\overline{y}$ と表されると推測されるが，推測が正しいとは限らないので

$$y=a'(x-\overline{x})+\overline{y}+b'$$

とおいて，a'，b' を求めることにする。前ページの ① の値は

$$\sum_{k=1}^{n}\{y_k-\{a'(x_k-\overline{x})+\overline{y}+b'\}\}^2=\sum_{k=1}^{n}\{a'(x_k-\overline{x})-(y_k-\overline{y})+b'\}^2$$

$$=\sum_{k=1}^{n}\{{a'}^2(x_k-\overline{x})^2-2a'(x_k-\overline{x})(y_k-\overline{y})+(y_k-\overline{y})^2$$

$$+2a'b'(x_k-\overline{x})-2b'(y_k-\overline{y})+{b'}^2\}$$

$$=n{s_x}^2{a'}^2-2ns_{xy}a'+n{s_y}^2+0-0+n{b'}^2=n{s_x}^2{a'}^2-2ns_xs_yra'+n{s_y}^2+n{b'}^2$$

$$=n(s_xa'-s_yr)^2+n{s_y}^2(1-r^2)+n{b'}^2$$

ゆえに，① の値は，$a'=\dfrac{s_y}{s_x}r$，$b'=0$ のとき最小になる。

よって，a，b の値は $a=\dfrac{s_y}{s_x}r$，$b=\overline{y}-a\overline{x}$ である。

【補足】 記号 $\sum_{k=1}^{n}$ については，第 1 章「数列」の 23～25 ページを参照。