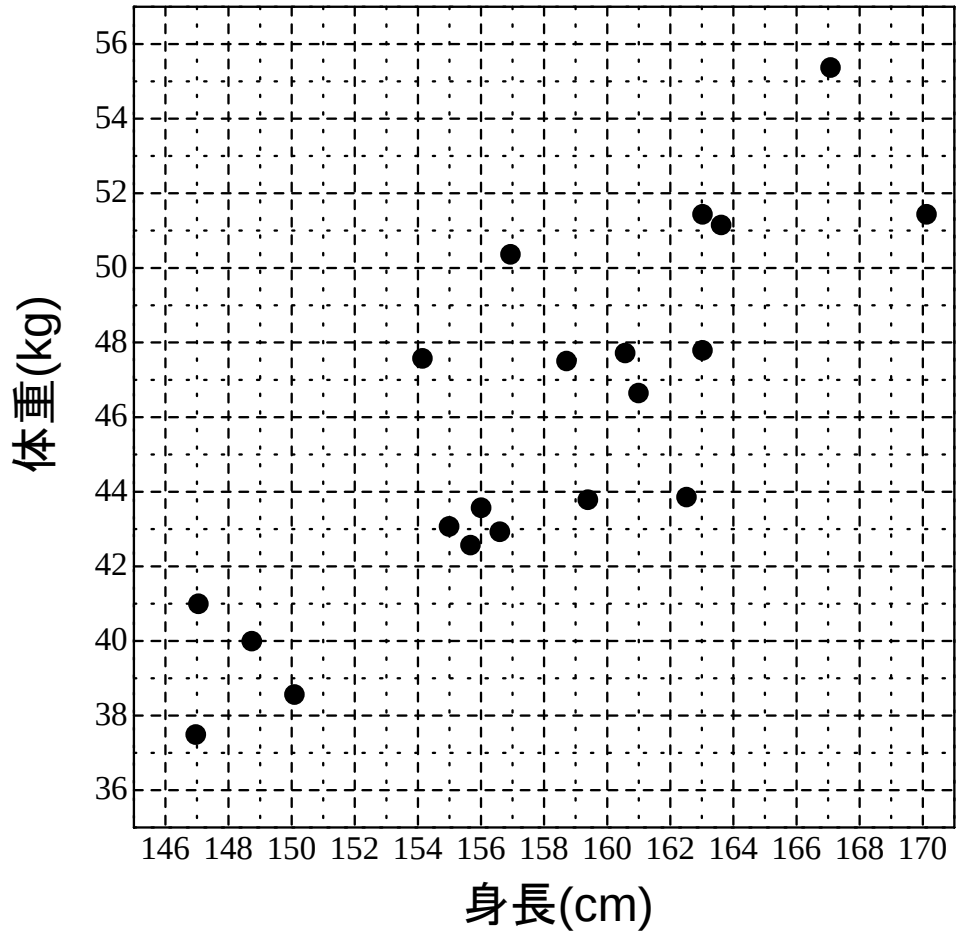
相関

中学校のあるクラスの女子生徒20人の身長と体重を調べた結果。

生徒	1	2	3	4	5	6	7	8	9	10
身長(cm)	146.8	156.3	163.0	170.1	161.1	155.2	163.2	150.2	156.0	158.7
体重(kg)	37.5	43.0	47.9	51.4	46.8	43.1	51.4	38.4	43.6	47.3

生徒	11	12	13	14	15	16	17	18	19	20
身長(cm)	162.5	147.1	167.0	157.0	163.8	160.4	155.4	154.0	148.7	159.2
体重(kg)	43.9	41.0	55.3	50.3	51.6	47.8	42.8	47.6	39.9	43.8



散布図：座標が(x_i, y_i)である観測値を点としてプロットした図。

x,yの一方が増加すれば他方が増加する傾向→正の相関。

一方が増加すれば他方が減少する傾向→負の相関。

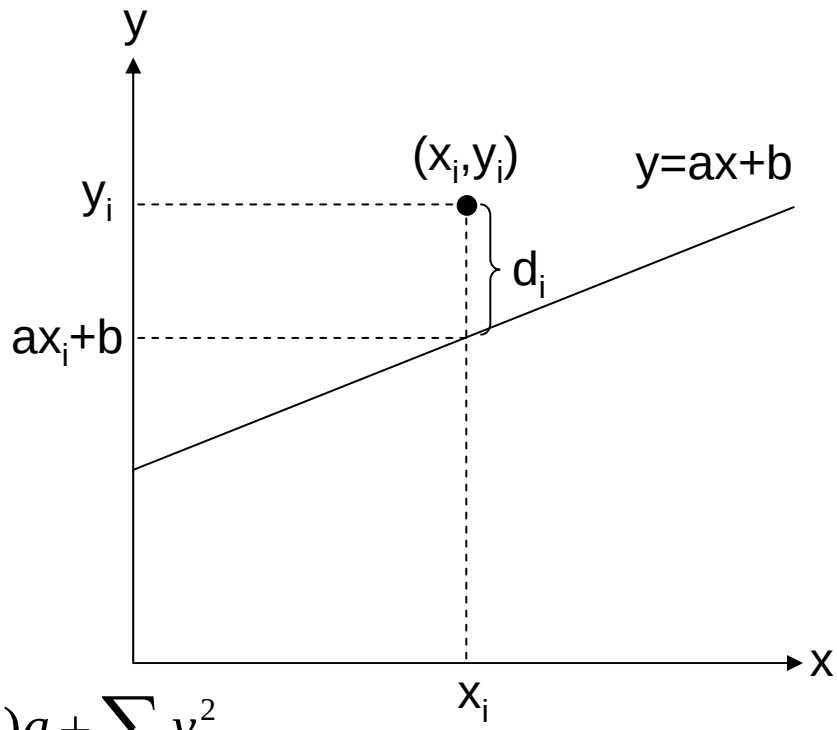
どちらもない→相関がない。

女子生徒20人の身長の体重の間には→正の相関がある。

回帰直線

散布図でデータが一つの直線のまわりに集まっていると考えられる場合、このデータを一本の直線の式で代表することを考える。

直線の式: y = ax+b



d_iの2乗和が最小になるように係数a,bを求める

$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \{y_i - (ax_i + b)\}^2$ であり、

$n\bar{x} = \sum_{i=1}^{n} x_i$ 、$n\bar{y} = \sum_{i=1}^{n} y_i$ の関係を利用。

$$\sum_{i=1}^{n} d_i^2 = nb^2 - 2n(\bar{y} - a\bar{x})b + (\sum x_i^2)a^2 - 2(\sum x_i y_i)a + \sum y_i^2$$

$$= n\{b - (\bar{y} - a\bar{x})\}^2 + (\sum x_i^2 - n\bar{x}^2)\left\{a - \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}\right\}^2 + \sum y_i^2 - n\bar{y}^2 - \frac{\left(\sum x_i y_i - n\bar{x}\bar{y}\right)^2}{\sum x_i^2 - n\bar{x}^2}$$

$b = \bar{y} - a\bar{x}$ 、 $a = \dfrac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$ を定めると $\sum_{i=1}^{n} d_i^2$ は最小になる。

$a=\dfrac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\overline{x}^2}$ の分母はxの分散に関係づけられ $\sum x_i^2 - n\overline{x}^2 = n\sigma_x^2$

また $\sigma_{xy}=\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\dfrac{1}{n}\sum_{i=1}^{n}(x_iy_i-x_i\overline{y}-\overline{x}y_i+\overline{xy})=\dfrac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{xy}$

とおくと分子は $n\sigma_{xy}$ と表される。 $a=\dfrac{\sigma_{xy}}{\sigma_x^2}$ 、 $b=\overline{y}-a\overline{x}$

求める式は、$y-\overline{y}=\dfrac{\sigma_{xy}}{\sigma_x^2}(x-\overline{x})$

定点 $(\overline{x},\overline{y})$ を通る傾き $\dfrac{\sigma_{xy}}{\sigma_x^2}$ の直線であることを示す。

σ_{xy} : 共分散という

例題1 : 上の例女子生徒20人の身長の体重の回帰直線の式を求めよ。またそれに基づき、身長160cmの人の体重を推定せよ。

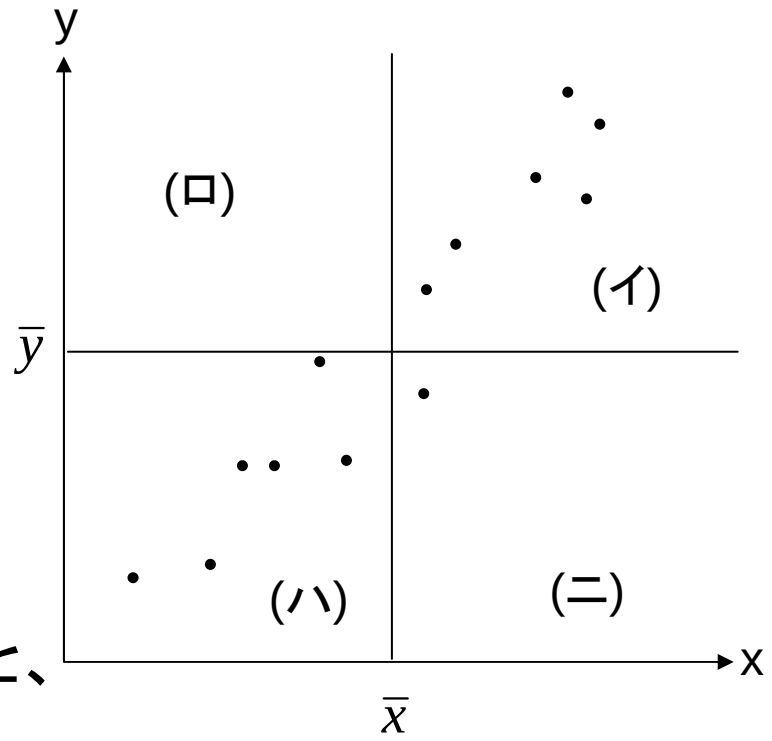
相関係数

点 (x_i, y_i) が(イ)(ハ)の部分にあるとき

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) > 0$$

点 (x_i, y_i) が(ロ)(ニ)の部分にあるとき

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) < 0$$



(イ)(ロ)(ハ)(ニ)に均等に散らばっていると、

σ_{xy} は0に近づく。

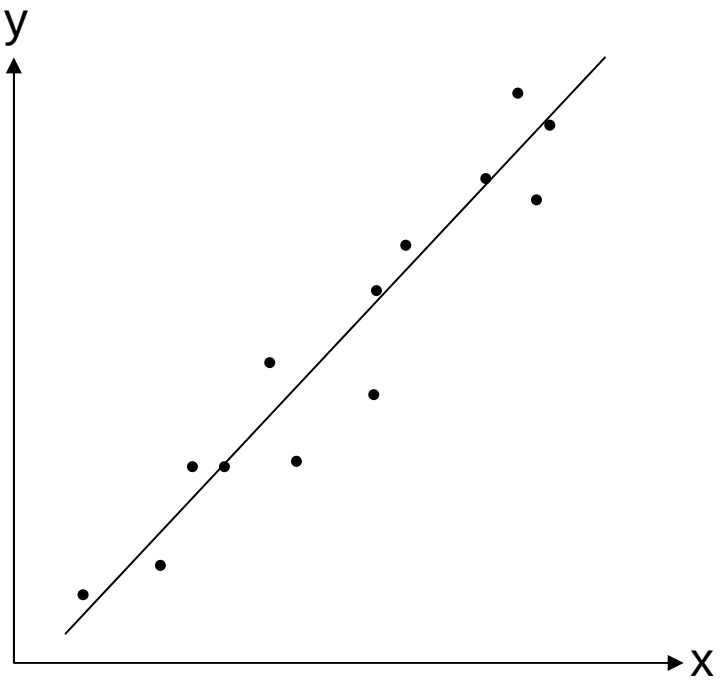
共分散 σ_{xy} は測定の単位に関係するので、これをなくすためそれぞれの標準偏差 σ_x, σ_y で割ったものを考える。

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{\sigma_x}\cdot\frac{(y_i - \bar{y})}{\sigma_y}$$

rを相関係数という。

rを用いると回帰直線の式は、 $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$ で表される。

誤差分散



各データ (x_i, y_i) が回帰直線の周りにどのくらい密集しているか考える。

$$S_y^2 = \frac{1}{n}\sum\left[y_i - \left\{ \bar{y} + r\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \right\} \right]^2$$

は、回帰直線の周りのyの分散。

S_y^2 は回帰直線から各x_iに対するyを推定したときの推定の誤差を与える（誤差分散）。

S_y^2 とrの関係： $S_y^2 = \sigma_y^2(1-r^2)$ 、$-1 \le r \le 1$

証明

$$S_y^2 = \frac{1}{n}\sum\left[(y_i - \bar{y}) - r\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \right]^2 = \frac{1}{n}\sum(y_i-\bar{y})^2 - 2r\frac{\sigma_y}{\sigma_x}\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y}) + r^2\frac{\sigma_y^2}{\sigma_x^2}\frac{1}{n}\sum(x_i-\bar{x})^2$$

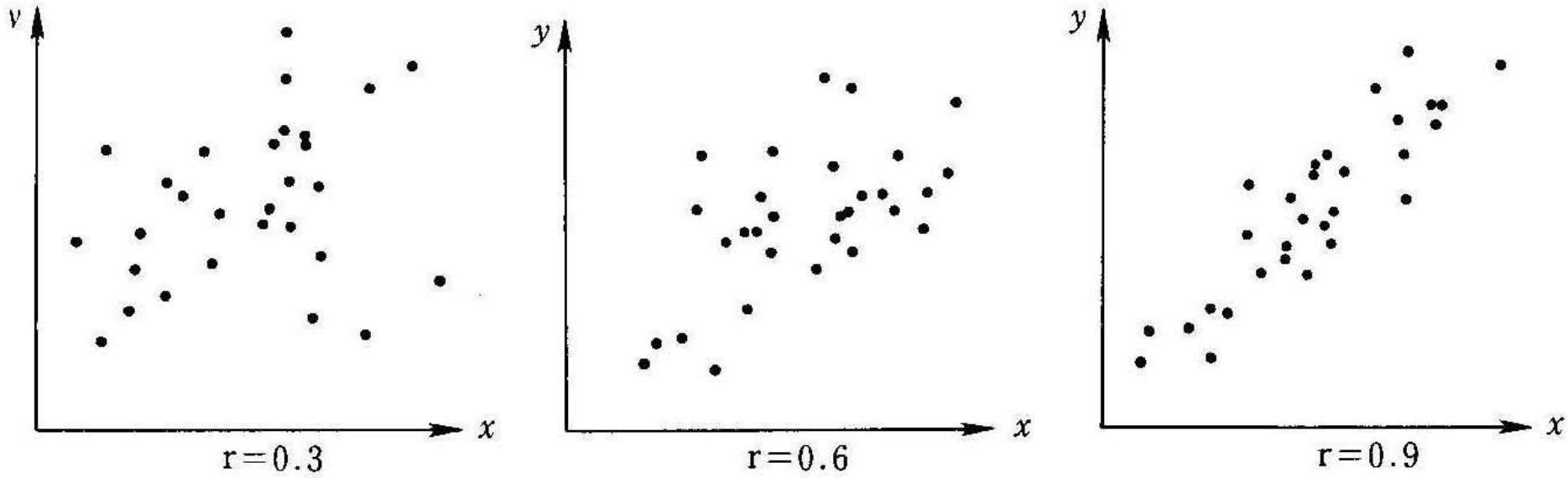
$$= \sigma_y^2 - 2r\frac{\sigma_y}{\sigma_x}\sigma_{xy} + r^2\frac{\sigma_y^2}{\sigma_x^2}\sigma_x^2 = \sigma_y^2 - 2r^2\sigma_y^2 + r^2\sigma_y^2 = \sigma_y^2(1-r^2)$$

$S_y^2 \ge 0$ で $s_y^2 \ge 0$ より、$(1-r^2) \ge 0$ したがって $-1 \le r \le 1$

相関係数の性質

r=±1のときに、S_y^2=0となり、全てのデータ(x_i,y_i)は回帰直線上にある。→相関は完全。

rが1に近いほど正の相関が強い。-1に近いほど負の相関が強い。



例題2：上の例女子生徒20人の身長の体重のデータについて相関係数を求めよ。