

平成27年度 数学科リレー講座

統計入門

3 日目「相関」

2015 年 8 月 26 日

平山 裕之

複数の変量をもつ資料

1 つの対象に複数の変量 (変数)

A 君の身長，体重，視力，...

B 君の身長，体重，視力，...

東京都の人口，面積，...

埼玉県の人ロ，面積，...

変量の間に関係があるか。

関係があれば，一方の変量から他の変量を予測できる。

10 人の身長と体重

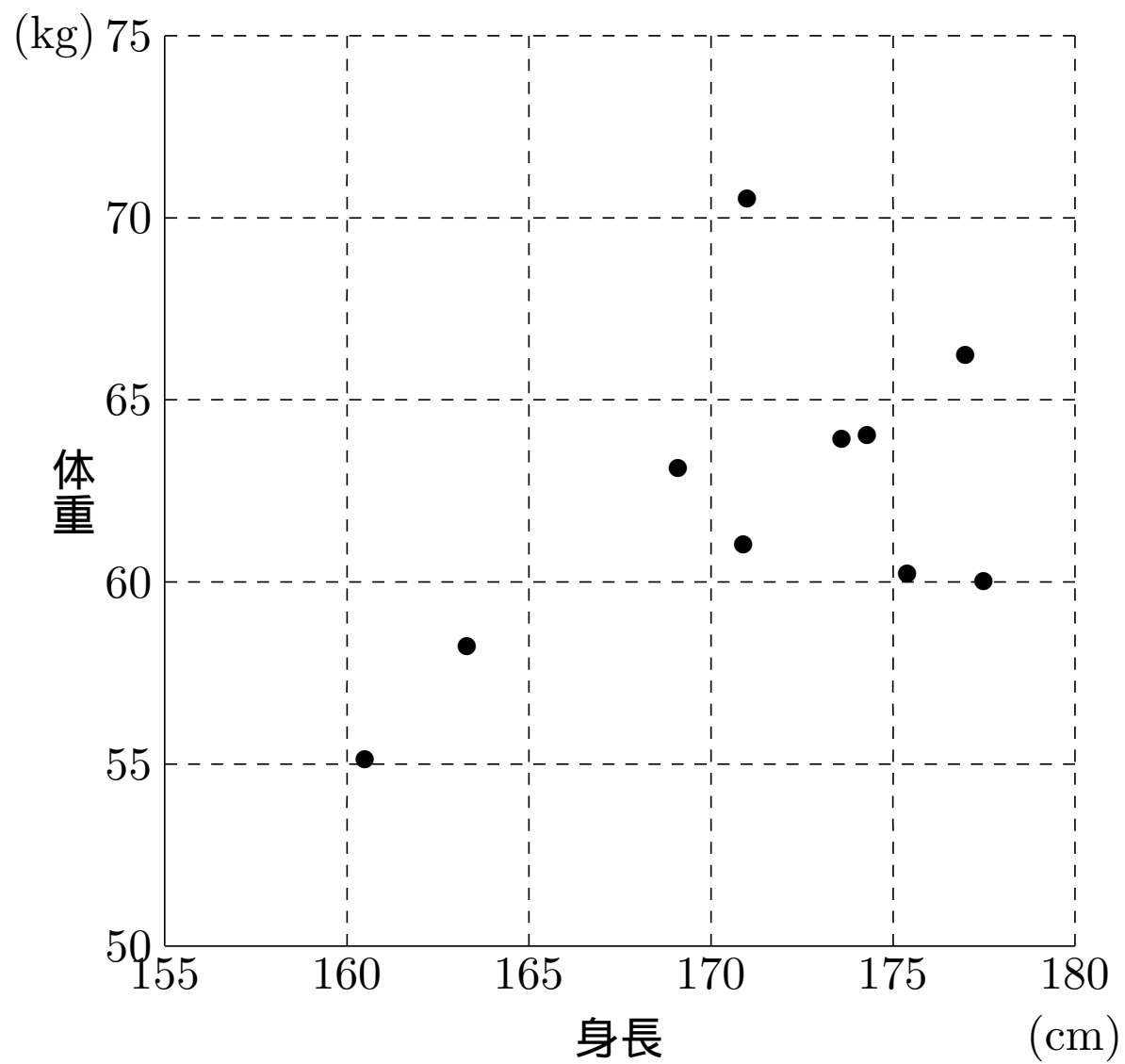
番号	身長	体重
1	177.0	66.2
2	177.5	60.0
3	163.3	58.2
4	170.9	61.0
5	169.1	63.1
6	171.0	70.5
7	173.6	63.9
8	174.3	64.0
9	160.5	55.1
10	175.4	60.2

身長を x 座標，体重を y 座標として，平面上の点で表す。

このような図を散布図または相関図という。

散布図をかけば，その関係が視覚的により明瞭に。

身長と体重



散布図を見る時には、

- 全体的な傾向をつかみ、その関係と強さを見る。
- 全体的な傾向から著しく離れたデータがあるかどうかを見る。

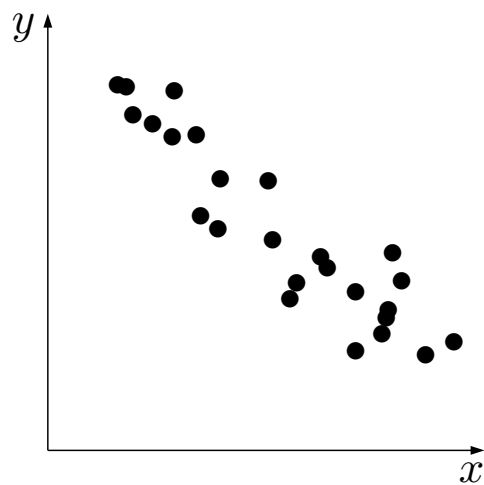
というポイントに注意する。

正の相関・負の相関

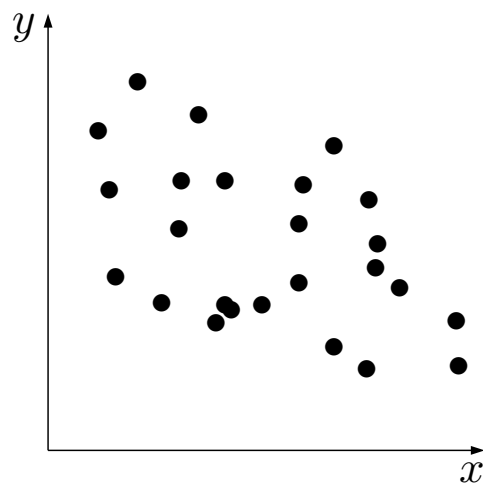
散布図において，

- (1) 一方が増加すると他方が直線的に増加する傾向があるとき，
正の相関がある。
- (2) 一方が増加すると他方が直線的に減少する傾向があるとき，
負の相関がある。
- (3) どちらでもないとき，相関関係はない(無相関)。

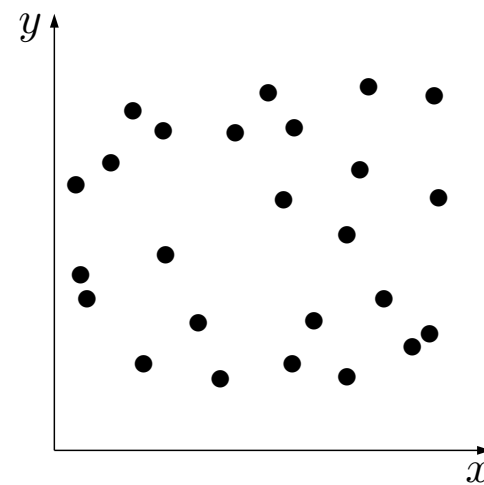
強い負の相関



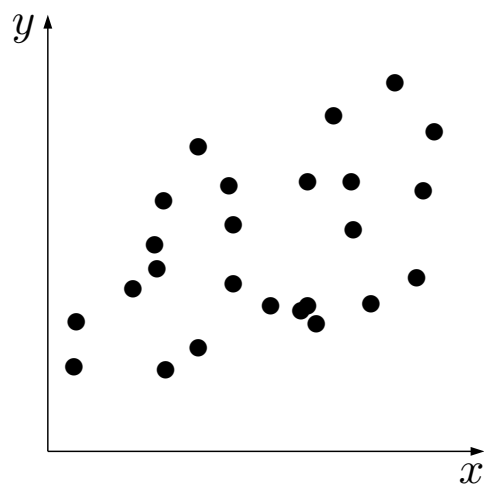
弱い負の相関



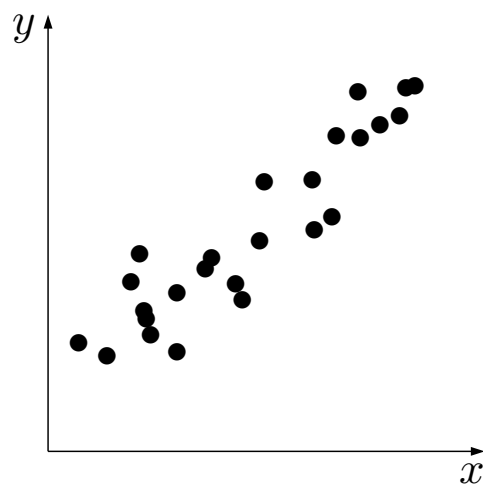
相関がない



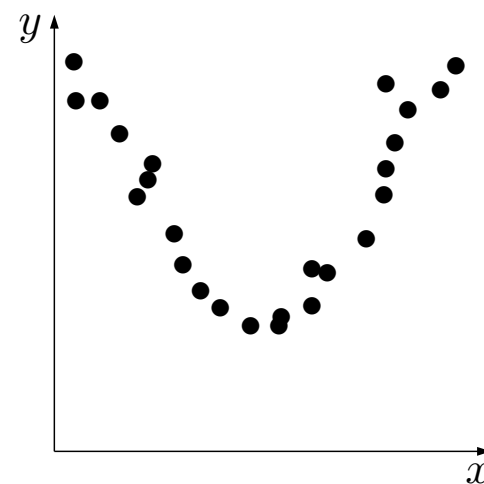
弱い正の相関



強い正の相関



直線的でない関係



2 つの変量 x, y 間にある相関関係の度合いを数値で表す。

番号	x	y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

それぞれ x, y の各変量について

変量 x の平均, 分散, 標準偏差を $\bar{x}, s_{xx}, s_x$

変量 y の平均, 分散, 標準偏差を $\bar{y}, s_{yy}, s_y$ とする。

平均

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}, \quad \bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

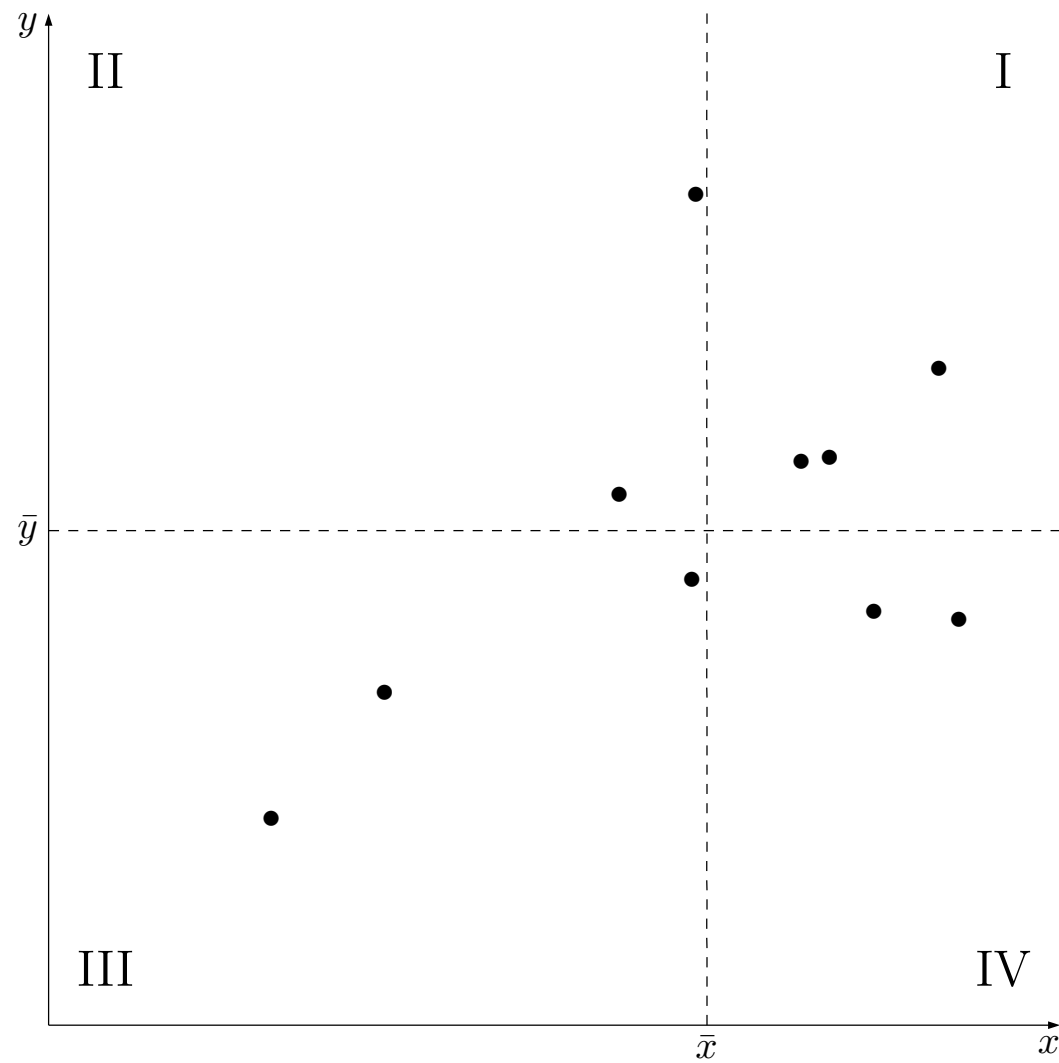
分散

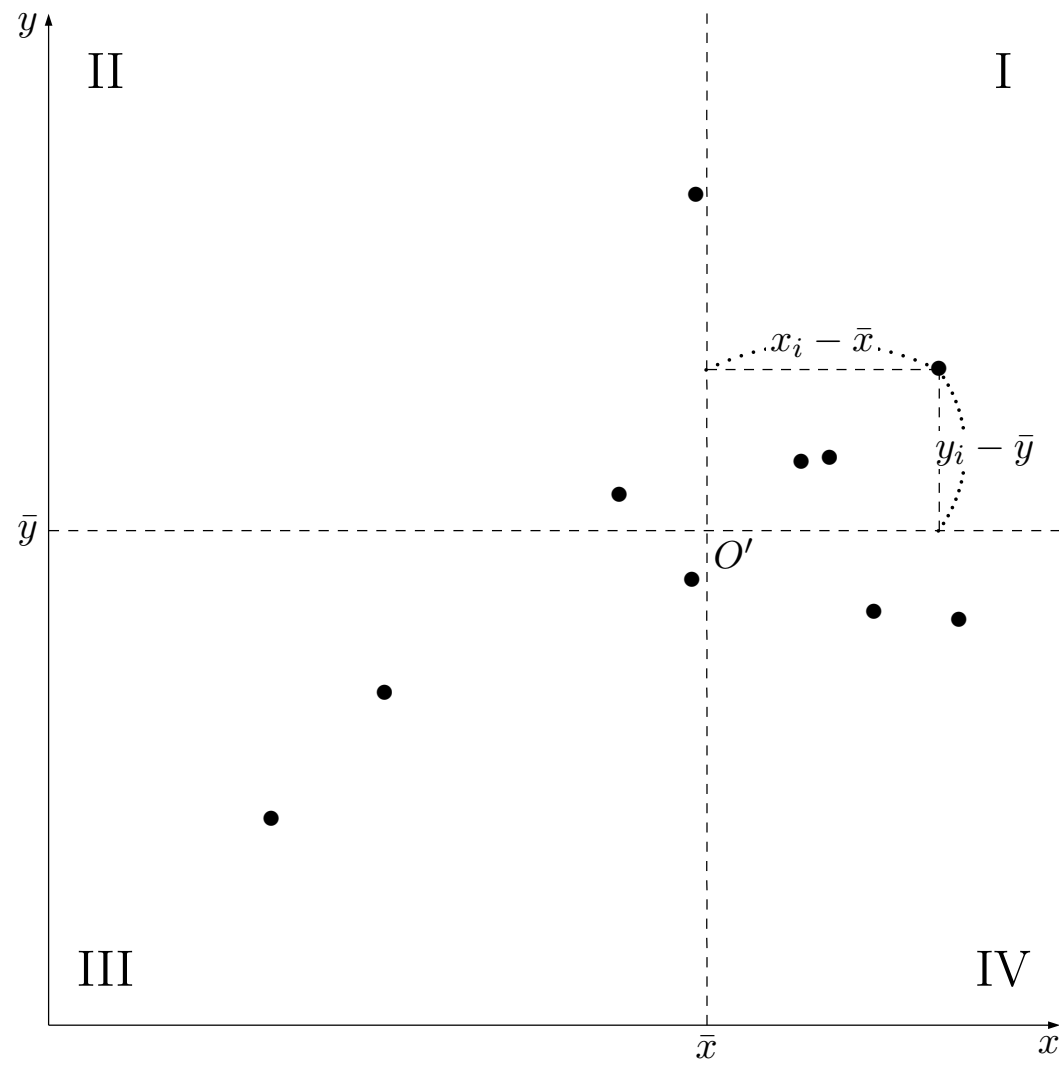
$$s_{xx} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

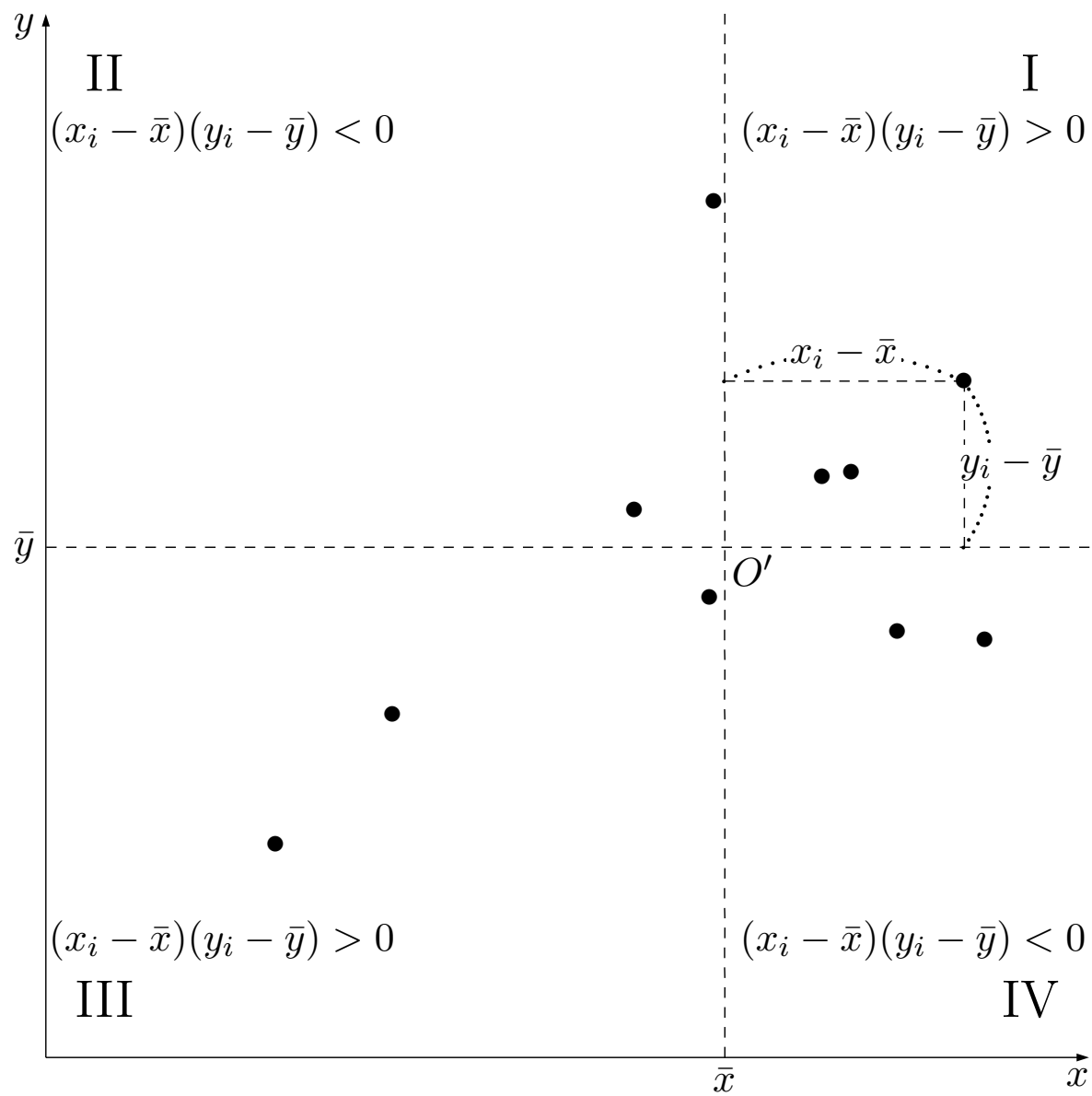
$$s_{yy} = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n}$$

標準偏差

$$s_x = \sqrt{s_{xx}}, \quad s_y = \sqrt{s_{yy}}$$







偏差の積 $(x_i - \bar{x})(y_i - \bar{y})$ の符号は

- (1) 第 I 象限に点があるとき , $(x_i - \bar{x}) > 0, (y_i - \bar{y}) > 0$ だから
 $(x_i - \bar{x})(y_i - \bar{y}) > 0$
- (2) 第 II 象限に点があるとき , $(x_i - \bar{x}) < 0, (y_i - \bar{y}) > 0$ だから
 $(x_i - \bar{x})(y_i - \bar{y}) < 0$
- (3) 第 III 象限に点があるとき , $(x_i - \bar{x}) < 0, (y_i - \bar{y}) < 0$ だから
 $(x_i - \bar{x})(y_i - \bar{y}) > 0$
- (4) 第 IV 象限に点があるとき , $(x_i - \bar{x}) > 0, (y_i - \bar{y}) < 0$ だから
 $(x_i - \bar{x})(y_i - \bar{y}) < 0$

偏差積の平均

$$s_{xy} = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

x と y の間に正の相関があるときは正の値。

x と y の間に負の相関があるときは負の値。

相関関係がないときは 0 に近い値。

s_{xy} は相関の有無を調べる指標で , x と y の共分散という。

相関係数の定義

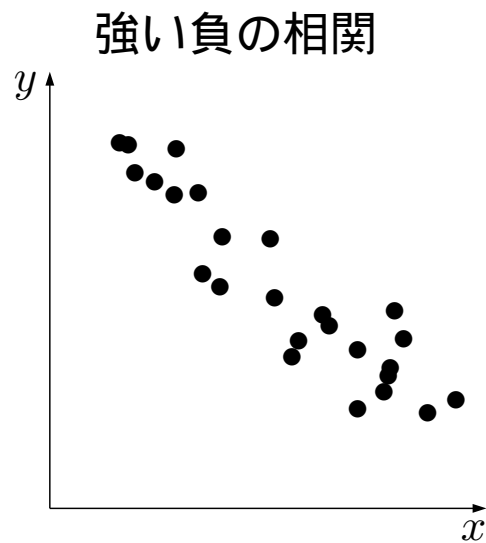
共分散を x と y の標準偏差で割る。

$$r = \frac{s_{xy}}{s_x s_y} \text{ (ピアソンの相関係数)}$$

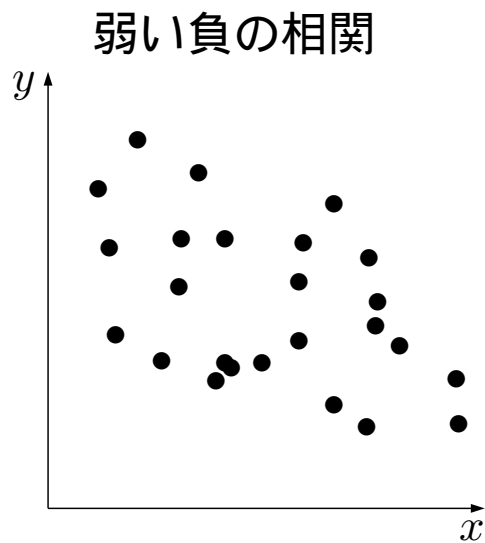
相関係数は $-1 \leq r \leq 1$ の値をとる。

$|r|$ が 1 に近いほど相関は強い。

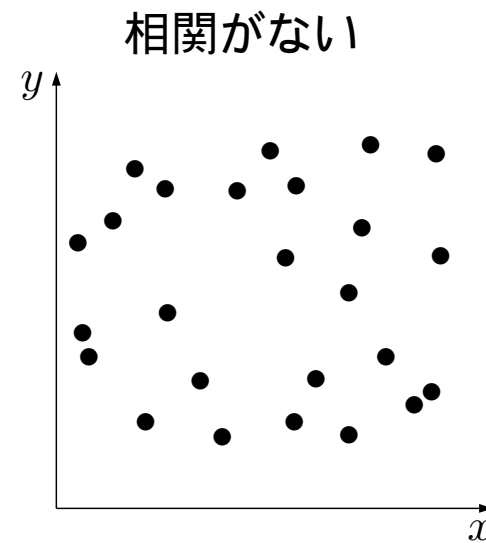
$r = \pm 1$ のとき , データは一直線上に並ぶ。(完全相関)



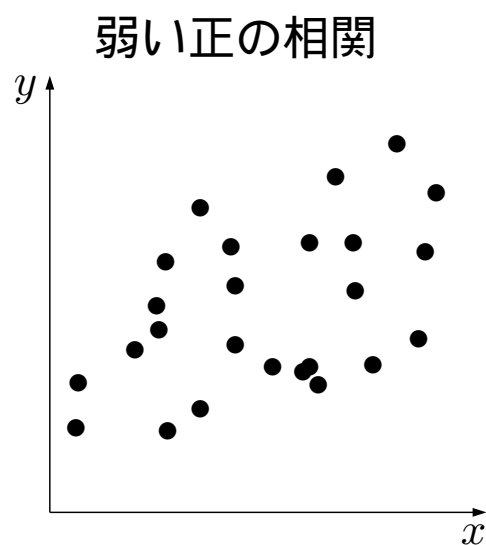
$$r = -0.92$$



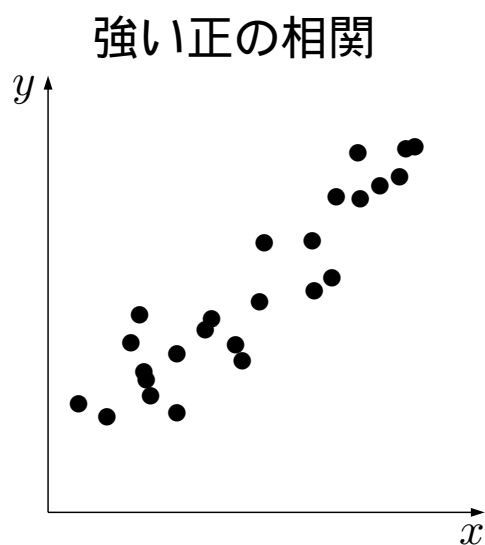
$$r = -0.48$$



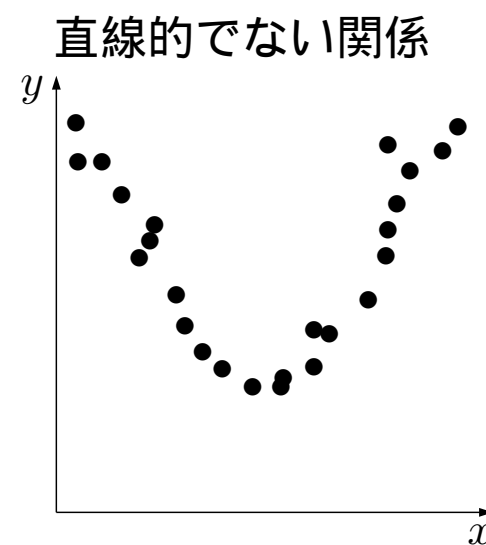
$$r = -0.02$$



$$r = 0.48$$



$$r = 0.92$$



$$r = 0.03$$

相関関係と因果関係

相関関係は因果関係を表してはいない。

2つの変量間に共通の別な変量関係して、見かけ上、相関関係がみられることがある。

擬似相関または偽相関という。

都道府県別のごみ総排出量と交通事故件数
人口という共通の原因があって、見かけ上の相関関係が生じている。

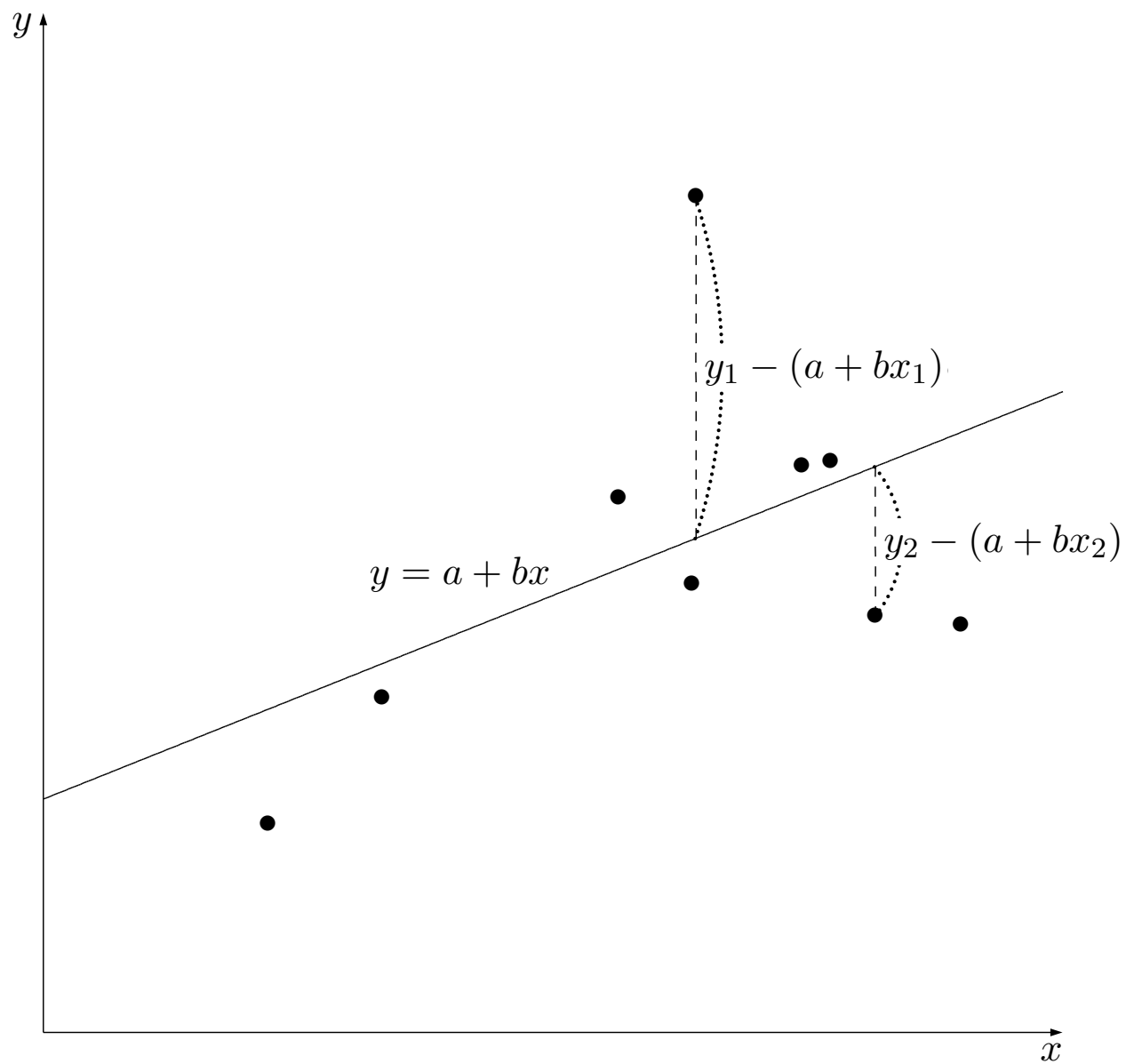
直線関係を近似する直線

x の値から y の値を求める，最もあてはまりのよい直線の式。

求める直線を $y = a + bx$ とおく。

実測値と計算値間の差 $\{y_i - (a + bx_i)\}$

この残差ができるだけ小さくなるような直線が，最もあてはまりのよい直線と考えられる。



回帰直線の決定

残差の平方和

$$Q = \{y_1 - (a + bx_1)\}^2 + \{y_2 - (a + bx_2)\}^2 + \cdots + \{y_n - (a + bx_n)\}^2$$

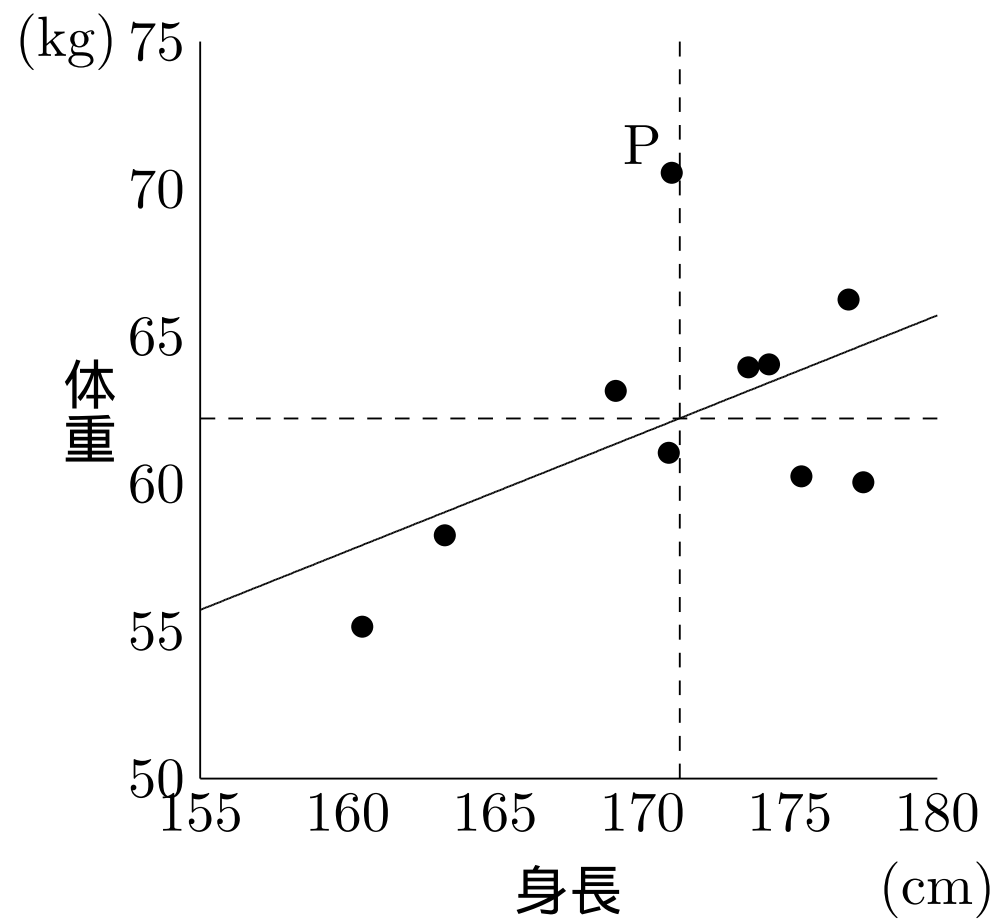
が最小になるようになるように a, b の値を決定する。

この方法を最小二乗法という。

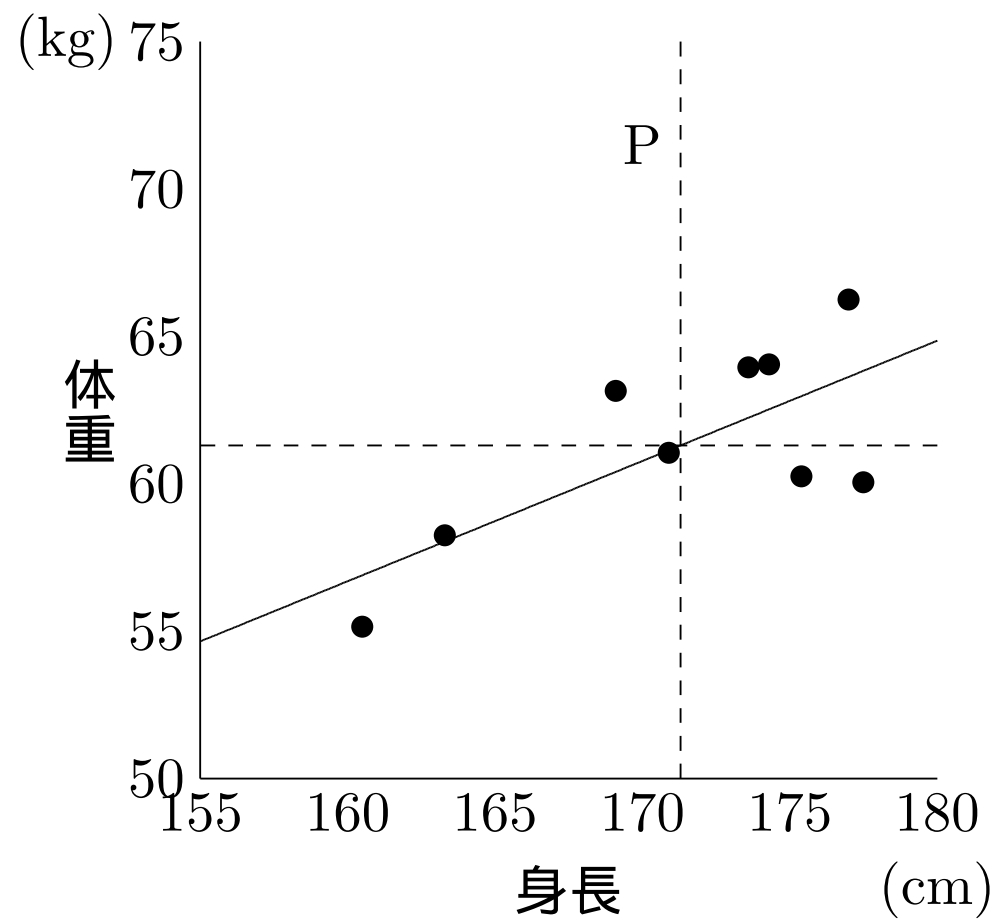
結果

$$b = \frac{s_{xy}}{s_{xx}} \text{ (回帰係数)}, \quad a = \bar{y} - b\bar{x}$$

x の分散と x, y の共分散から計算できる。



$$y = 0.40x - 6.21, \quad r = 0.52$$



$$y = 0.41x - 8.58, \quad r = 0.72$$

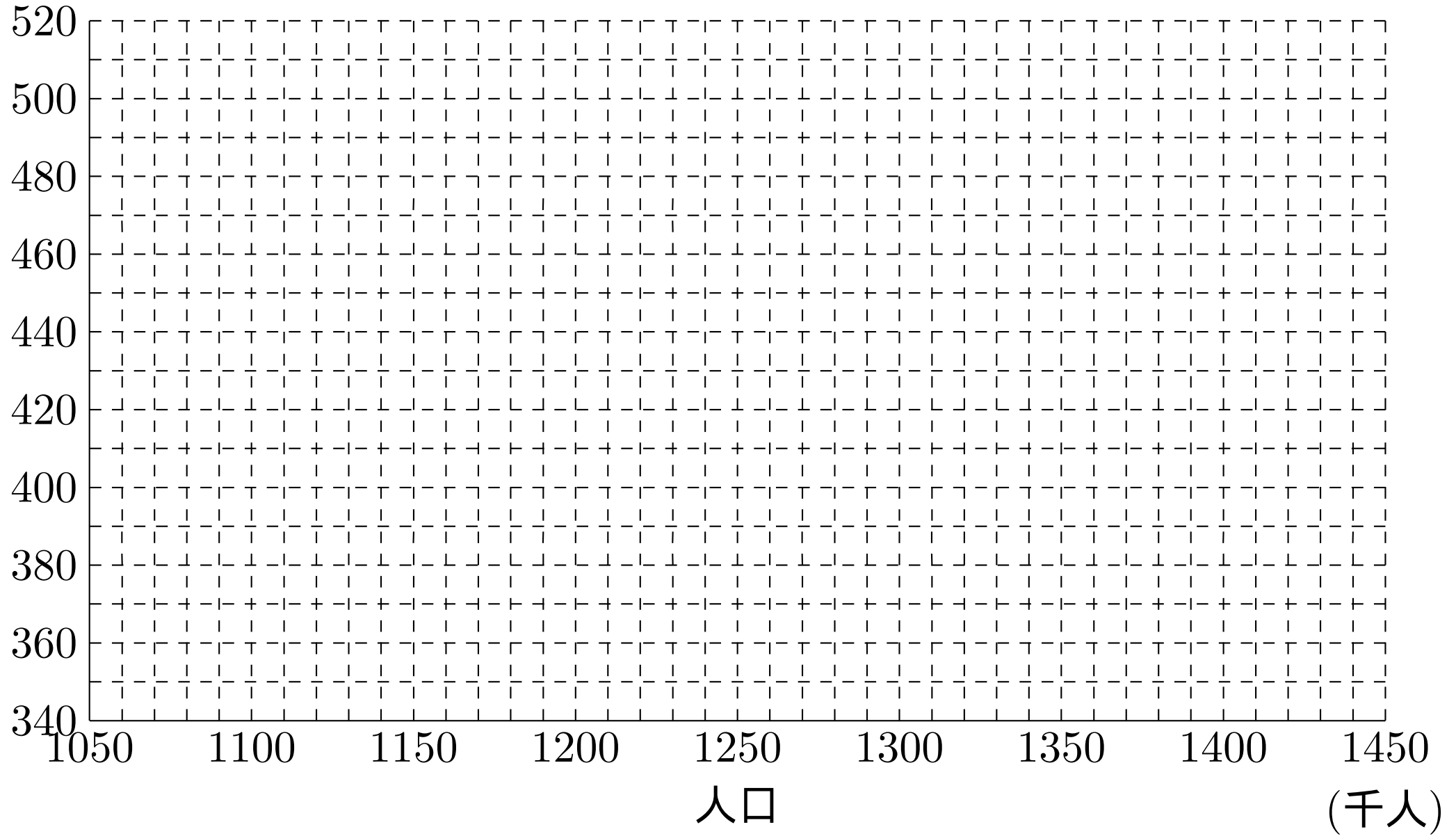
都道府県	人口 (千人)	ごみ総排出量 (千トン)
岩手	1303	449
山形	1152	377
石川	1163	425
滋賀	1415	454
長崎	1408	497
東京	13230	?
富山	?	402

- (1) 人口を x 軸に , ごみ総排出量を y 軸にとって散布図をかく。
- (2) 平均, 分散, 相関係数, 回帰直線を作業表の各欄をうめながら求める。
- (3) 資料の空欄の値を予測する。

都道府県別 人口とごみ総排出量

(千トン)

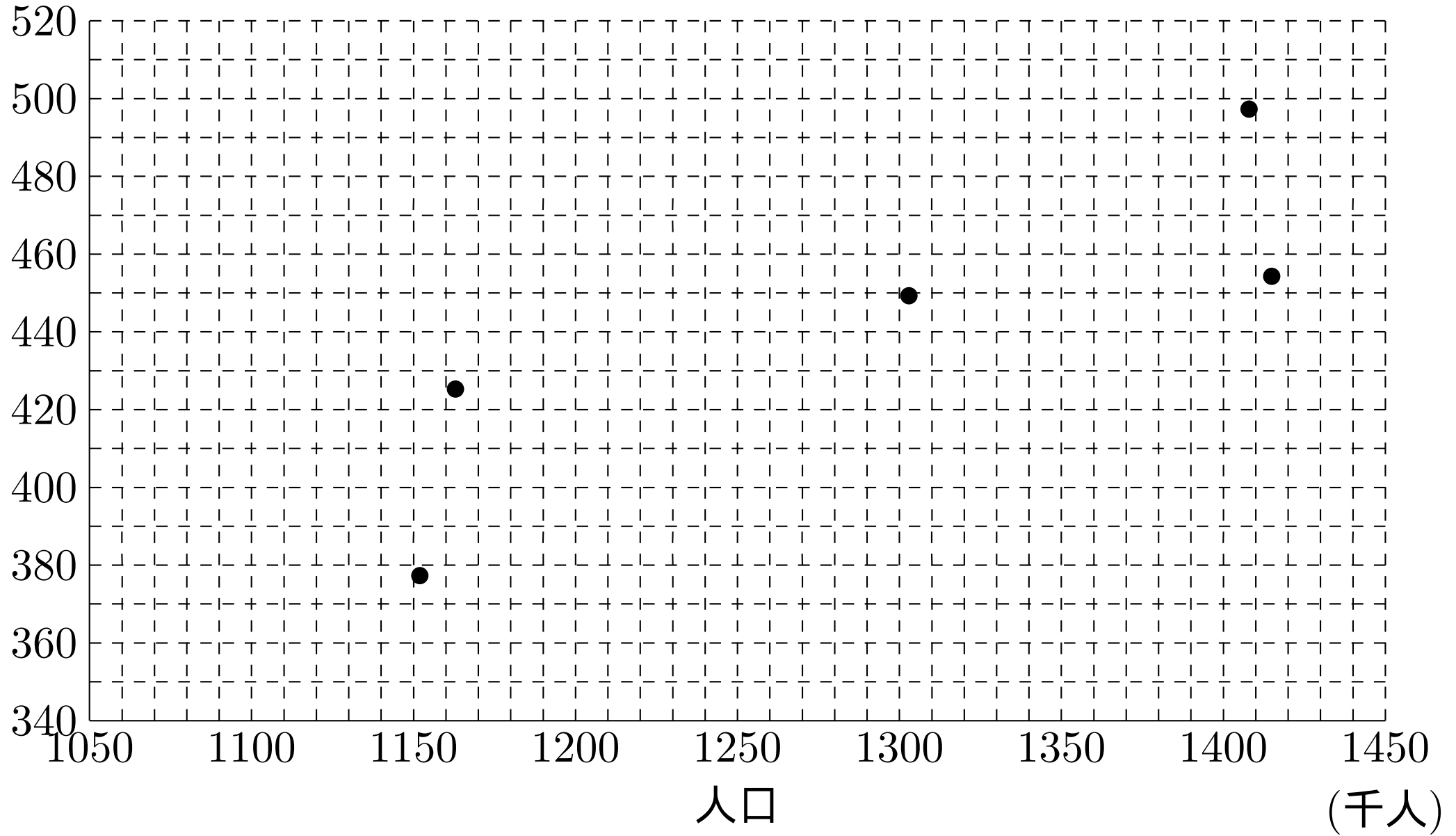
ごみ総排出量



都道府県別 人口とごみ総排出量

(千トン)

ごみ総排出量



都道府県	x	y	x^2	y^2	xy
岩手	1303	449			
山形	1152	377			
石川	1163	425			
滋賀	1415	454			
長崎	1408	497			
合計	①	②	③	④	⑤

都道府県	x	y	x^2	y^2	xy
岩手	1303	449	1697809	201601	585047
山形	1152	377	1327104	142129	434304
石川	1163	425	1352569	180625	494275
滋賀	1415	454	2002225	206116	642410
長崎	1408	497	1982464	247009	699776
合計	6441 ①	2202 ②	8362171 ③	977480 ④	2855812 ⑤

$$ns_{xx} = \textcircled{3} 8362171 - \frac{\textcircled{1} 6441^2}{5} = \textcircled{6} \boxed{}$$

$$ns_{yy} = \textcircled{4} 977480 - \frac{\textcircled{2} 2202^2}{5} = \textcircled{7} \boxed{}$$

$$ns_{xy} = \textcircled{5} 2855812 - \frac{\textcircled{1} 6441 \times \textcircled{2} 2202}{5} = \textcircled{8} \boxed{}$$

$$\text{相関係数 } r = \frac{\textcircled{8}19195.6}{\sqrt{\textcircled{6}64874.8 \times \textcircled{7}7719.2}} = \textcircled{r} \boxed{}$$

$$b = \frac{\textcircled{8}19195.6}{\textcircled{6}64874.8} = \textcircled{9} \boxed{}$$

$$a = \frac{\textcircled{2}2202 - \textcircled{9} \boxed{} \times \textcircled{1}6441}{5}$$

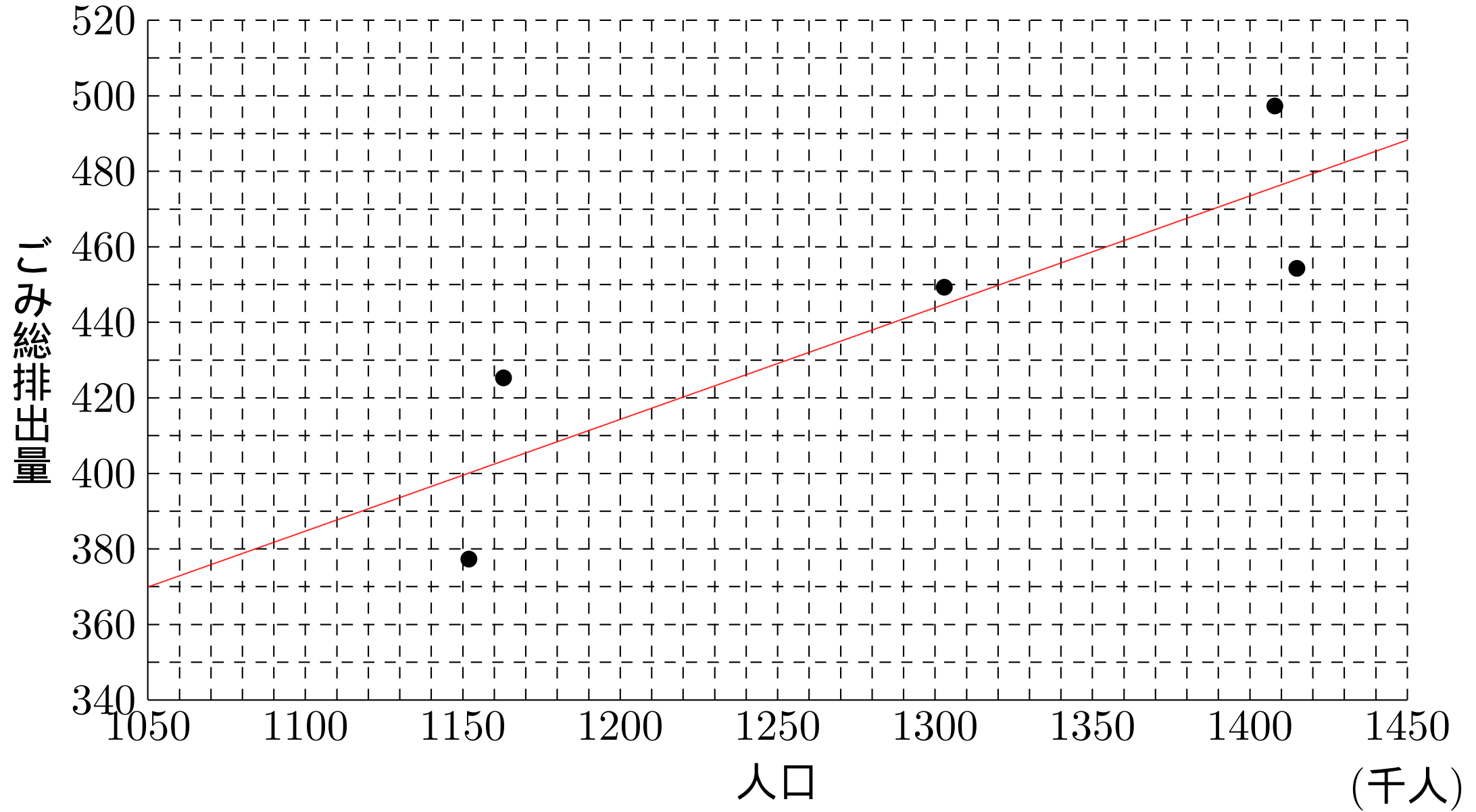
$$= \textcircled{10} \boxed{}$$

相関係数 $r = 0.86$

回帰直線 $y = 59.24 + 0.30x$

都道府県別 人口とごみ総排出量

(千トン)



回帰直線 $y = 59.24 + 0.30x$

東京都の人口 $x = 13230$ に対して, $y =$

富山県のごみ総排出量 $y = 402$ に対して, $x =$

$$nS_{xx} = \sum_i^n (x_i - \bar{x})^2 = \sum_i^n x_i^2 - \frac{\left(\sum_i^n x_i\right)^2}{n}$$

$$nS_{xy} = \sum_i^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_i^n x_i y_i - \frac{\left(\sum_i^n x_i\right)\left(\sum_i^n y_i\right)}{n}$$