デジタル情報処理

最小二乗法

佐藤　嘉伸

yoshi@image.med.osaka-u.ac.jp

http://www.image.med.osaka-u.ac.jp/member/yoshi/

日本語ページ　→　授業の資料　→　デジタル情報処理

デジタル情報処理：授業の予定

- 最小二乗法
  - データ分析法の実践
  - (多変数の微分の基礎、線形代数の基礎)
- 直交関数展開
- フーリエ解析
- 標本化定理
- ＪＰＥＧ方式

高校 数II のレベルを前提とする。
データ処理、信号処理ソフトを使った演習を行いながら、
重要項目に重点を絞って、授業を進める。

デジタル情報処理：授業の情報

- Excel, Mathematica を使って授業を行う。
  (Mathematica を使えるようにしておくこと。必要があれば、サポートセンターに今すぐいくこと。)

  所属　Osaka Electro-Communication University

  パスワード(2007 Apr)　09849-31503-64015-30704-18799-390

- **講義の資料・情報は、以下のホームページに掲載される。**

  http://www.image.med.osaka-u.ac.jp/member/yoshi/

  日本語ページ　→　授業の資料　→　デジタル情報処理

- **成績評価は、授業中に課す演習問題（2，3回）、および、演習課題の発表（プレゼン）で評価する。**

デジタル情報処理：授業の予定

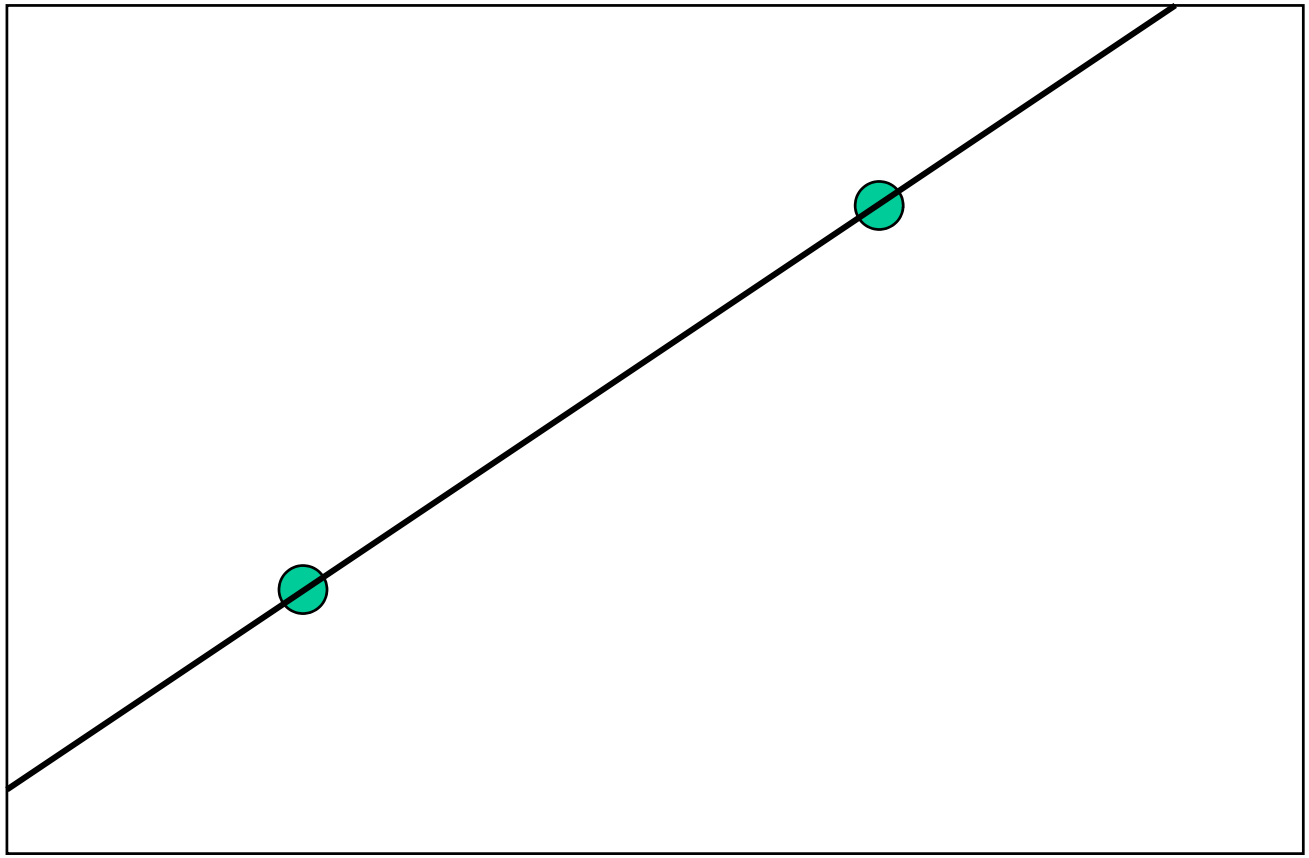
- 最小二乗法
  - データ分析法の実践
  - （多変数の微分の基礎、線形代数の基礎）
- 直交関数展開
- フーリエ解析
- 標本化定理
- （主成分分析）

高校 数II のレベルを前提とする。
データ処理、信号処理ソフトを使った演習を行いながら、
重要項目に重点を絞って、授業を進める。

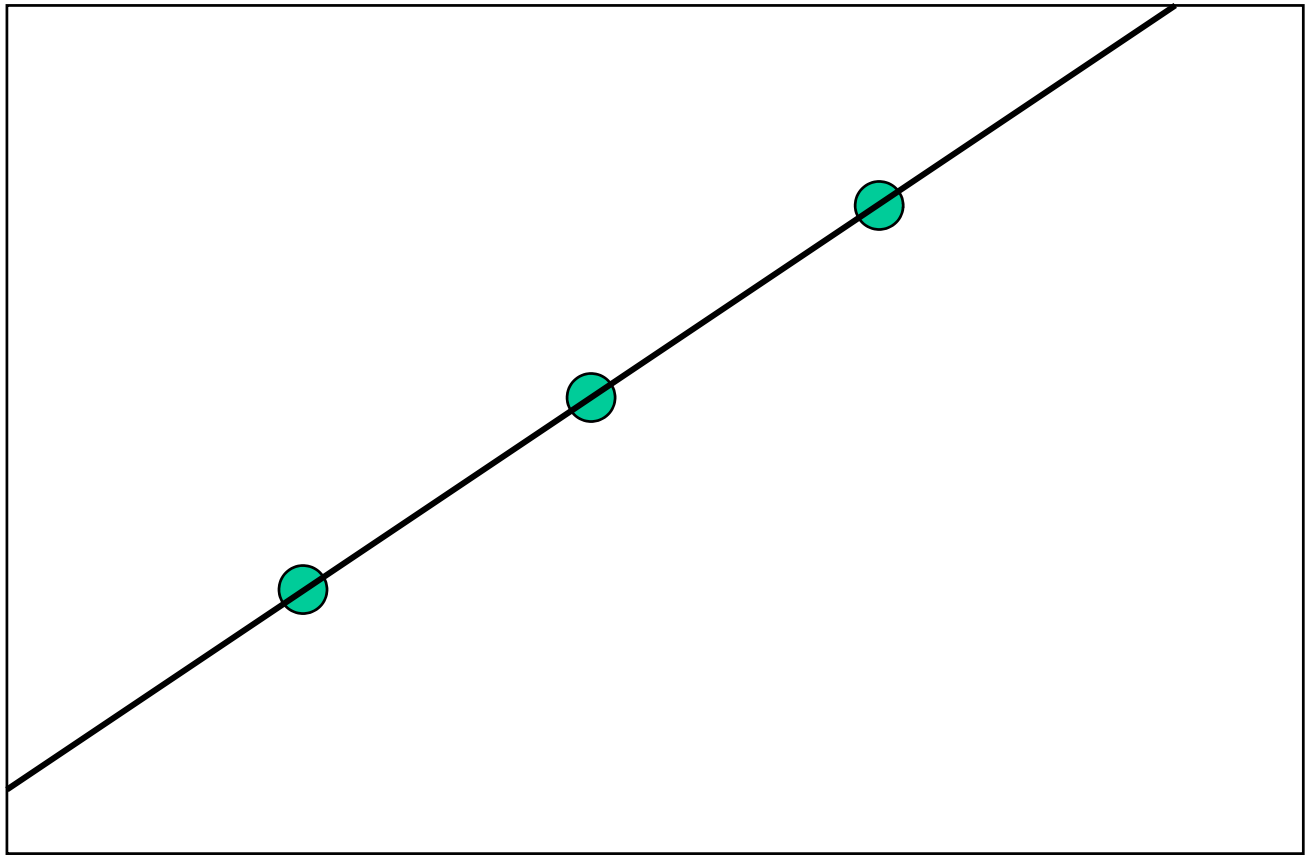
最小二乗法

- 直線当てはめ
  - 直線当てはめとは？
  - 直線当てはめの応用
  - 直線の数式表現
  - 最小二乗基準
  - 極小値の復習
  - 偏微分
  - 最小二乗法による直線当てはめ
  - Excelによる演習
- 多項式当てはめ

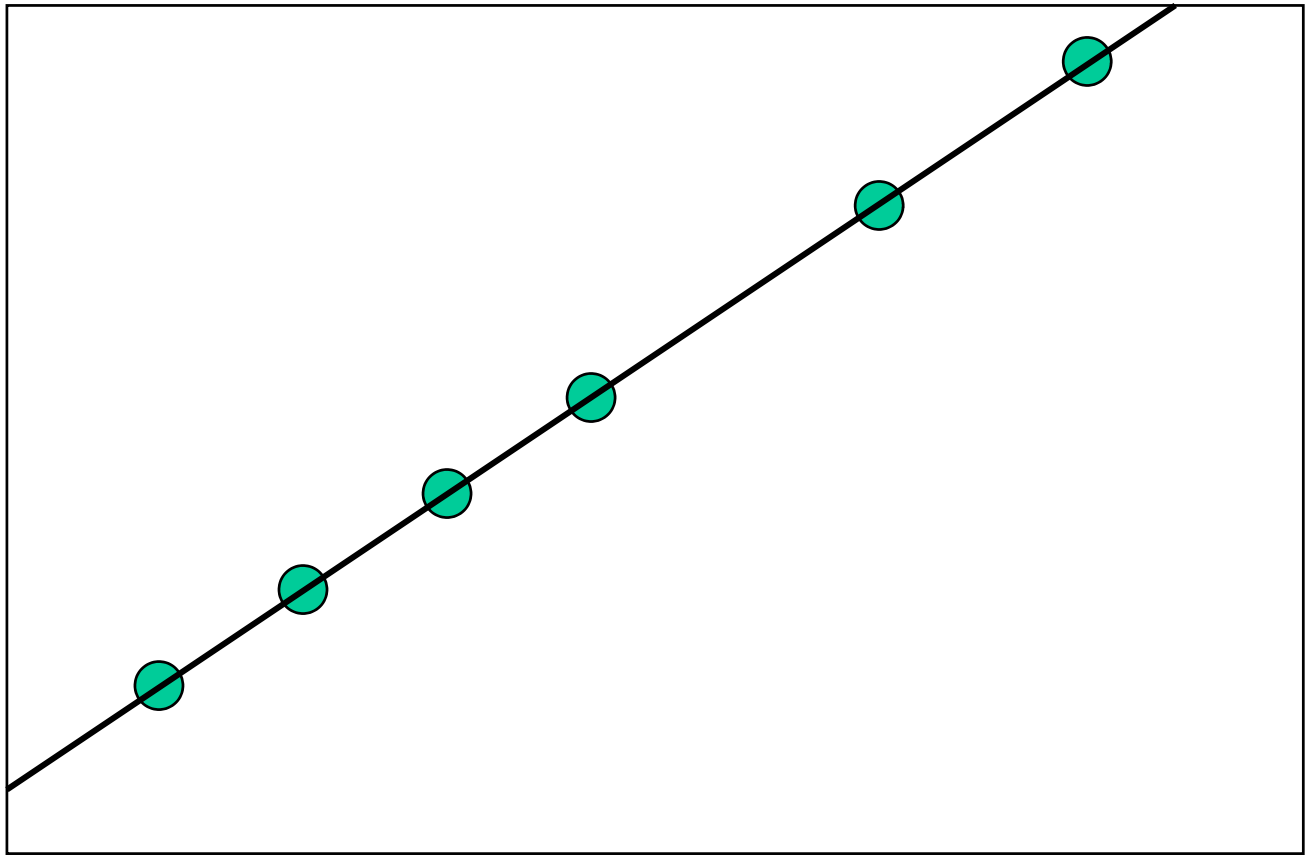
直線当てはめとは



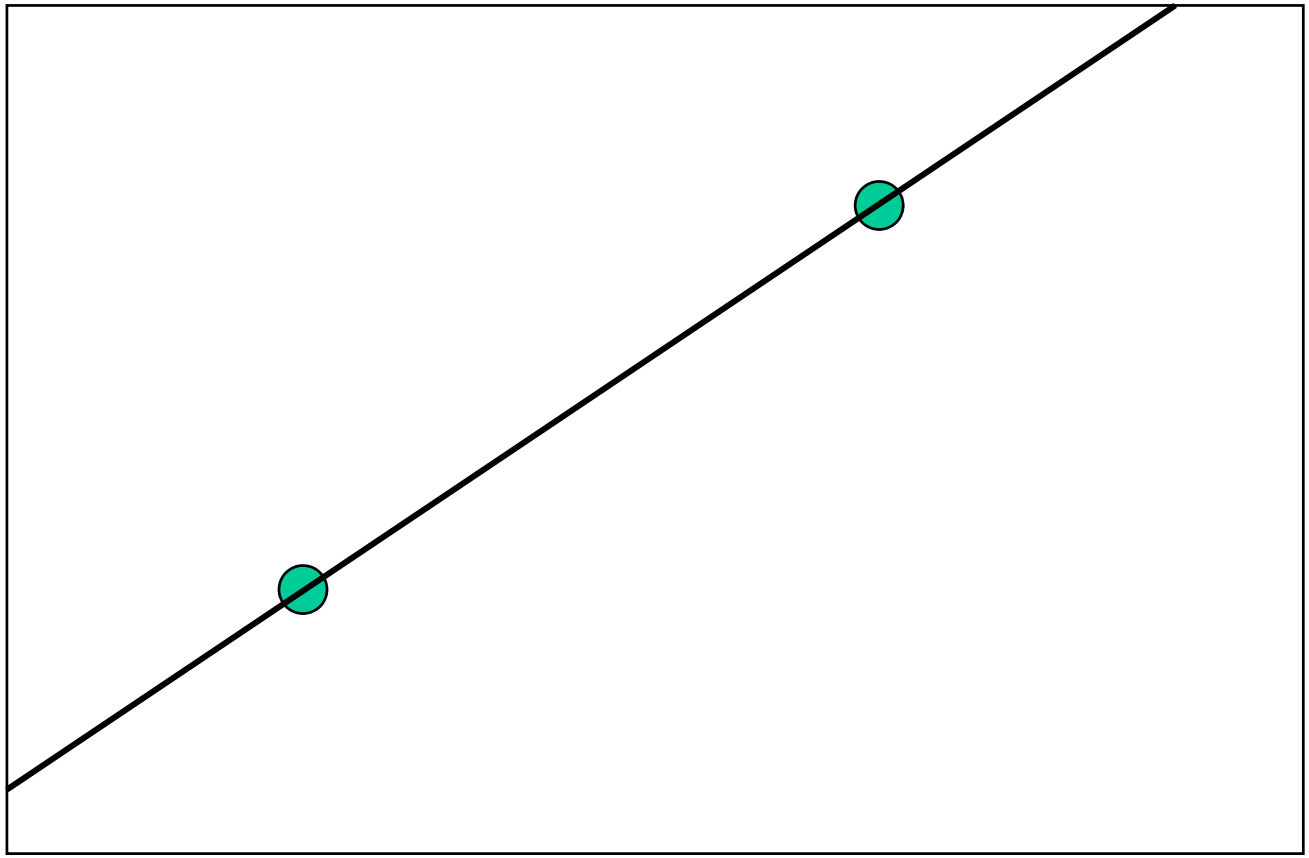
直線当てはめとは



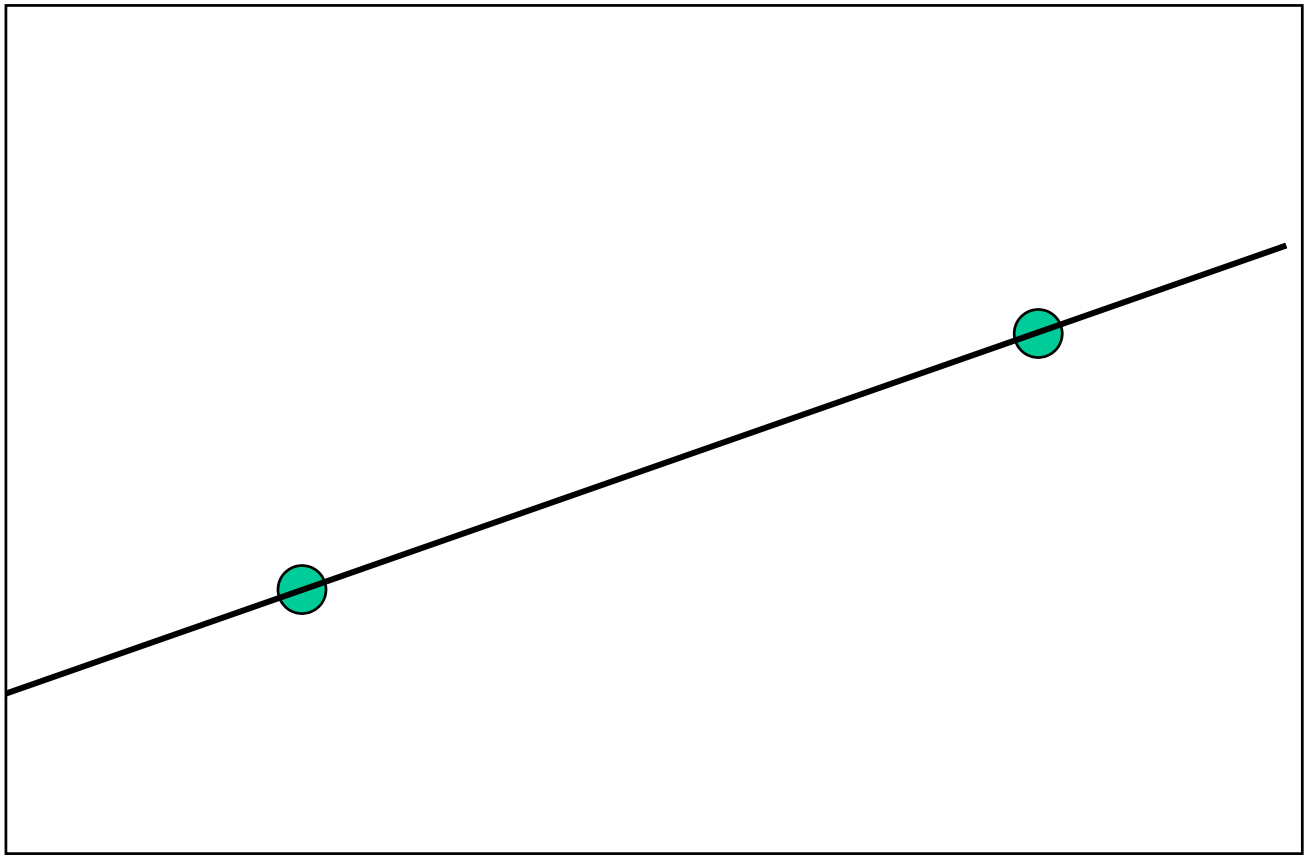
直線当てはめとは



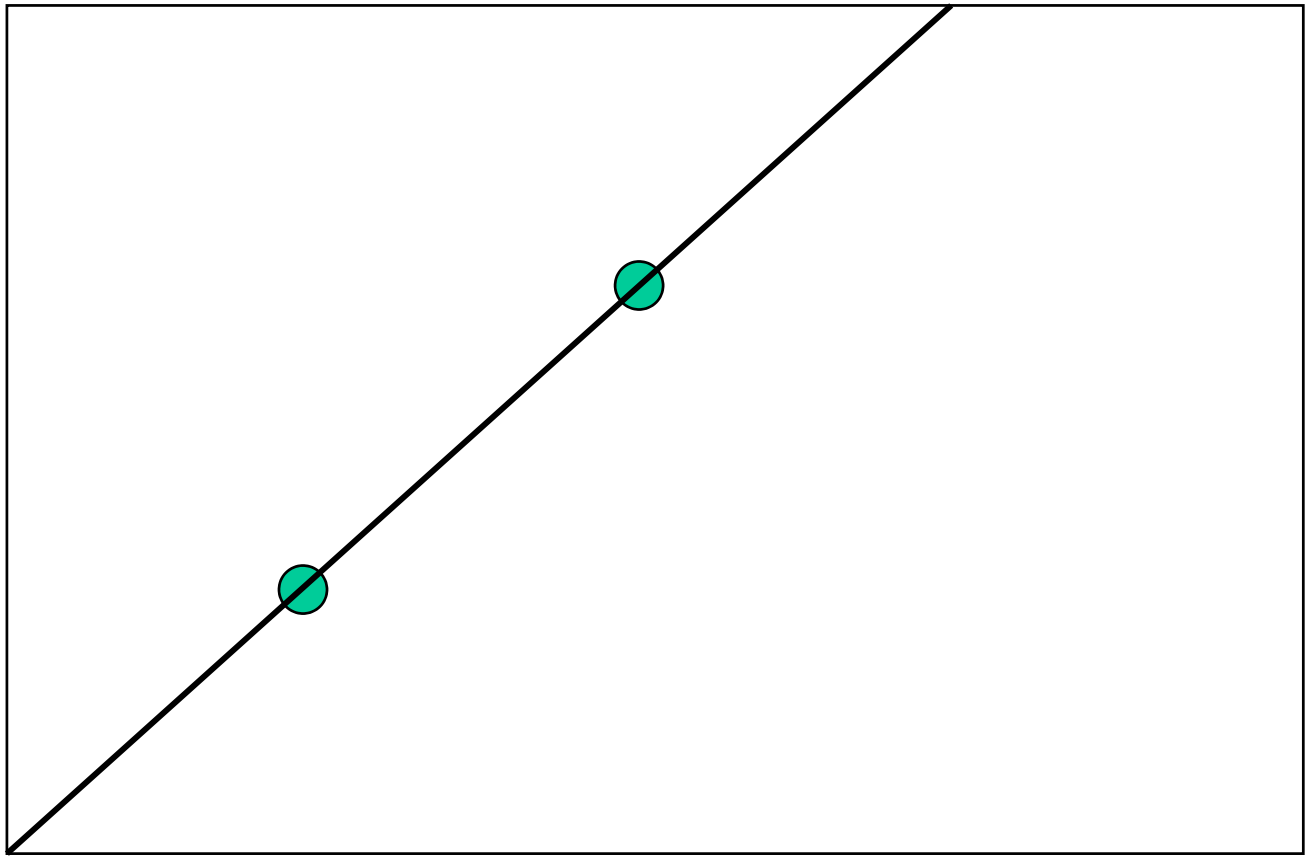
直線当てはめとは



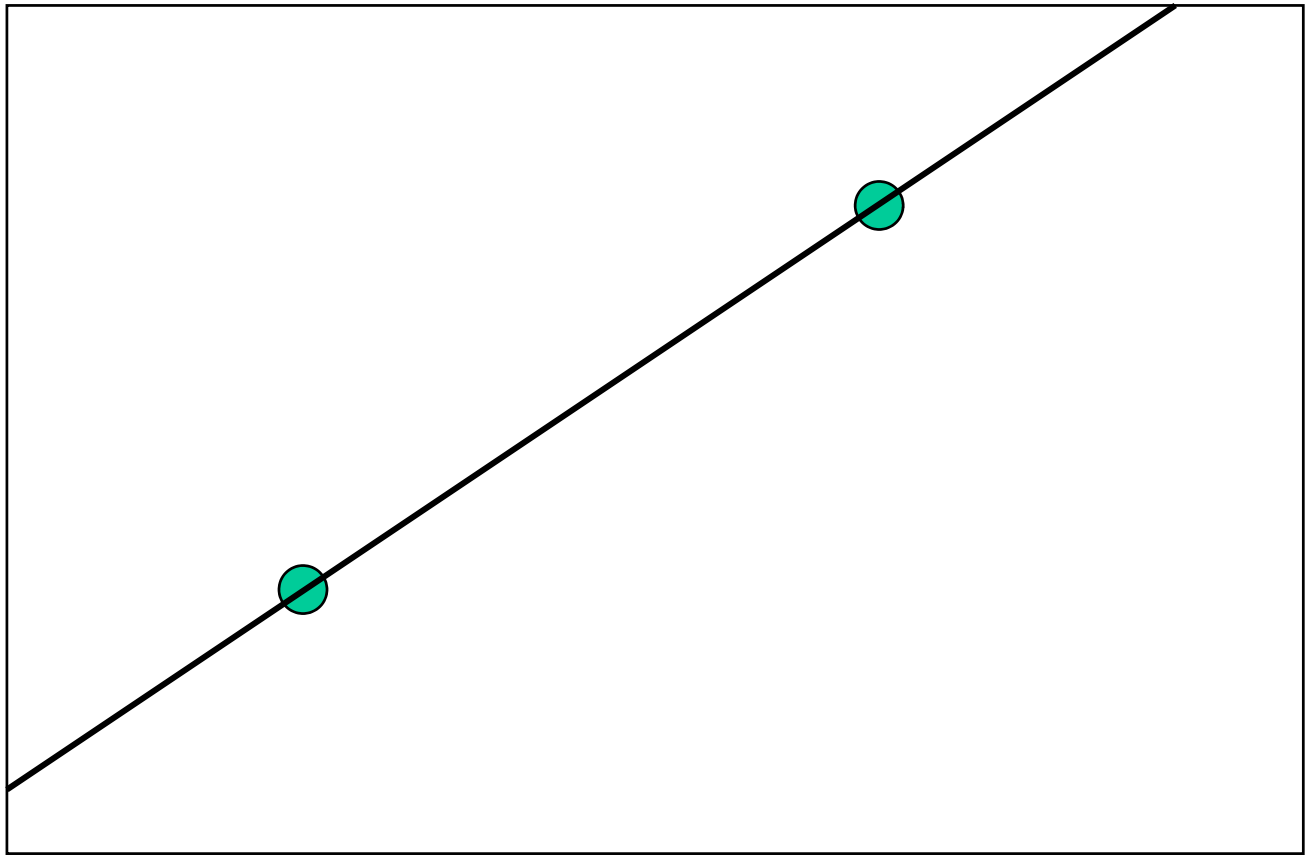
直線当てはめとは



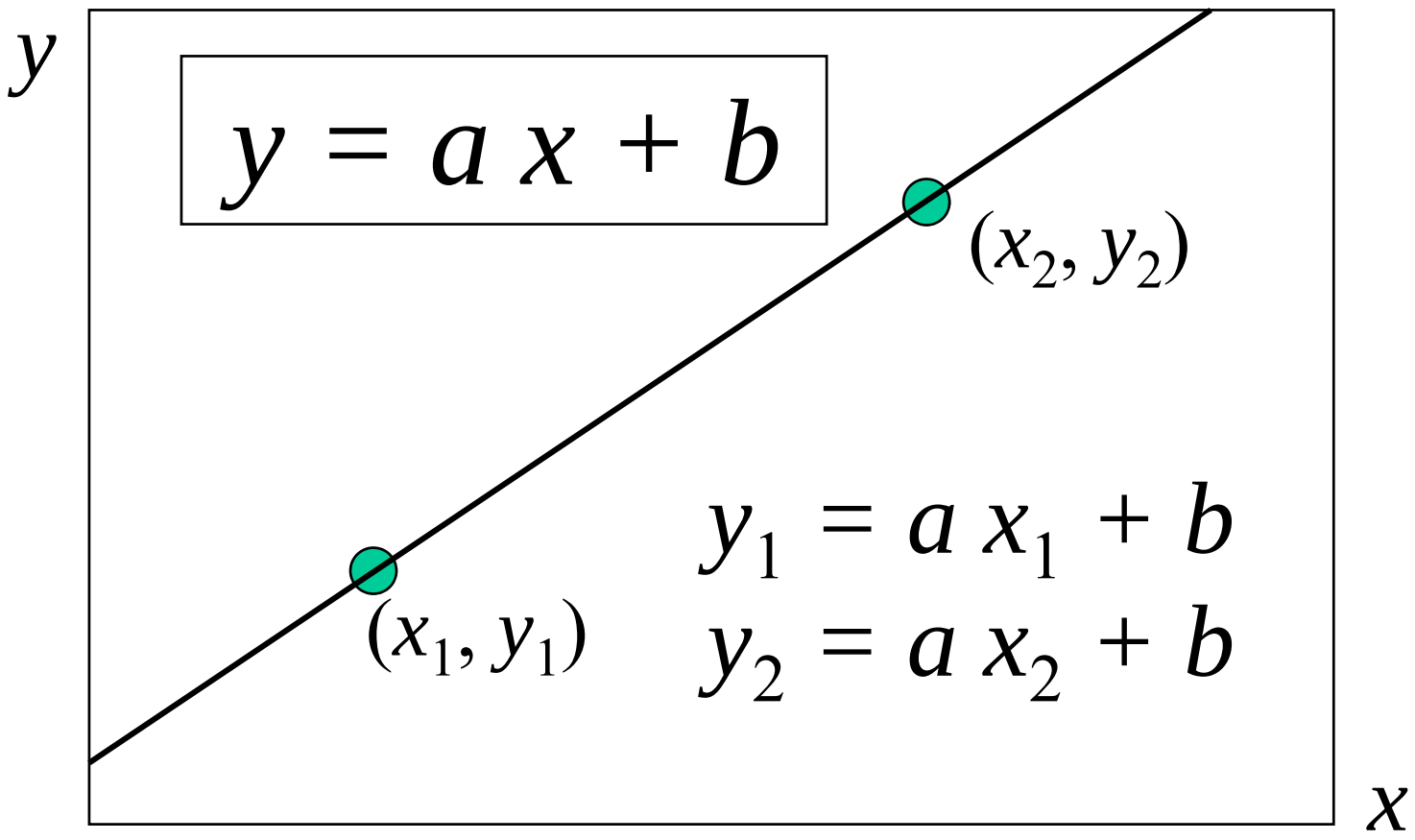
直線当てはめとは



直線当てはめとは



直線当てはめとは



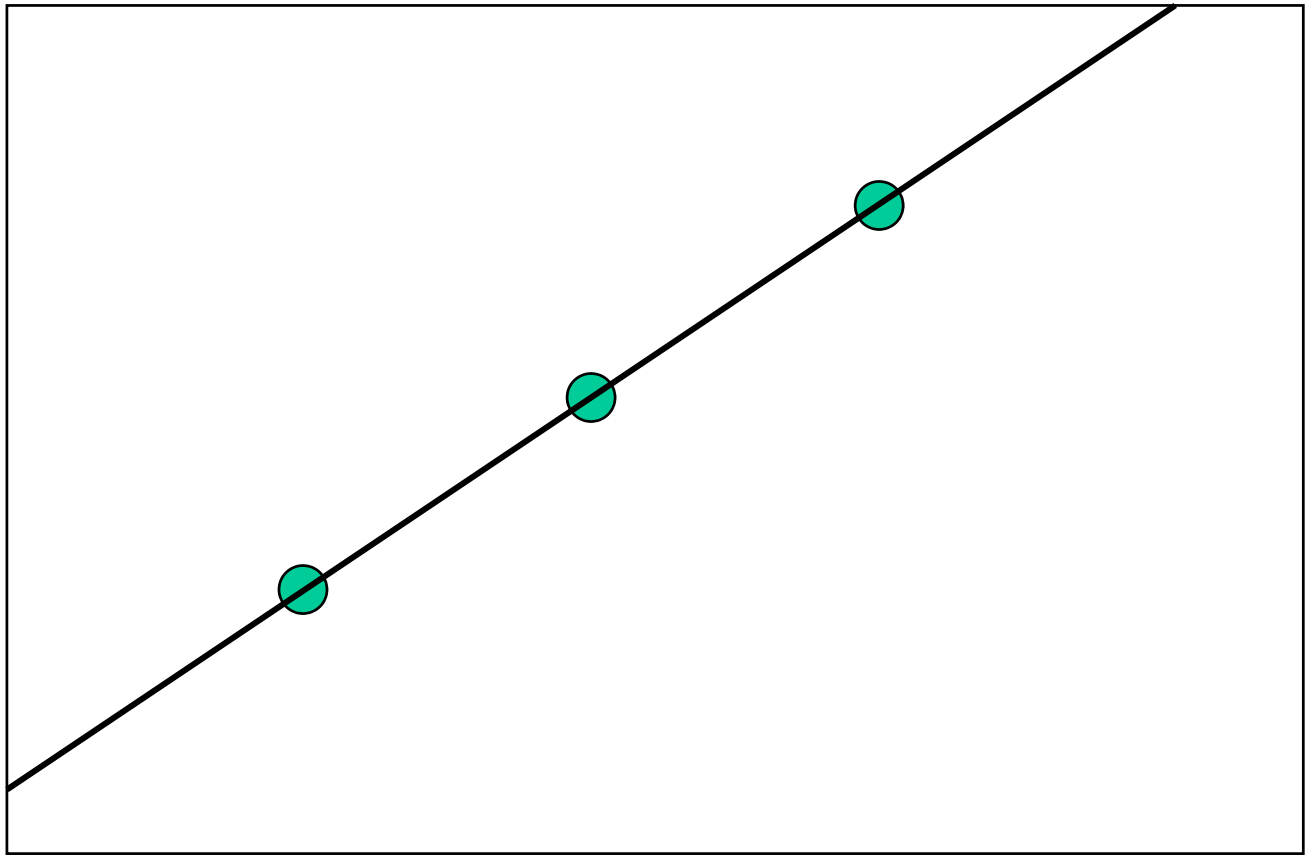
2点の座標値から直線を求める

$$y_1 = a\,x_1 + b$$
$$y_2 = a\,x_2 + b$$

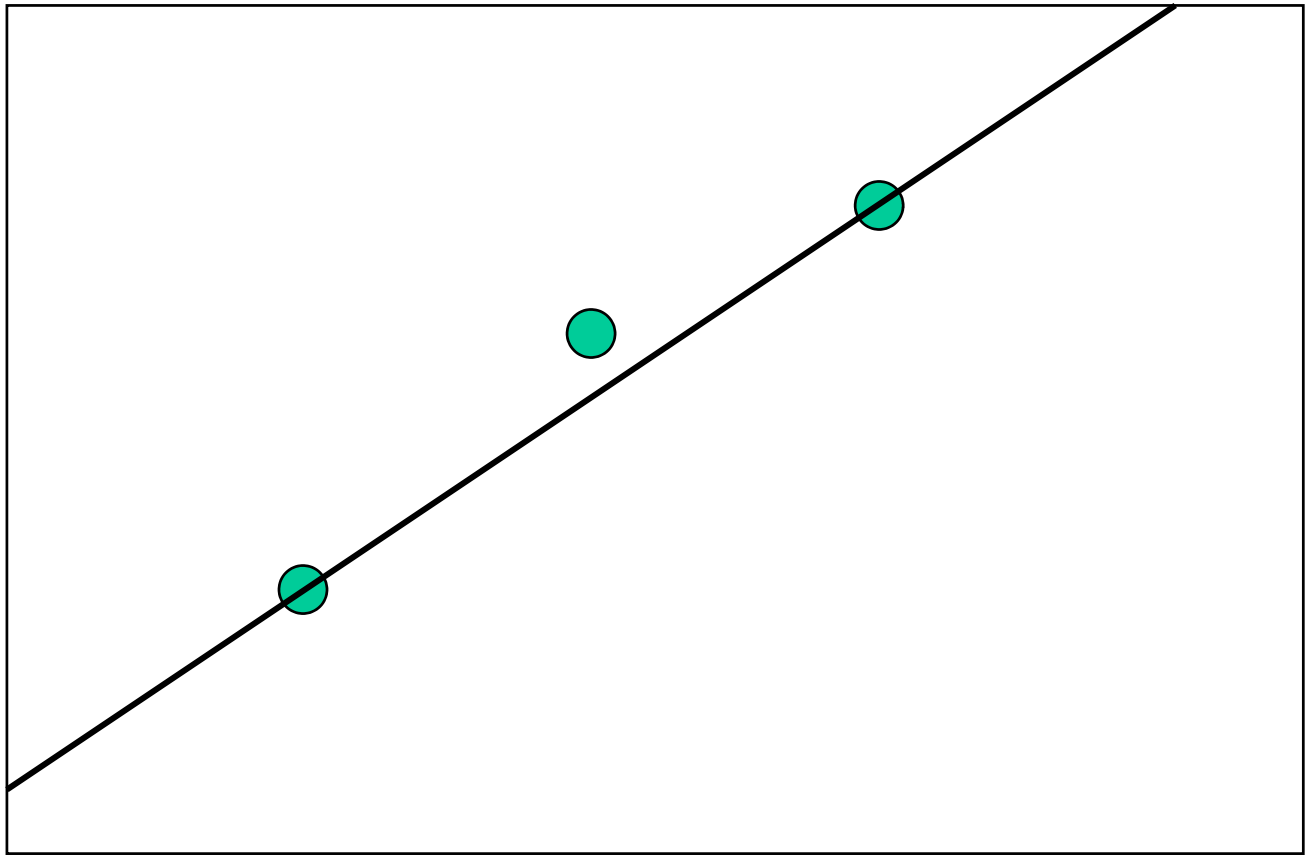
$$\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

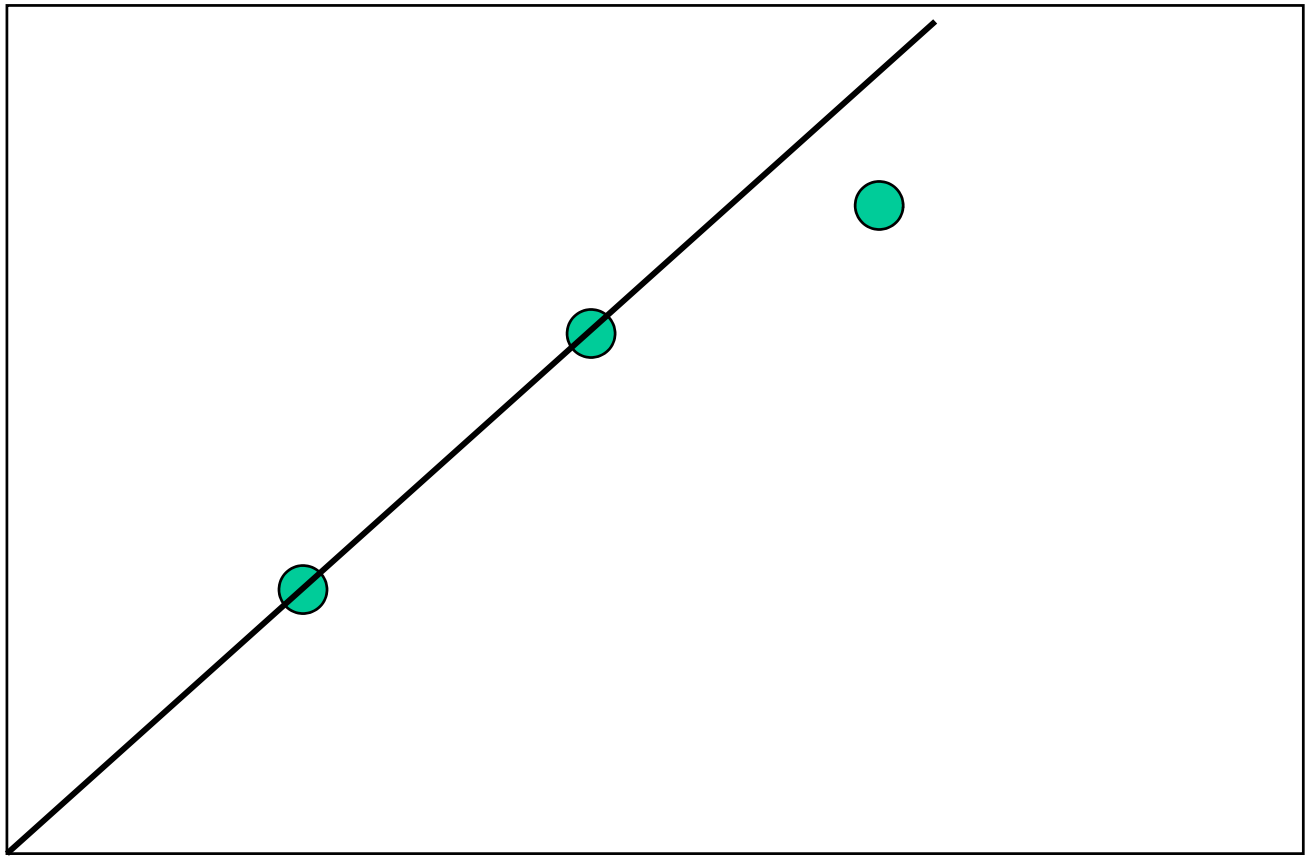
直線当てはめとは



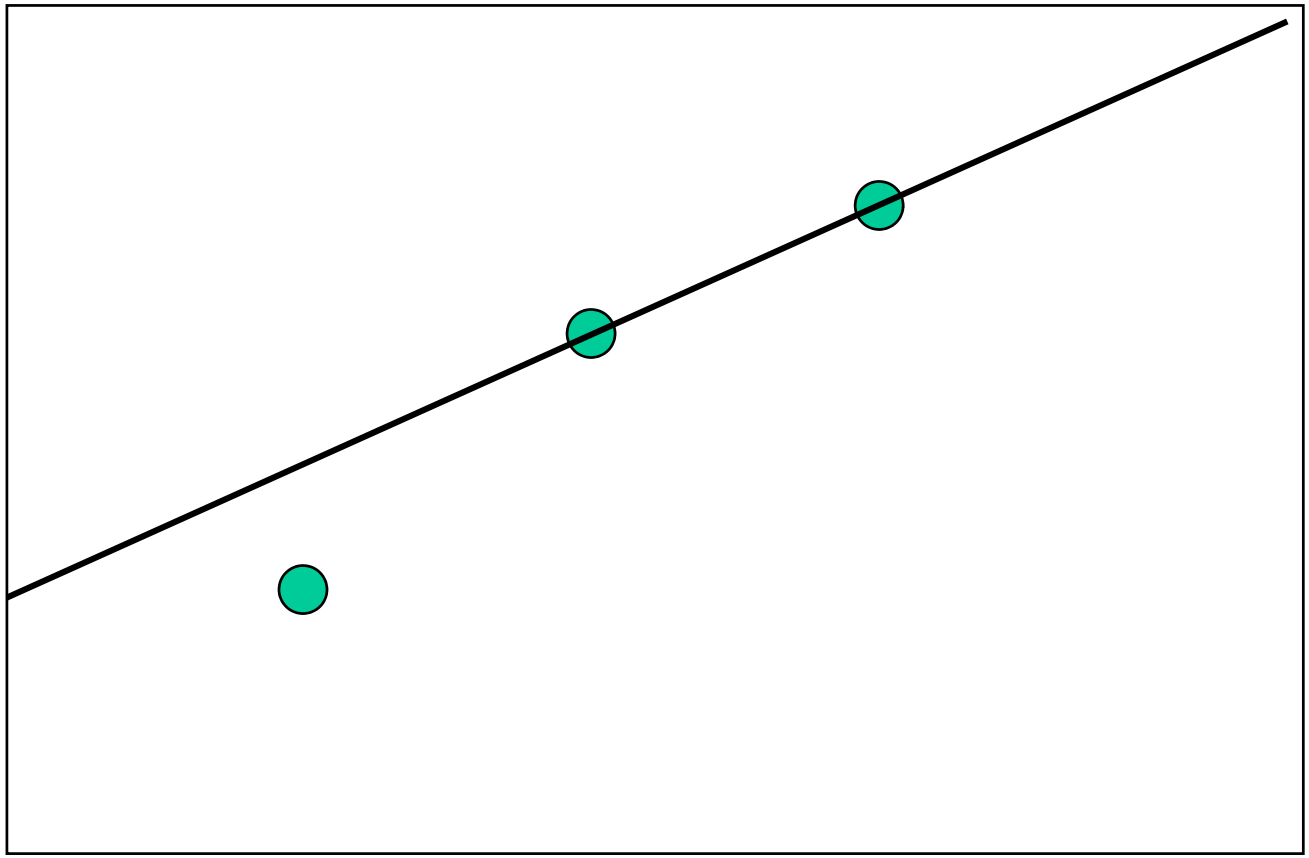
直線当てはめとは



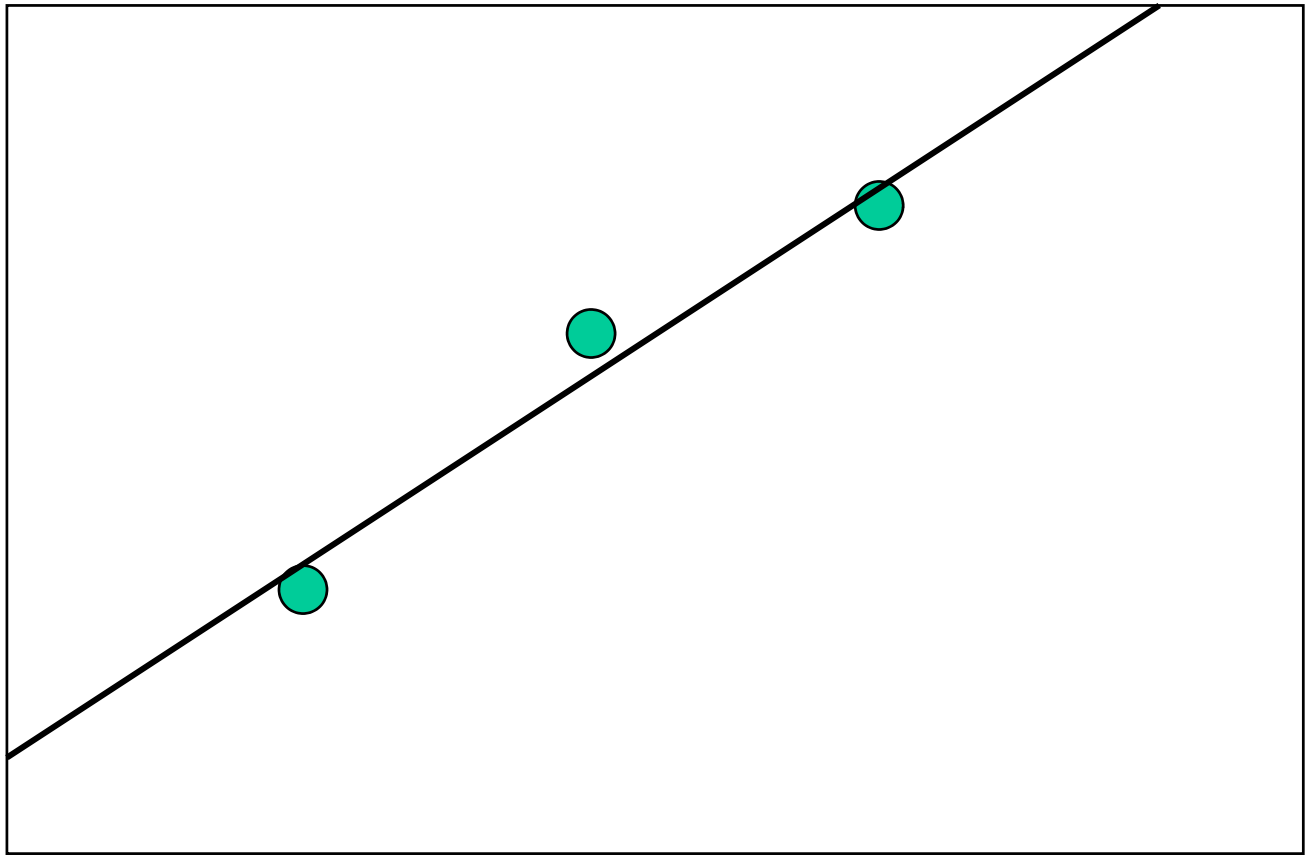
直線当てはめとは



直線当てはめとは

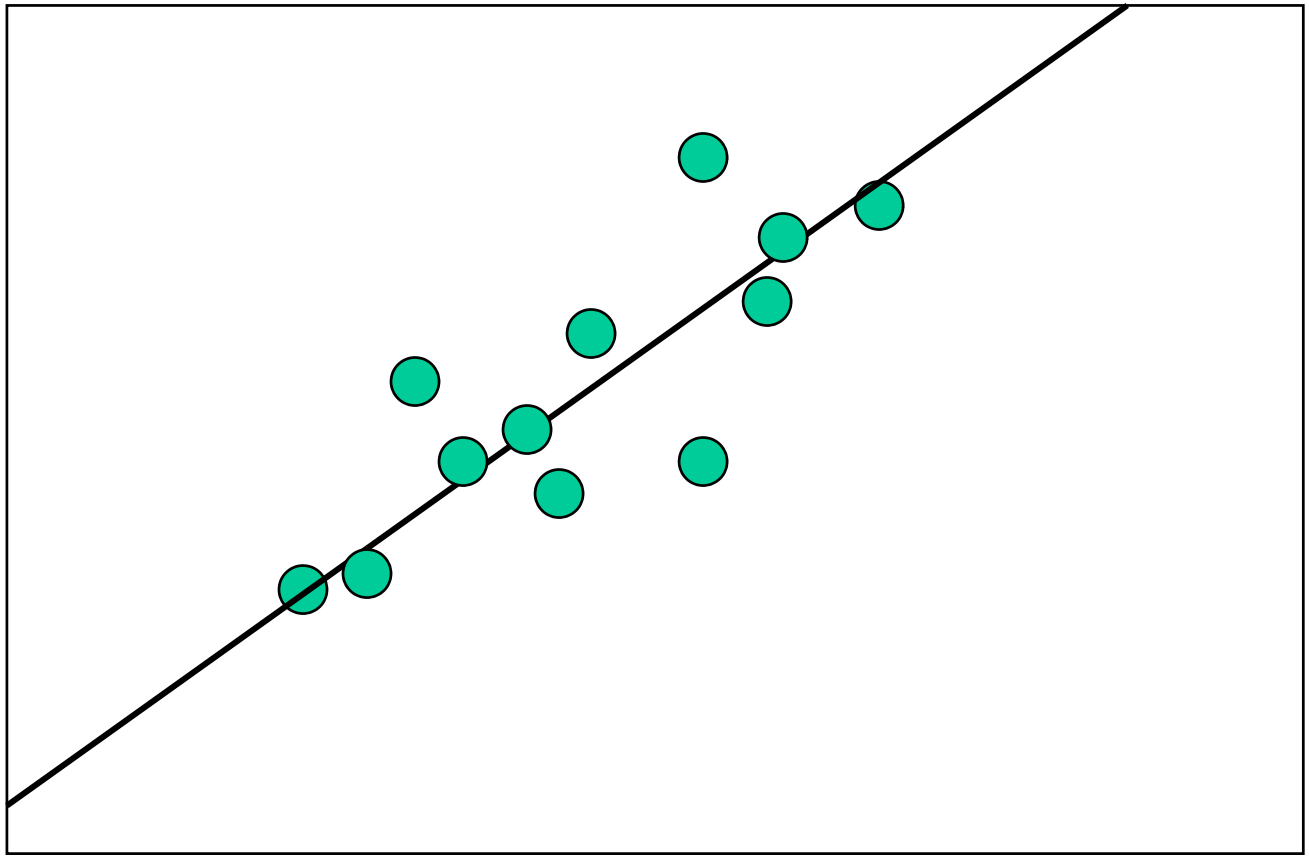


直線当てはめとは



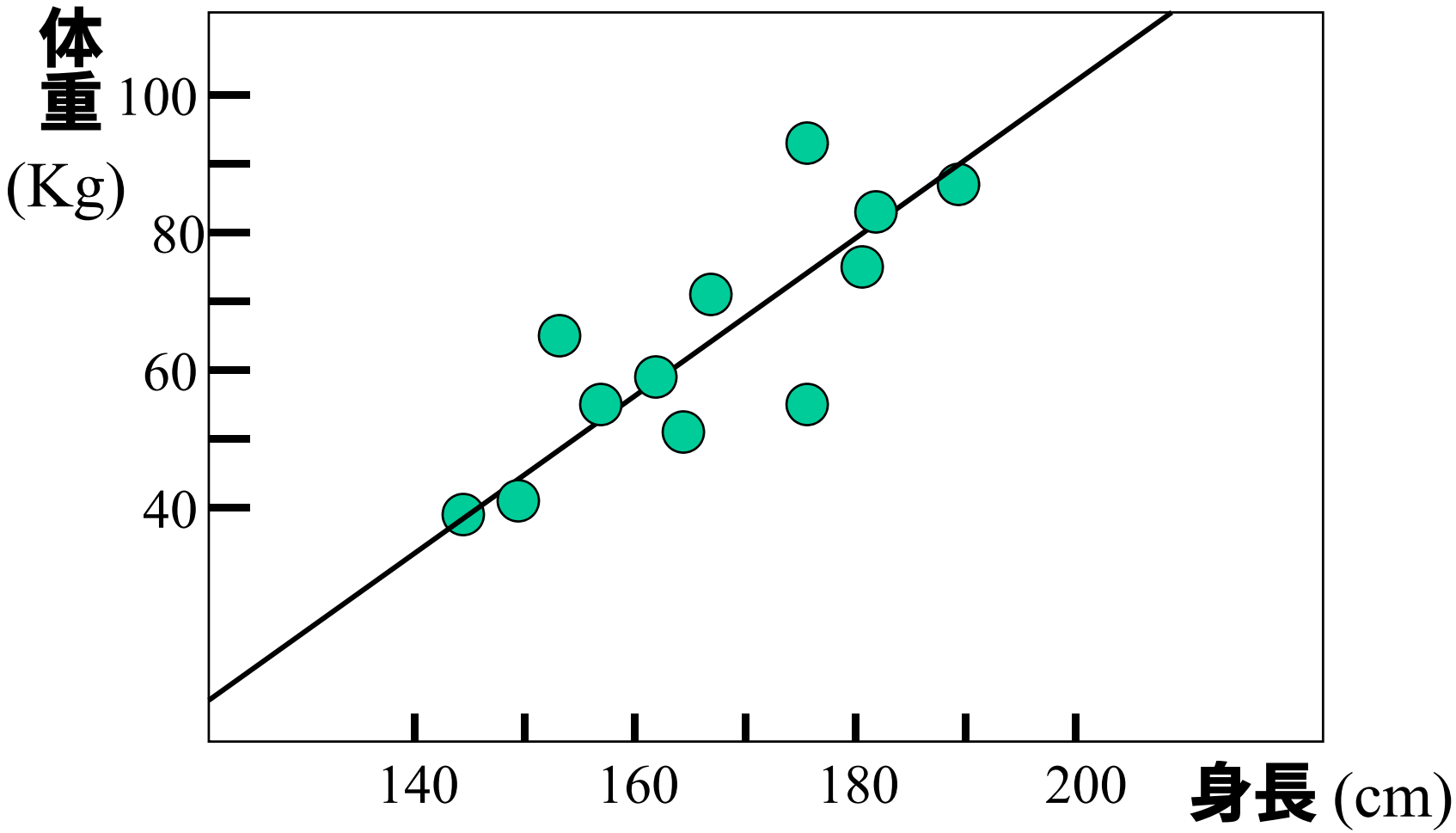
点の数が3以上になれば、直線当てはめを行うためには、
何らかの妥協が必要である。

直線当てはめとは

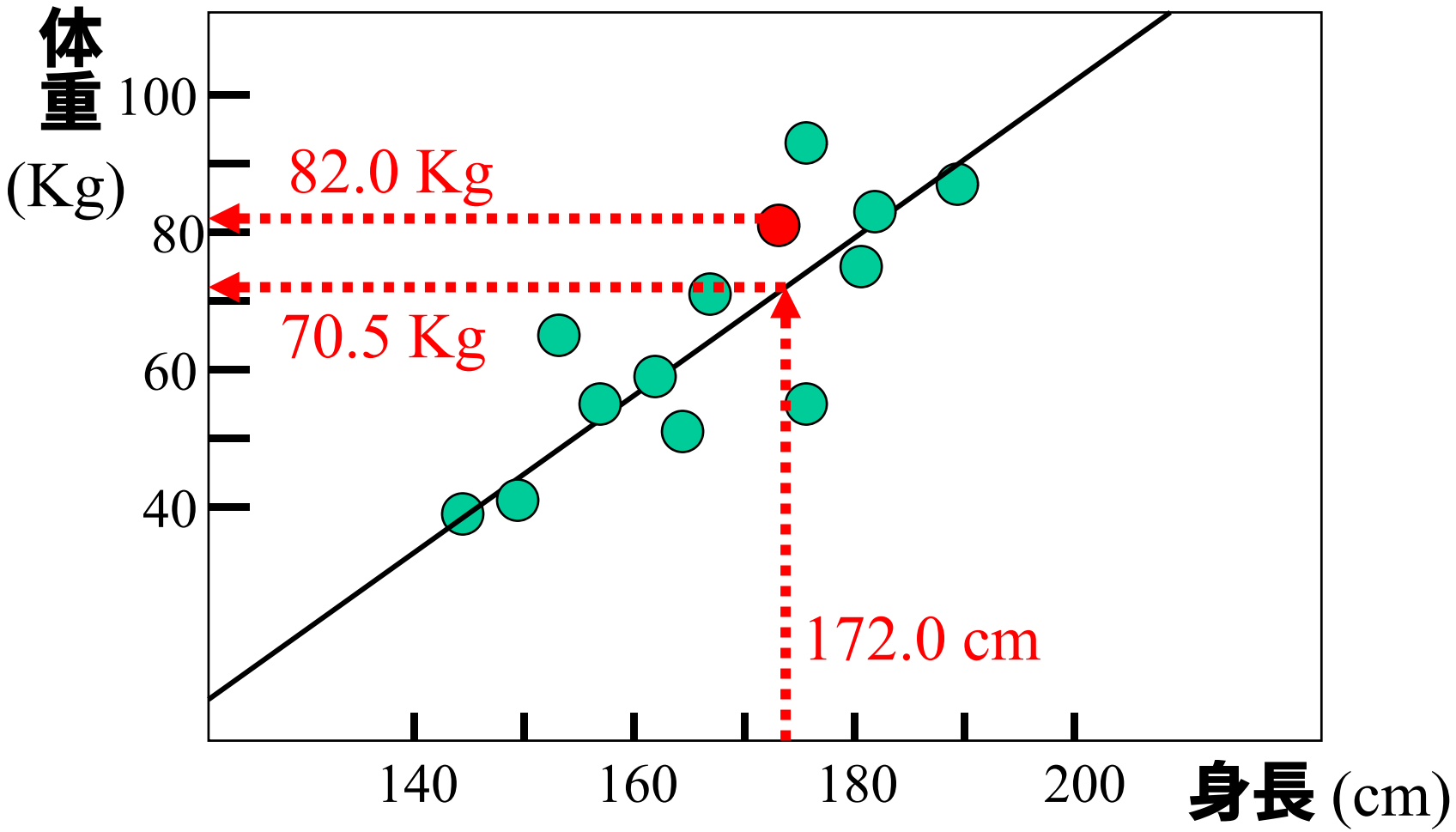


点の数が多くなって、妥当な直線当てはめを行うための基準が必要である。

直線当てはめの応用

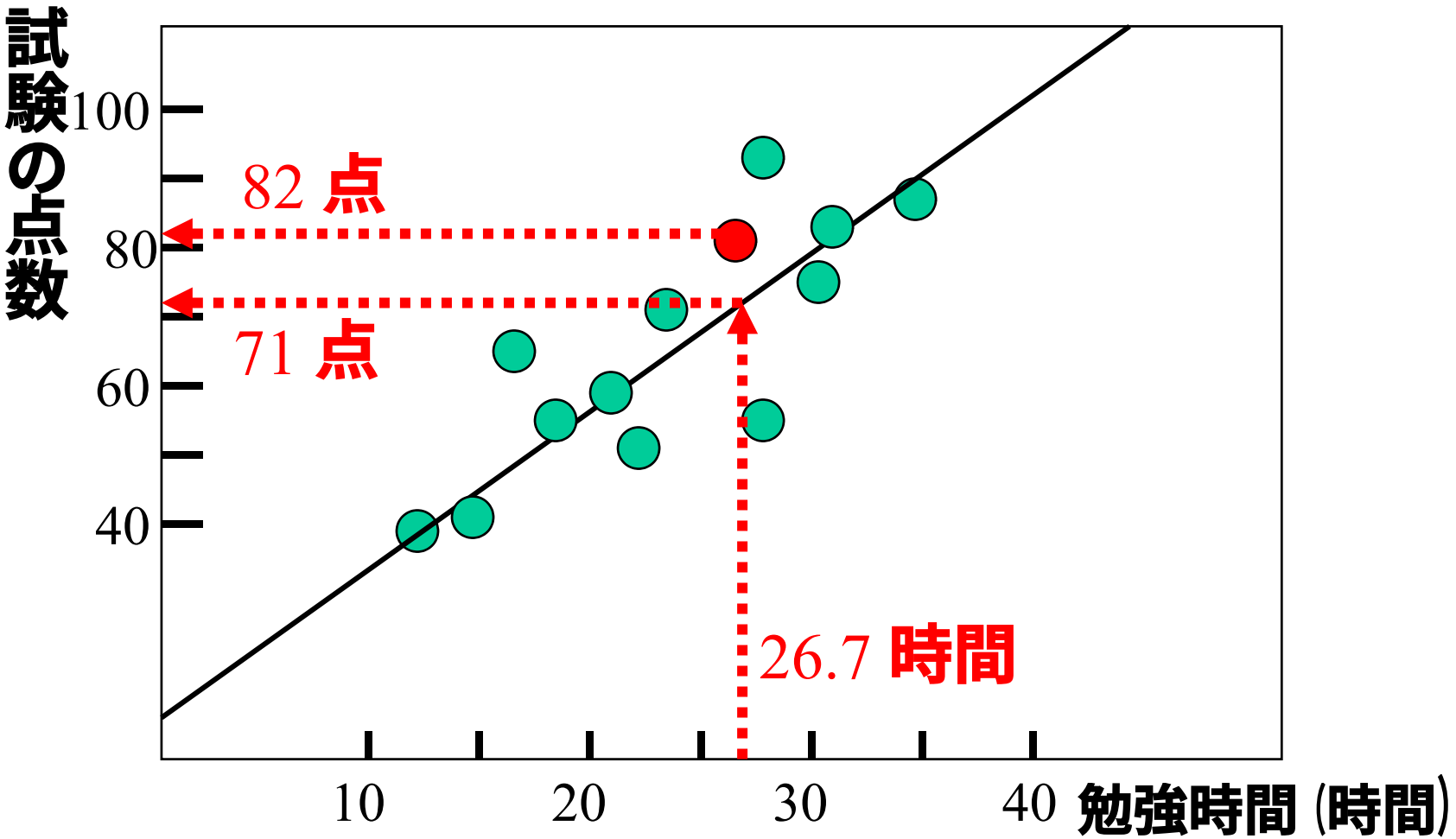


直線当てはめの応用



自分の身長から平均的体重を知ることができる。

直線当てはめの応用

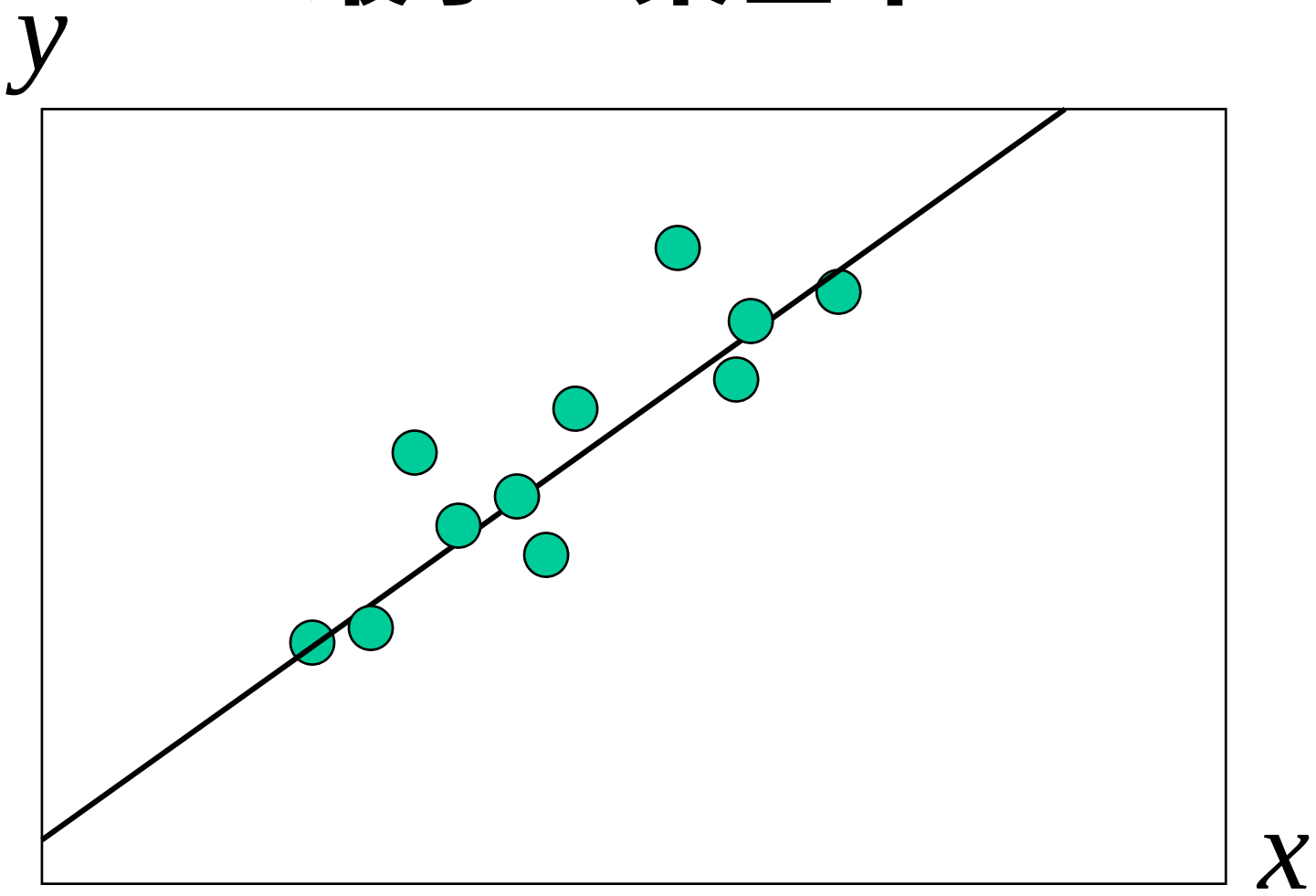


長い時間勉強をすれば、試験で良い点がとれる。

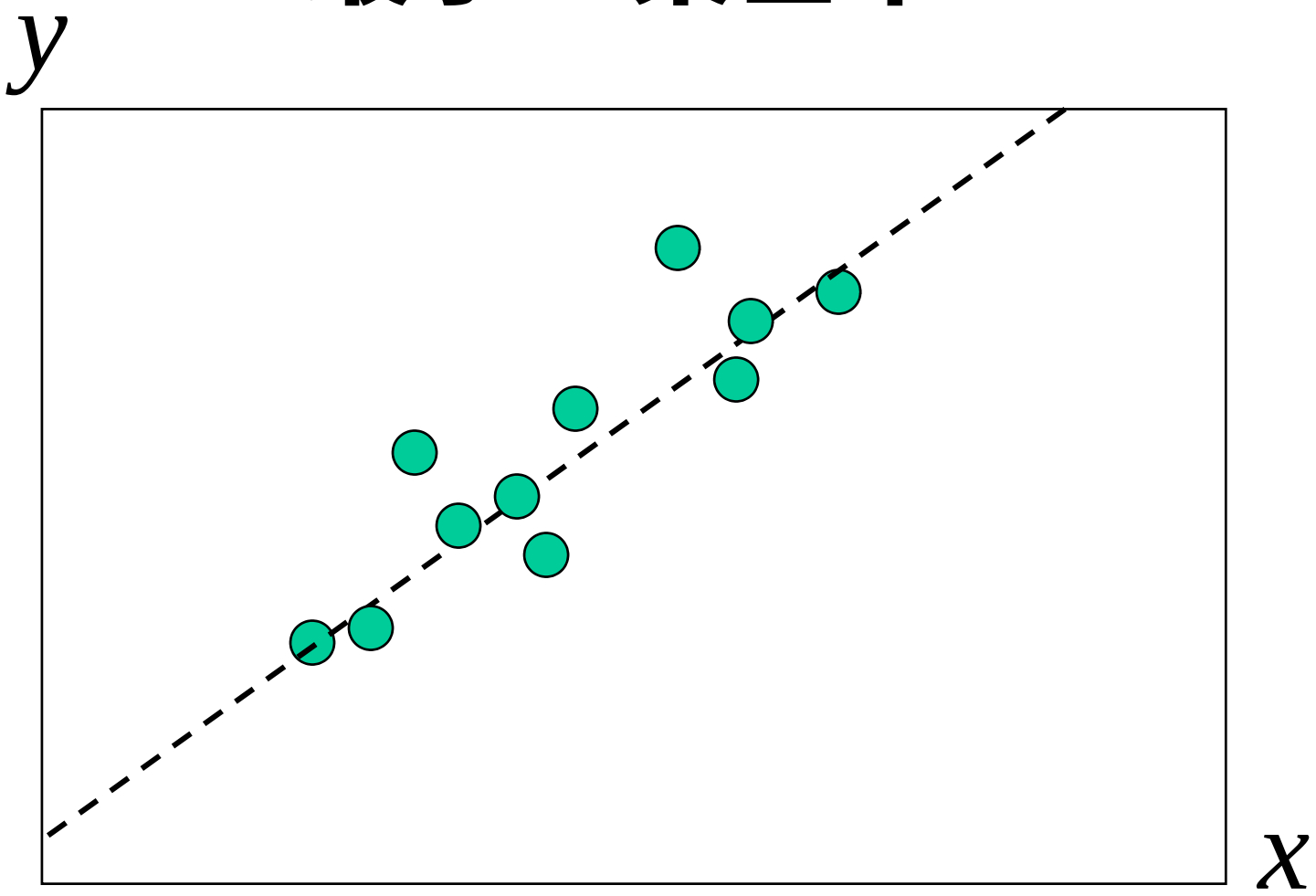
直線当てはめの応用

- 直線当てはめを行う対象としてどのようなものがあるか？　(縦軸と横軸にどのような量をとれば、おもしろいか?／意味があるか？
  - プラズマテレビの値段と画面インチ数
  - 車の値段とエンジン排気量
  - プロ野球選手のホームラン数 (打数) と打点数 (安打数)
  - ある果物の直径と水分含有率
  - 夏の気温とある果物の糖度

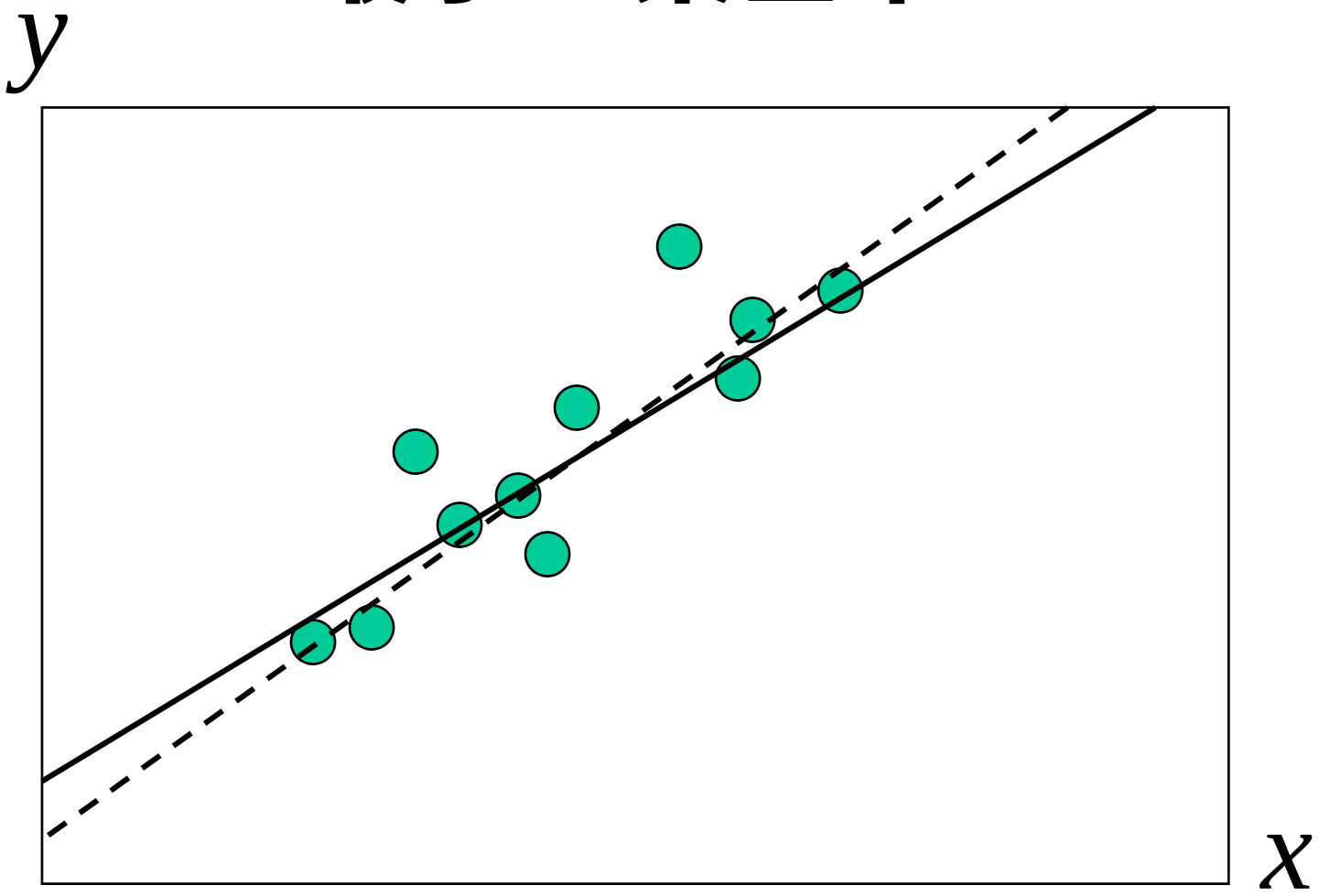
最小二乗基準



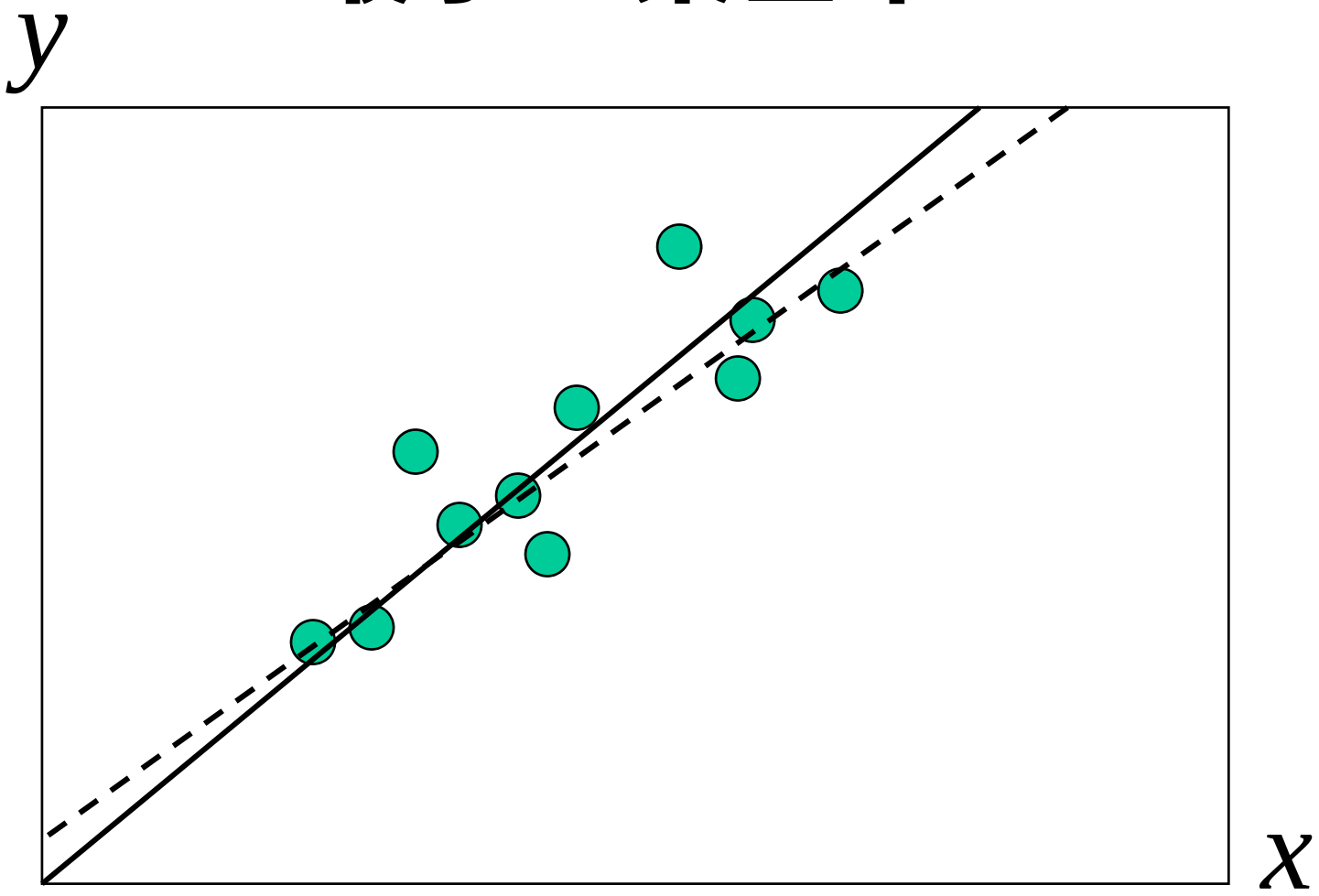
最小二乗基準



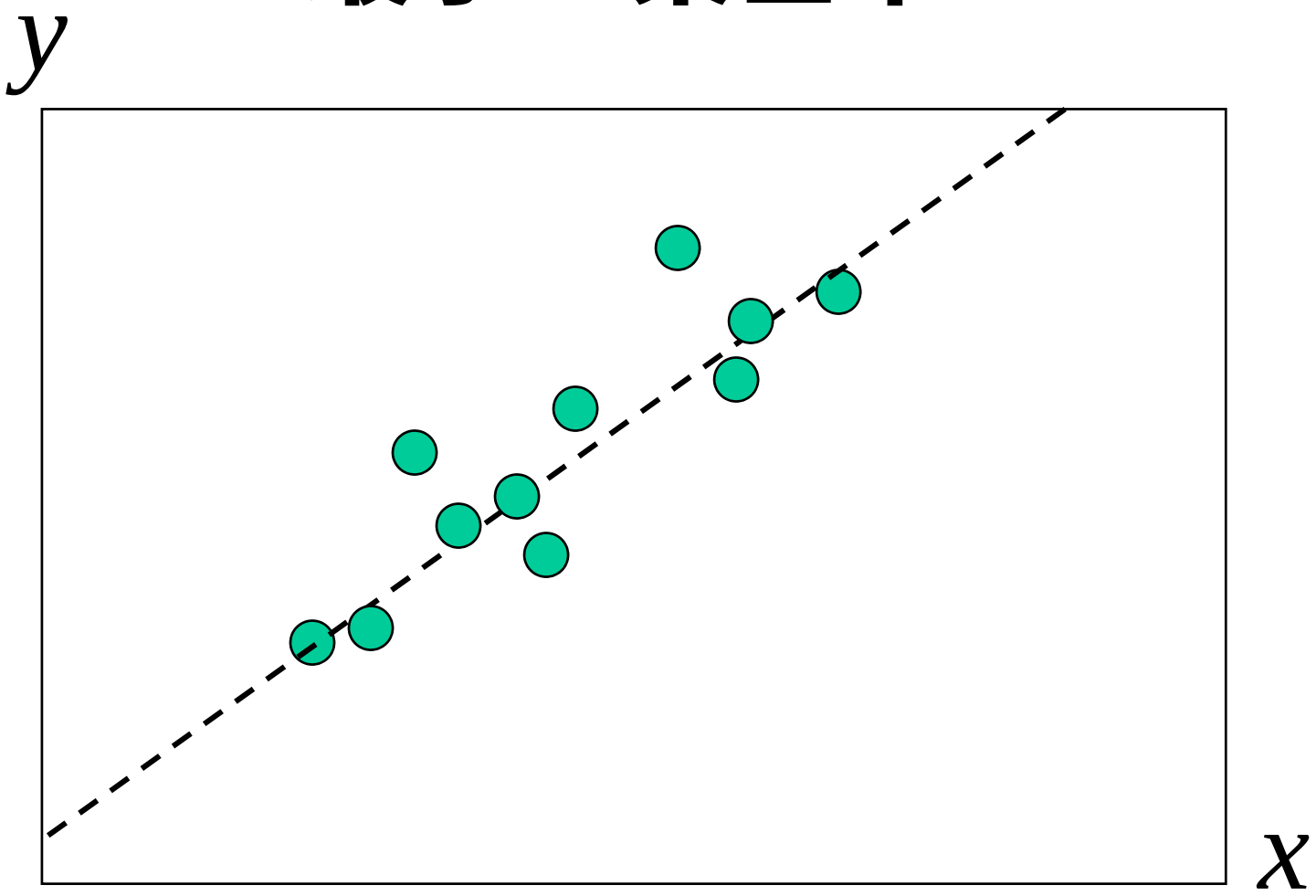
最小二乗基準



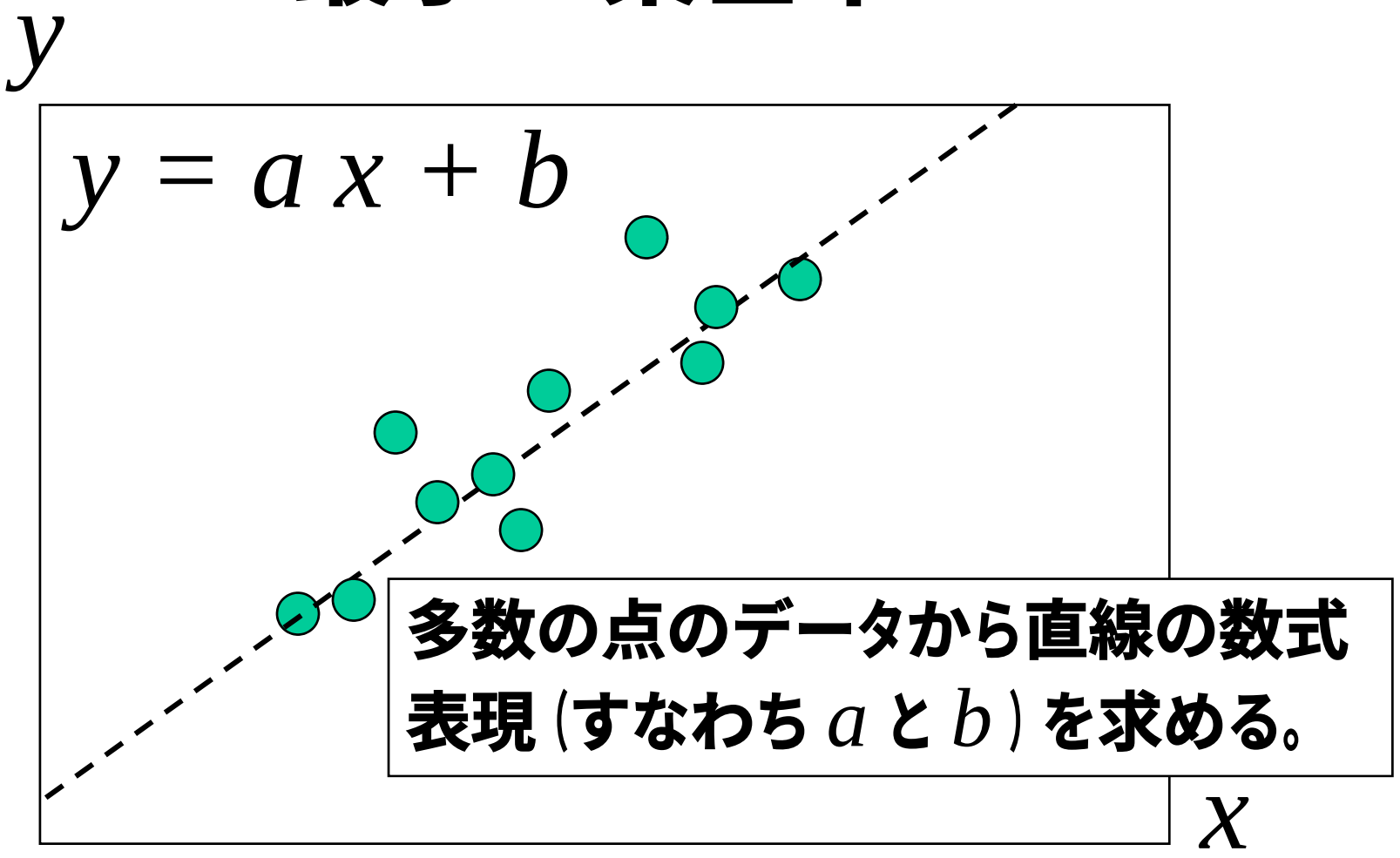
最小二乗基準



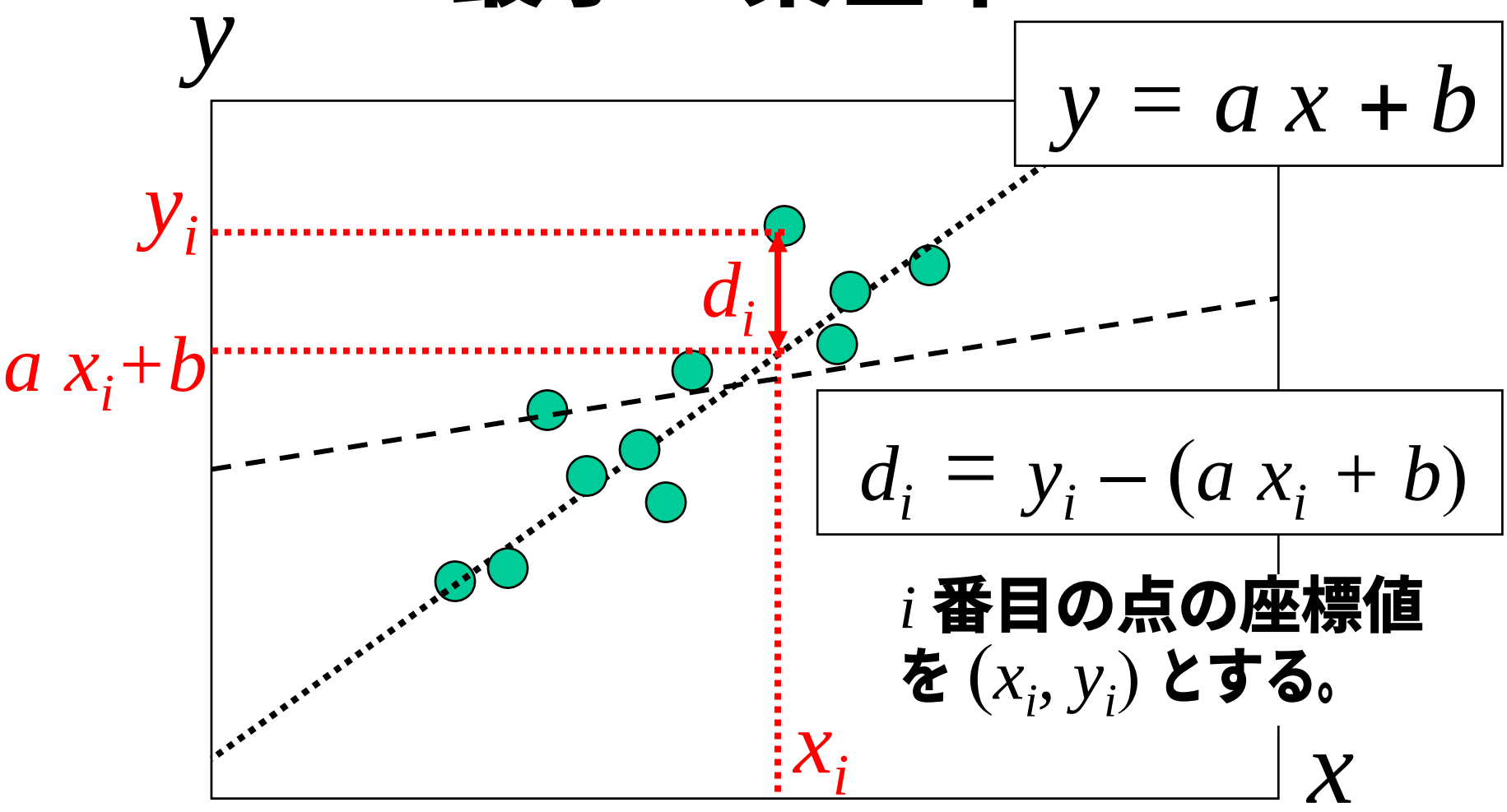
最小二乗基準



最小二乗基準



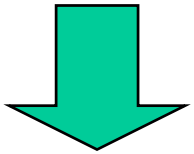
最小二乗基準



$\sum d_i^2 = \sum \{ y_i - (a\,x_i + b) \}^2$ **を最小にする** a, b **を求める。**

最小二乗基準

- すべての点が、直線 $y = a\,x + b$ の上になるべく〈のっている。



$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \{ y_i - (a\,x_i + b) \}^2$ を最小にする a, b を求める。

n 個の点があり、i 番目の点の座標値を (x_i, y_i) とする。

最小二乗基準

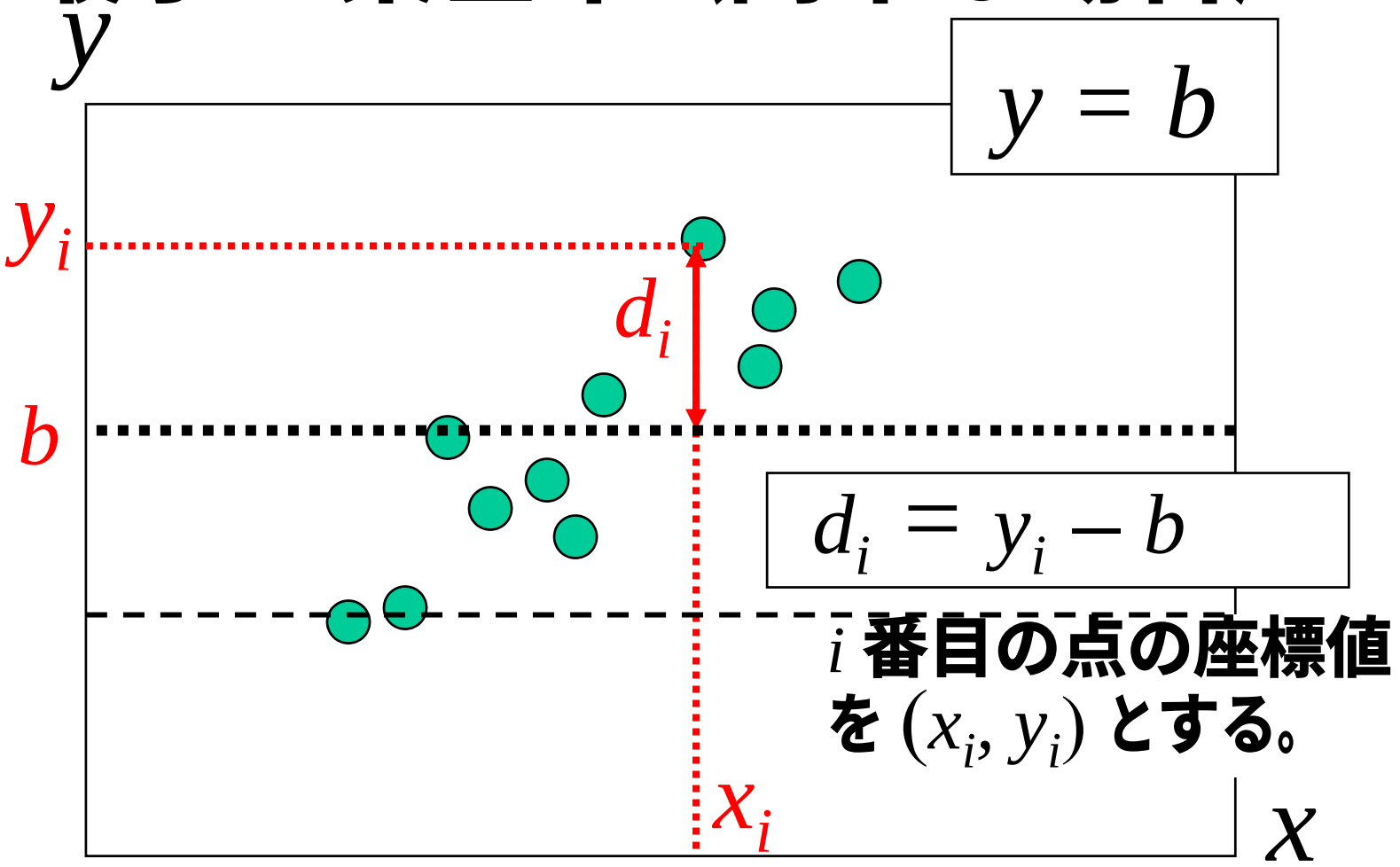
$f(a,b)=\sum_{i=1}^{n}\{y_i-(a\,x_i+b)\}^2$ を最小にする $a,\ b$ を求める。

$f(a,b)$ を最小にする $a,\ b$ を求める。

$$\frac{\partial}{\partial a}f(a,b)=0 \qquad \frac{\partial}{\partial b}f(a,b)=0$$

$a,\ b$ に関する偏微分が 0 になる。

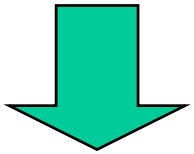
最小二乗基準（簡単な場合）



$\sum d_i^2 = \sum \{y_i - b\}^2$ を最小にする b を求める。

最小二乗基準 (簡単な場合)

- すべての点が、直線 $y = b$ の上になるべくのっている。



$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \{ y_i - b \}^2$ を最小にする b を求める。

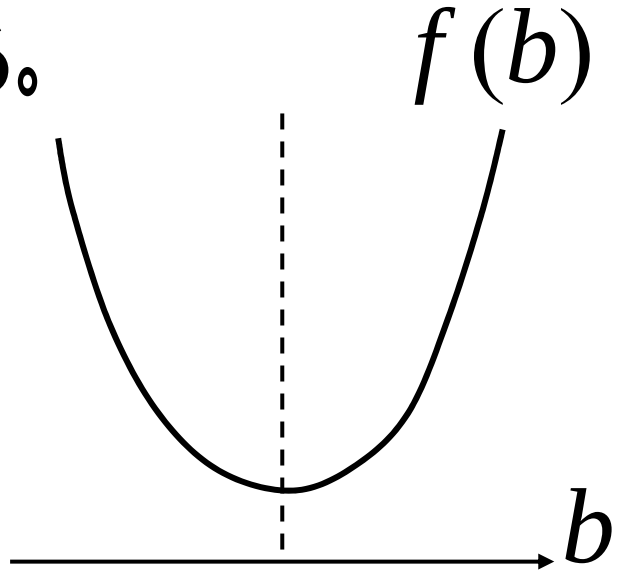
n 個の点があり、i 番目の点の座標値を (x_i, y_i) とする。

最小二乗基準 (簡単な場合)

$f(b)=\sum_{i=1}^{n}\{y_i-b\}^2$ を最小にする b を求める。

$f(b)$ を最小にする b を求める。

$$\frac{\mathrm{d}}{\mathrm{d}b}f(b)=0$$



b に関する微分が 0 になる。

高校数学の復習

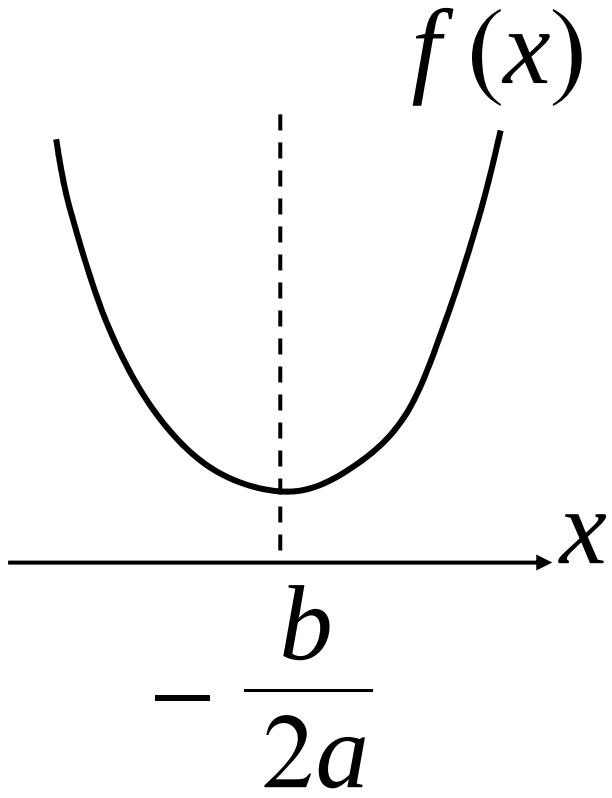
- **2次関数 の最小値問題**

$$f(x) = ax^2 + bx + c$$

最小値をとる条件

$$f'(x) = 2ax + b = 0$$

$$x = -\frac{b}{2a}$$

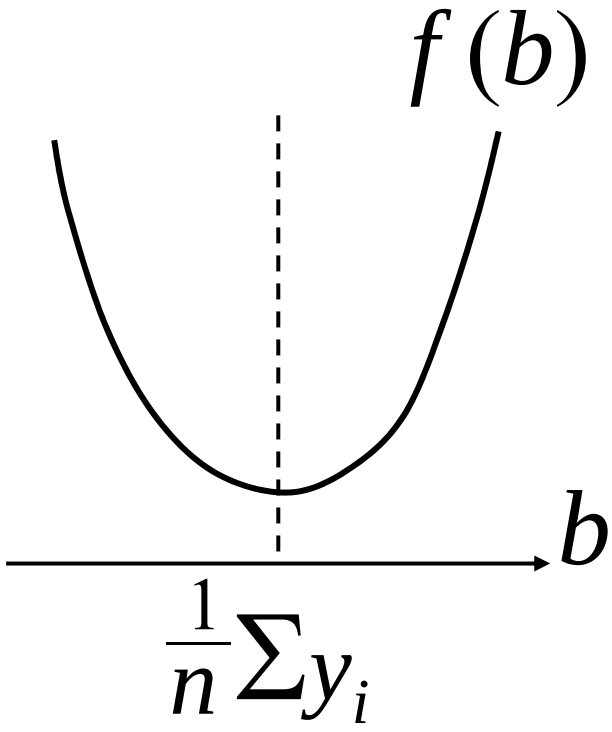


最小二乗基準 (簡単な場合)

$$f(b) = \sum_{i=1}^{n} \{ y_i - b \}^2$$

$$= \sum \{ y_i{}^2 - 2 y_i b + b^2 \}$$

$$= \sum y_i{}^2 - 2b \sum y_i + nb^2$$

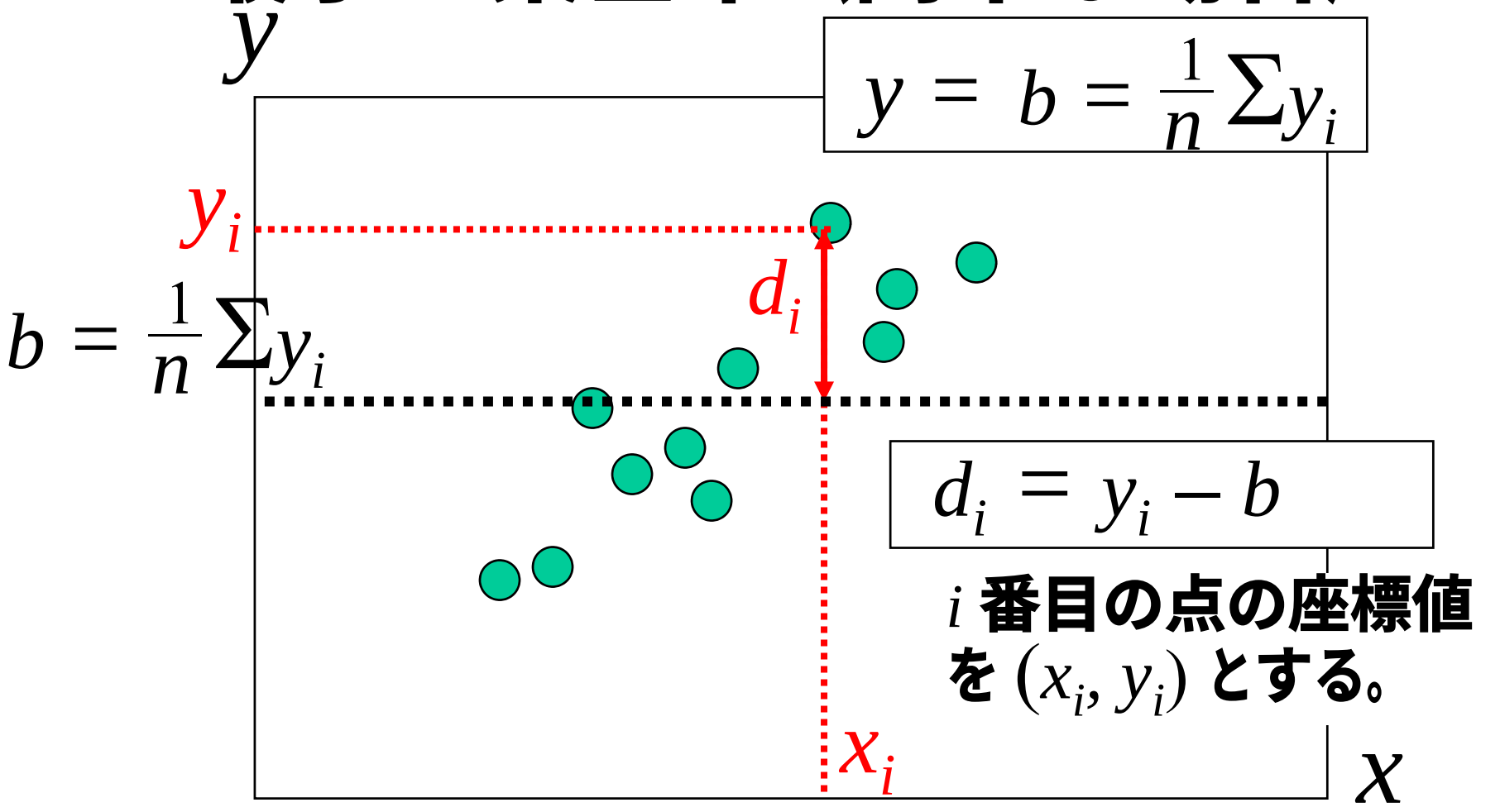


最小値をとる条件

$$\frac{\mathrm{d}}{\mathrm{d}b} f(b) = 2nb - 2\sum y_i = 2(nb - \sum y_i) = 0$$

$$b = \frac{1}{n} \sum y_i \quad (b \text{ は } y_i \text{ の平均値})$$

最小二乗基準（簡単な場合）



$\sum d_i^2 = \sum \{ y_i - b \}^2$ を最小にする b は $\frac{1}{n}\sum y_i$

最小二乗基準

$f(a,b)=\sum_{i=1}^{n}\{y_i-(a\,x_i+b)\}^2$ を最小にする $a,\ b$ を求める。

$f(a,b)$ を最小にする $a,\ b$ を求める。

$$\frac{\partial}{\partial a}f(a,b)=0 \qquad \frac{\partial}{\partial b}f(a,b)=0$$

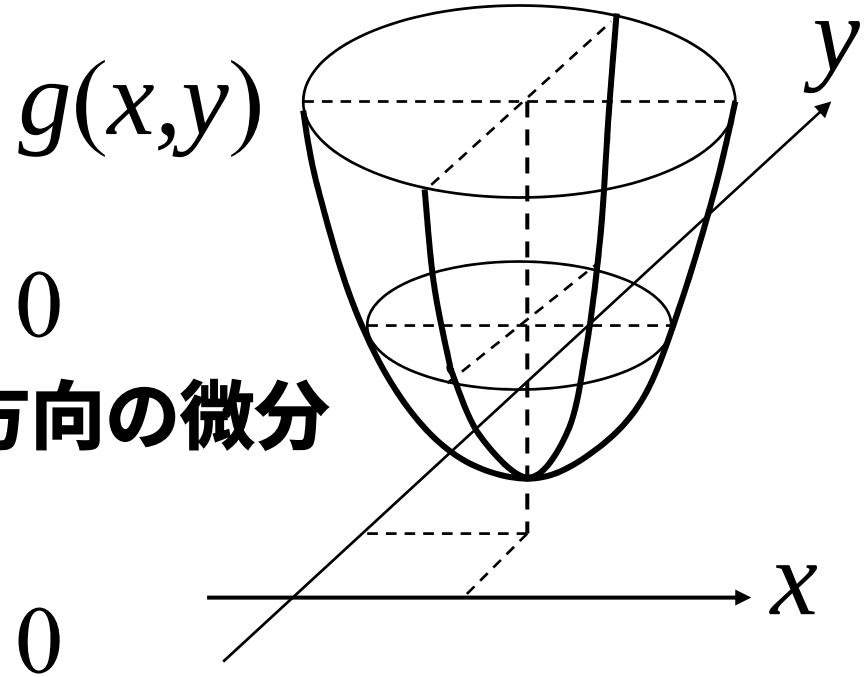
$a,\ b$ に関する偏微分が 0 になる。

大学の数学：多変数関数の場合

- 2次関数 の最小値問題

$$g(x,y) = ax^2 + bxy + cy^2 + dx + ey + f$$

最小値をとる条件



$$\frac{\partial}{\partial x} g(x,y) = g_x(x,y) = 0$$

xに関する偏微分：x軸方向の微分

$$\frac{\partial}{\partial y} g(x,y) = g_y(x,y) = 0$$

yに関する偏微分：y軸方向の微分

大学の数学：多変数関数の場合

$$g(x,y) = ax^2 + bxy + cy^2 + dx + ey + f$$

$$g(y) = cy^2 + (bx + e)y + (ax^2 + dx + f)$$

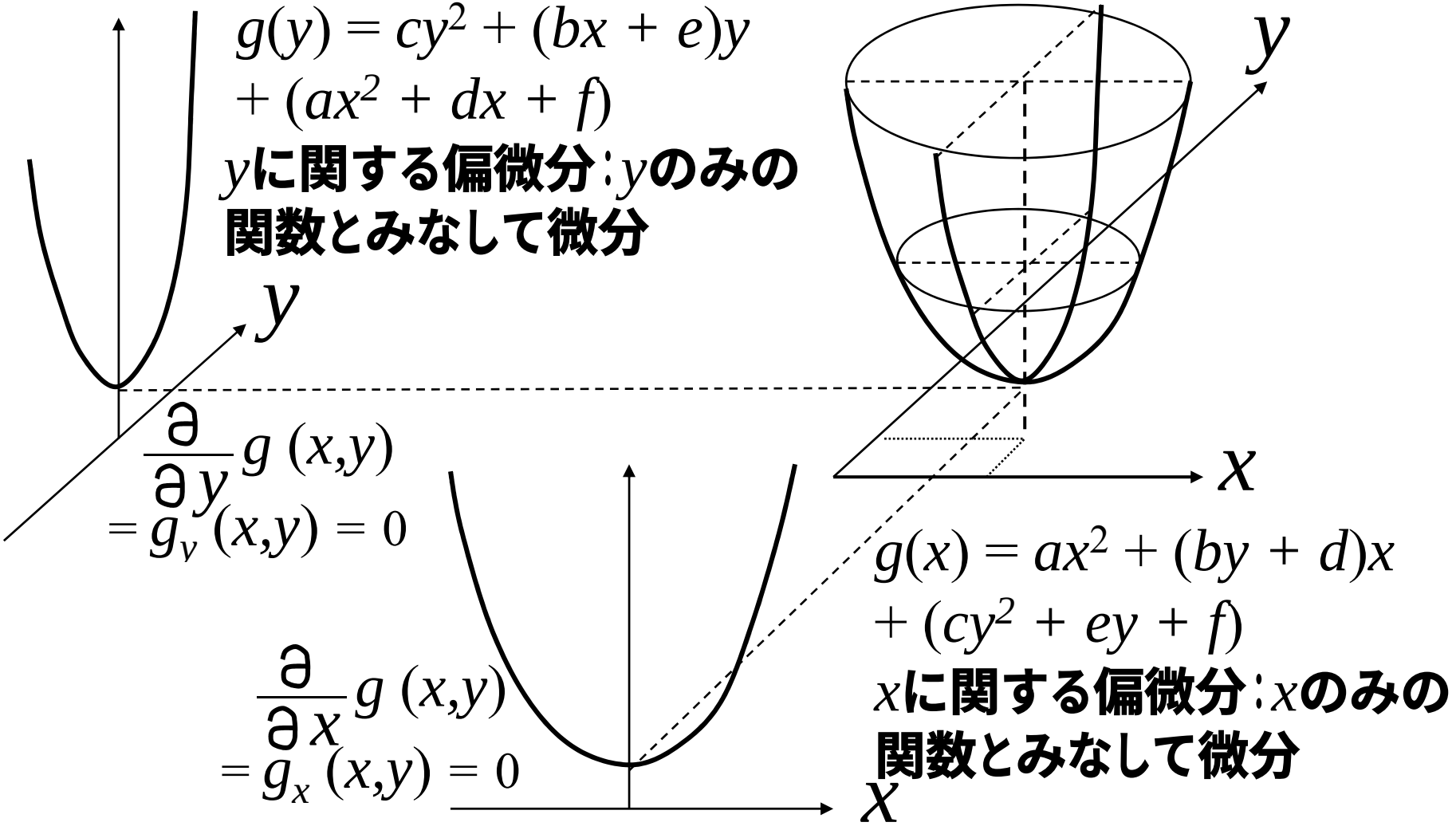
yに関する偏微分：yのみの関数とみなして微分

$$\frac{\partial}{\partial y} g(x,y) = g_y(x,y) = 0$$

$$g(x) = ax^2 + (by + d)x + (cy^2 + ey + f)$$

xに関する偏微分：xのみの関数とみなして微分

$$\frac{\partial}{\partial x} g(x,y) = g_x(x,y) = 0$$

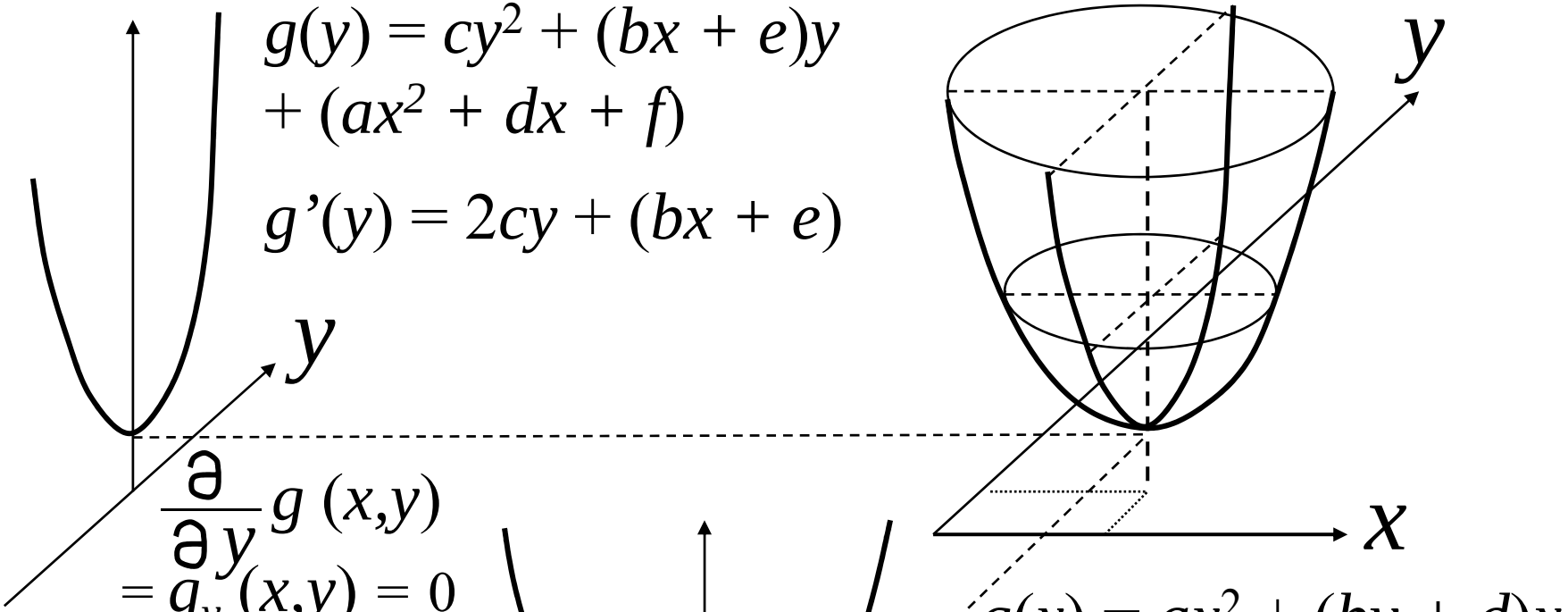


偏微分

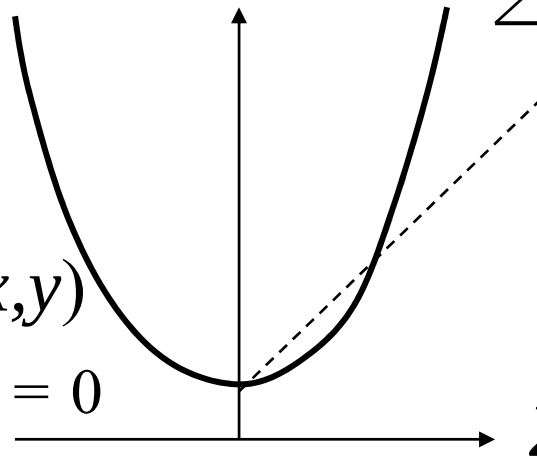
$$g(x,y) = ax^2 + bxy + cy^2 + dx + ey + f$$

$$g(y) = cy^2 + (bx + e)y + (ax^2 + dx + f)$$

$$g'(y) = 2cy + (bx + e)$$



$$\frac{\partial}{\partial y} g(x,y) = g_y(x,y) = 0$$



$$\frac{\partial}{\partial x} g(x,y) = g_x(x,y) = 0$$

$$g(x) = ax^2 + (by + d)x + (cy^2 + ey + f)$$

$$g'(x) = 2ax + (by + d)$$

大学の数学：多変数関数の場合

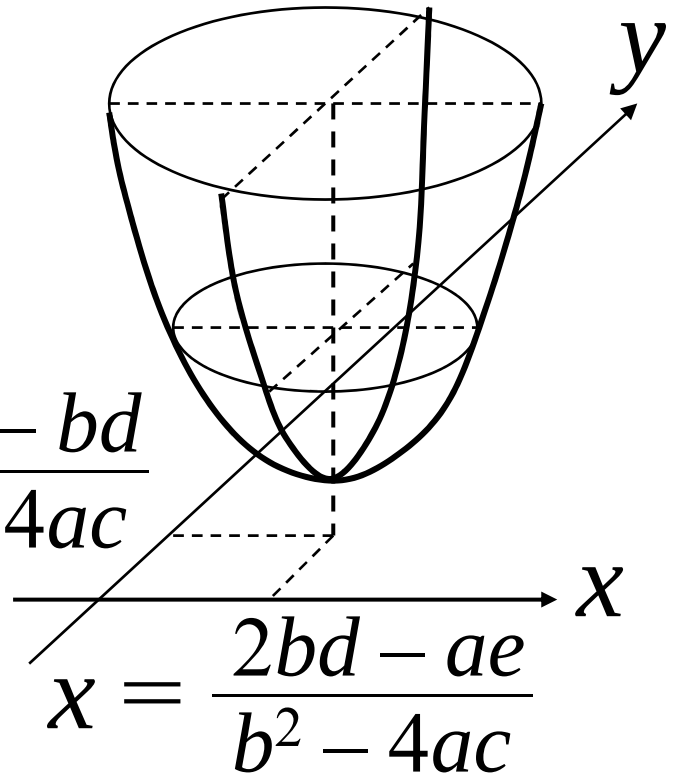
- 2次関数 の最小値問題

$$g(x,y) = ax^2 + bxy + cy^2 + dx + ey + f$$

最小値をとる条件

$$g_x(x,y) = 2ax + by + d = 0$$

$$g_y(x,y) = 2cy + bx + e = 0$$



$$y = \frac{2ae - bd}{b^2 - 4ac}$$

$$x = \frac{2bd - ae}{b^2 - 4ac}$$

偏微分の演習問題

- 以下を x, y それぞれについて偏微分せよ。

$$f(x, y) = 3x^2 + 5xy + 2y^2 - 5x - 8y + 6$$

$$f(x, y) = 3x^2a^2 + 5xya - 2cx - 2b^2y + 6$$

最小二乗基準

$f(a,b)=\sum_{i=1}^{n}\{y_i-(a\,x_i+b)\}^2$ を最小にする $a,\ b$ を求める。

$f(a,b)$ を最小にする $a,\ b$ を求める。

$$\frac{\partial}{\partial a}f(a,b)=0 \qquad \frac{\partial}{\partial b}f(a,b)=0$$

$a,\ b$ に関する偏微分が 0 になる。

演習問題

A. $f(a,b)=\sum_{i=1}^{n}\{y_i-(a\,x_i+b)\}^2$ をa, b に関して偏微分せよ。

B. $f(a,b)$のa, b に関する偏微分を、それぞれ $f_a(a,b)$, $f_b(a,b)$ とすると、$f(a,b)$ を最小にする条件は、以下で与えられる。

$$f_a(a,b)=0$$
$$f_b(a,b)=0$$

この2つの条件式から、未知数を a, b とする2元連立1次方程式がえれらる。この連立方程式を行列を用いて表現せよ。

演習問題Aの解答

$$f(a,b) = \sum \{ y_i - (a\,x_i + b) \}^2$$

$$= \sum \{ y_i^{\,2} - 2y_i (a\,x_i + b) + (a\,x_i + b)^2 \}$$

$$= \sum \{ y_i^{\,2} - 2ax_iy_i - 2by_i + a^2x_i^{\,2} + 2abx_i + b^2 \}$$

（各項を別々にする）

$$= \sum y_i^{\,2} - \sum 2ax_iy_i - \sum 2by_i + \sum a^2x_i^{\,2} + \sum 2abx_i + \sum b^2$$

（x_i と y_i 以外の変数は、Σの外に出せる。$n = \sum 1$）

$$= \sum y_i^{\,2} - 2a\sum x_iy_i - 2b\sum y_i + a^2\sum x_i^{\,2} + 2ab\sum x_i + nb^2$$

$$= a^2\sum x_i^{\,2} + 2ab\sum x_i + nb^2 - 2a\sum x_iy_i - 2b\sum y_i + \sum y_i^{\,2}$$

演習問題Aの解答 (続き1)

$$f(a,b) = \sum \{ y_i - (a\, x_i + b) \}^2$$

$$= a^2 \sum x_i^2 + 2ab \sum x_i + nb^2 - 2a \sum x_i y_i - 2b \sum y_i + \sum y_i^2$$

$$= a^2 \sum x_i^2 + 2ab \sum x_i - 2a \sum x_i y_i + nb^2 - 2b \sum y_i + \sum y_i^2$$

$$= a^2 \sum x_i^2 + 2a(b \sum x_i - \sum x_i y_i) + nb^2 - 2b \sum y_i + \sum y_i^2$$

$$f_a(a,b) = 2a \sum x_i^2 + 2(b \sum x_i - \sum x_i y_i)$$

演習問題Aの解答（続き2）

$$f(a,b) = \sum \{ y_i - (a\, x_i + b) \}^2$$

$$= a^2 \sum x_i^2 + 2ab \sum x_i + nb^2 - 2a \sum x_i y_i - 2b \sum y_i + \sum y_i^2$$

$$= nb^2 + 2ab \sum x_i - 2b \sum y_i + a^2 \sum x_i^2 - 2a \sum x_i y_i + \sum y_i^2$$

$$= nb^2 + 2b(a \sum x_i - \sum y_i) + a^2 \sum x_i^2 - 2a \sum x_i y_i + \sum y_i^2$$

$$f_b(a,b) = 2nb + 2(a \sum x_i - \sum y_i)$$

演習問題Bの解答

$$f_a(a,b) = 2a\Sigma x_i^2 + 2b\Sigma x_i - 2\Sigma x_i y_i = 0$$

$$f_b(a,b) = 2a\Sigma x_i + 2nb - 2\Sigma y_i = 0$$

$$\begin{pmatrix} \Sigma x_i^2 & \Sigma x_i \\ \Sigma x_i & n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \Sigma x_i y_i \\ \Sigma y_i \end{pmatrix}$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \Sigma x_i^2 & \Sigma x_i \\ \Sigma x_i & n \end{pmatrix}^{-1} \begin{pmatrix} \Sigma x_i y_i \\ \Sigma y_i \end{pmatrix}$$

行列演算のおさらい

- **行列の足し算**

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} a+e & b+f \\ c+g & d+h \end{pmatrix}$$

- **行列の掛け算　(横列×縦列)**

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e \\ f \end{pmatrix} = \begin{pmatrix} ae+bf \\ ce+df \end{pmatrix}$$

- **単位行列**　$a \times 1 = a$

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix} \qquad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

- **逆行列**　$a^{-1} \times a = \frac{1}{a} \times a = 1$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

行列演算のおさらい

- **逆行列**

$$a^{-1} \times a = \frac{1}{a} \times a = 1$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} e \\ f \end{pmatrix}$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} \begin{pmatrix} e \\ f \end{pmatrix}$$

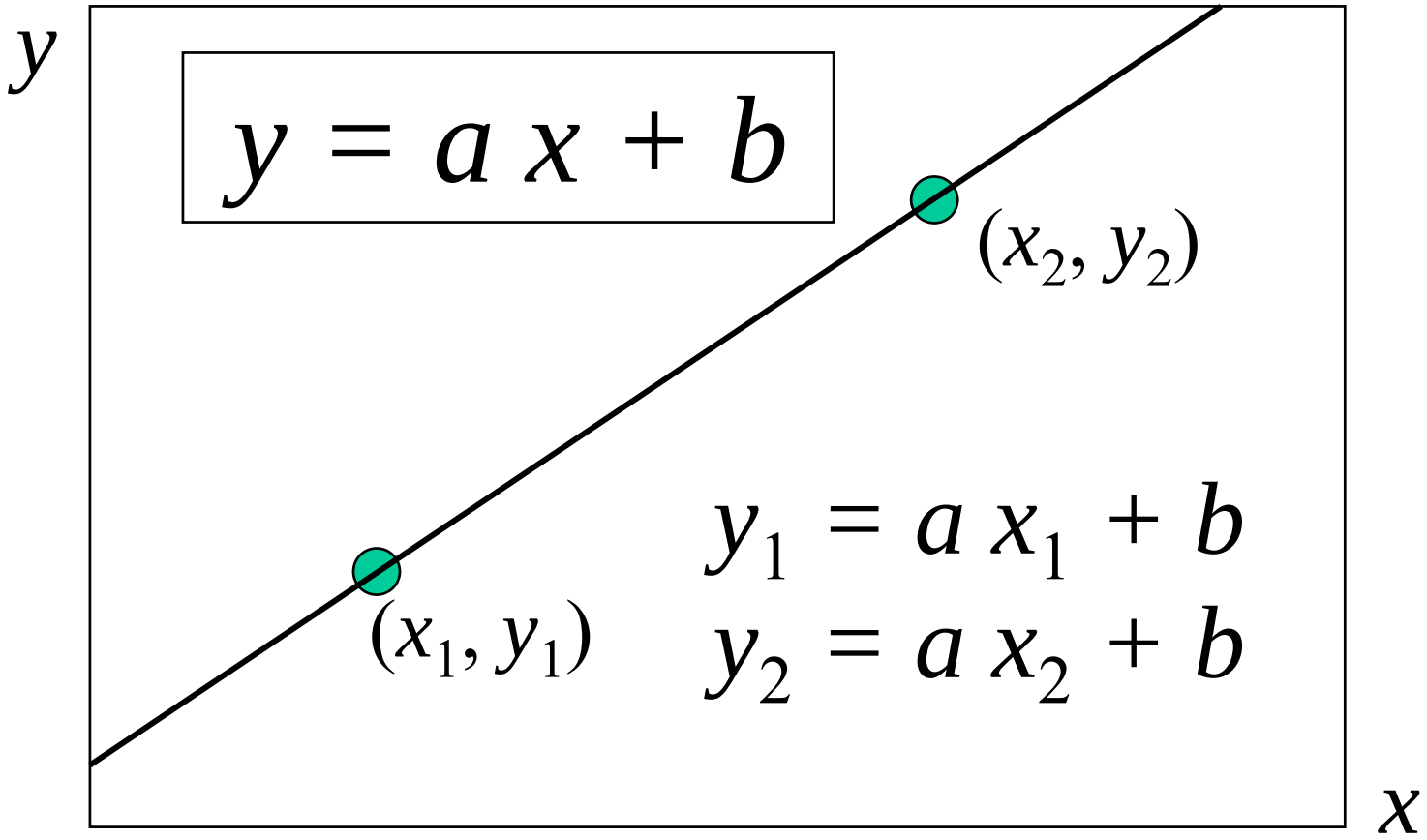
$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} \begin{pmatrix} e \\ f \end{pmatrix}$$

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} \begin{pmatrix} e \\ f \end{pmatrix}$$

最小二乗法の別の定式化

- 一般逆行列（擬逆行列）による定式化
  - 最小二乗基準を偏微分して導かれた最小二乗条件を、行列の式変形だけで導く。

直線当てはめ:2点の場合



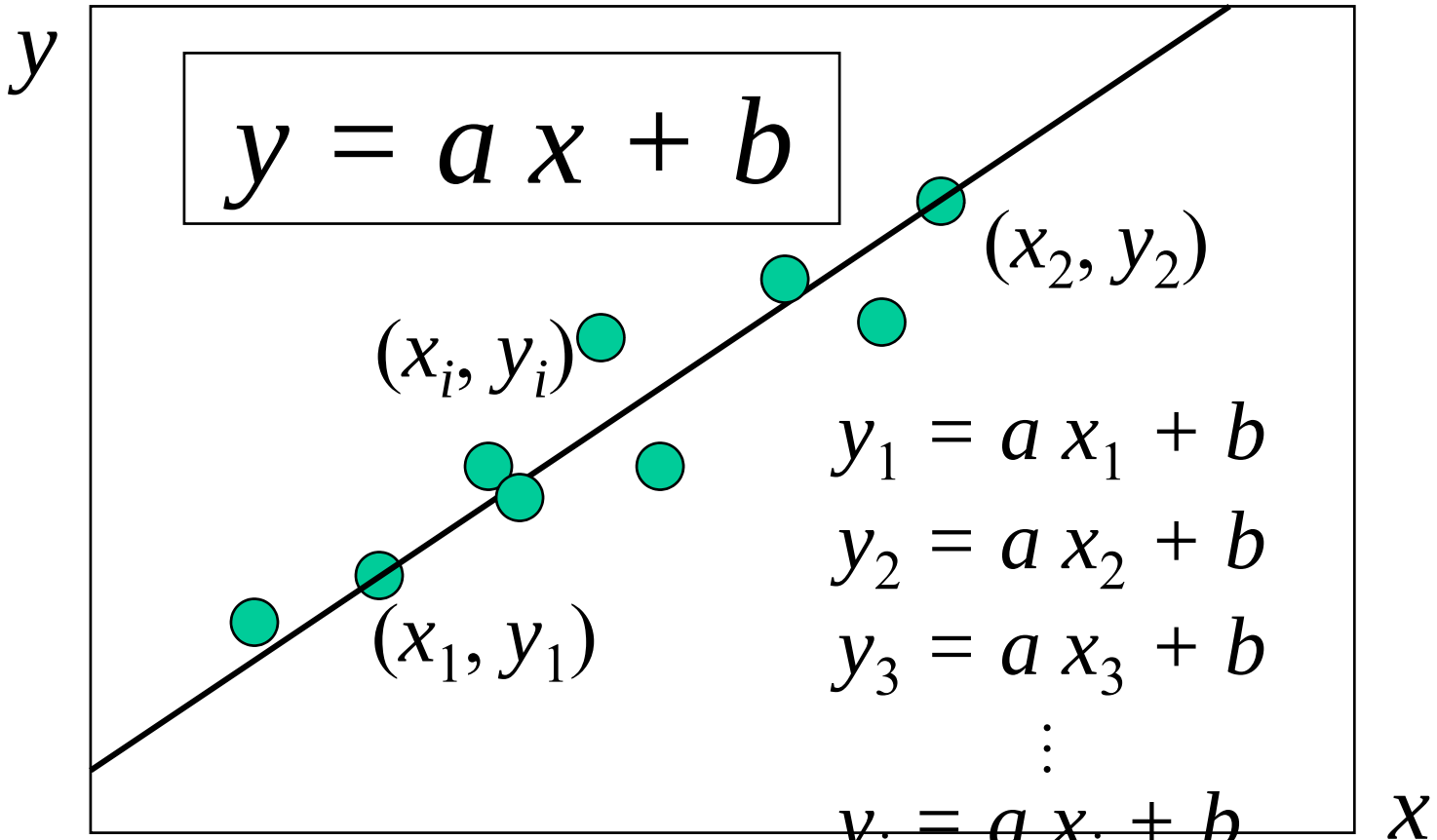
直線当てはめの定式化
2点の場合

$$y_1 = a\,x_1 + b$$
$$y_2 = a\,x_2 + b$$

$$\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

直線当てはめ：n点の場合



$$y_1 = a\,x_1 + b$$
$$y_2 = a\,x_2 + b$$
$$y_3 = a\,x_3 + b$$
$$\vdots$$
$$y_i = a\,x_i + b$$
$$\vdots$$
$$y_n = a\,x_n + b$$

最小二乗法の異なる定式化
一般逆行列 (擬逆行列)

$$
\begin{aligned}
y_1 &= a\,x_1 + b \\
y_2 &= a\,x_2 + b \\
y_3 &= a\,x_3 + b \\
&\vdots \\
y_i &= a\,x_i + b \\
&\vdots \\
y_n &= a\,x_n + b
\end{aligned}
\quad\Rightarrow\quad
\begin{aligned}
a\,x_1 + b &= y_1 \\
a\,x_2 + b &= y_2 \\
a\,x_3 + b &= y_3 \\
&\vdots \\
a\,x_i + b &= y_i \\
&\vdots \\
a\,x_n + b &= y_n
\end{aligned}
\quad\Rightarrow\quad
\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ \vdots & \\ x_i & 1 \\ \vdots & \\ x_n & 1 \end{pmatrix}
\begin{pmatrix} a \\ b \end{pmatrix}
=
\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}
$$

最小二乗法の異なる定式化
一般逆行列 (擬逆行列)

$$\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \\ x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$\begin{pmatrix} x_1 & x_2 & \dots & x_n \\ 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \\ x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

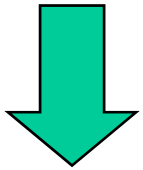
最小二乗法の異なる定式化
一般逆行列 (擬逆行列)

$$\begin{pmatrix} x_1 & x_2 & \dots & x_n \\ 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \\ x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$\begin{pmatrix} \Sigma x_i^2 & \Sigma x_i \\ \Sigma x_i & n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \Sigma x_i y_i \\ \Sigma y_i \end{pmatrix}$$

最小二乗法の異なる定式化
一般逆行列 (擬逆行列)

$$\begin{pmatrix} \Sigma x_i^2 & \Sigma x_i \\ \Sigma x_i & n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \Sigma x_i y_i \\ \Sigma y_i \end{pmatrix}$$



$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \Sigma x_i^2 & \Sigma x_i \\ \Sigma x_i & n \end{pmatrix}^{-1} \begin{pmatrix} \Sigma x_i y_i \\ \Sigma y_i \end{pmatrix}$$

最小二乗法による直線当てはめ
補足：用語について

- 最小二乗法による直線当てはめは、「回帰分析」とも呼ばれる。
- 直線あてはめを行うことは、「回帰」と呼ばれる。
- 直線の傾き a は、「回帰係数」と呼ばれる。
- 当てはめた直線は、「回帰直線」とも呼ばれる。
- 誤差は、「残差」とも呼ばれる。
- データ点の分布を示したグラフは、「散布図」と呼ばれる。

Excelによる最小二乗法演習

(野球に興味のない人、ごめんなさい)

- 目的：現実のデータに、直線当てはめを行い、その結果を考察して、最小二乗法による直線当てはめ（回帰分析）の意義を確認する。
- 演習：打点の多い打者の傾向
  - 打率
  - 長打率
  - 打点
- 演習：イチロー選手の打撃傾向
  - ゴロ／フライ率
  - 打率
  - 本塁打率

演習：打点の多い打者の傾向分析

（ＥＳＰＮ　ホームページより）　2004年シーズン

打者	打率 （ヒット数／打数）	長打率 （総累打数／打数）	打点
シェフィールド	．２９０	．５３４	１２１
松井秀	．２９８	．５２２	１０８
ロドリゲス	．２８６	．５１２	１０６
ポサーダ	．２７２	．４８１	８１
ジーター	．２９２	．４７１	７８
ウィリアムス	．２６２	．４３５	７０

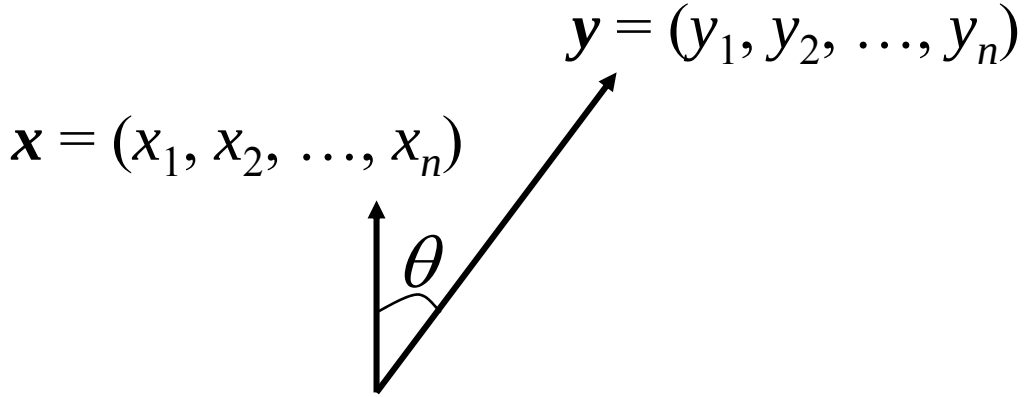
- 以下のことがいえるか？
  - 打率が高いほど打点が多い。
  - 長打率が高いほど打点が多い。

相関係数 R　決定係数 R^2値

- **相関係数 R とは？**

$$R = \cos\theta = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}| \cdot |\boldsymbol{y}|}$$

$$-1 \leqq R \leqq 1$$



ベクトルの内積

$$\boldsymbol{x} \cdot \boldsymbol{y} = \Sigma x_i y_i = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$$
$$= |\boldsymbol{x}| \cdot |\boldsymbol{y}| \cos\theta$$

ベクトル大きさ

$$|\boldsymbol{x}|^2 = \Sigma x_i^2 = x_1^2 + x_2^2 + \ldots + x_n^2$$

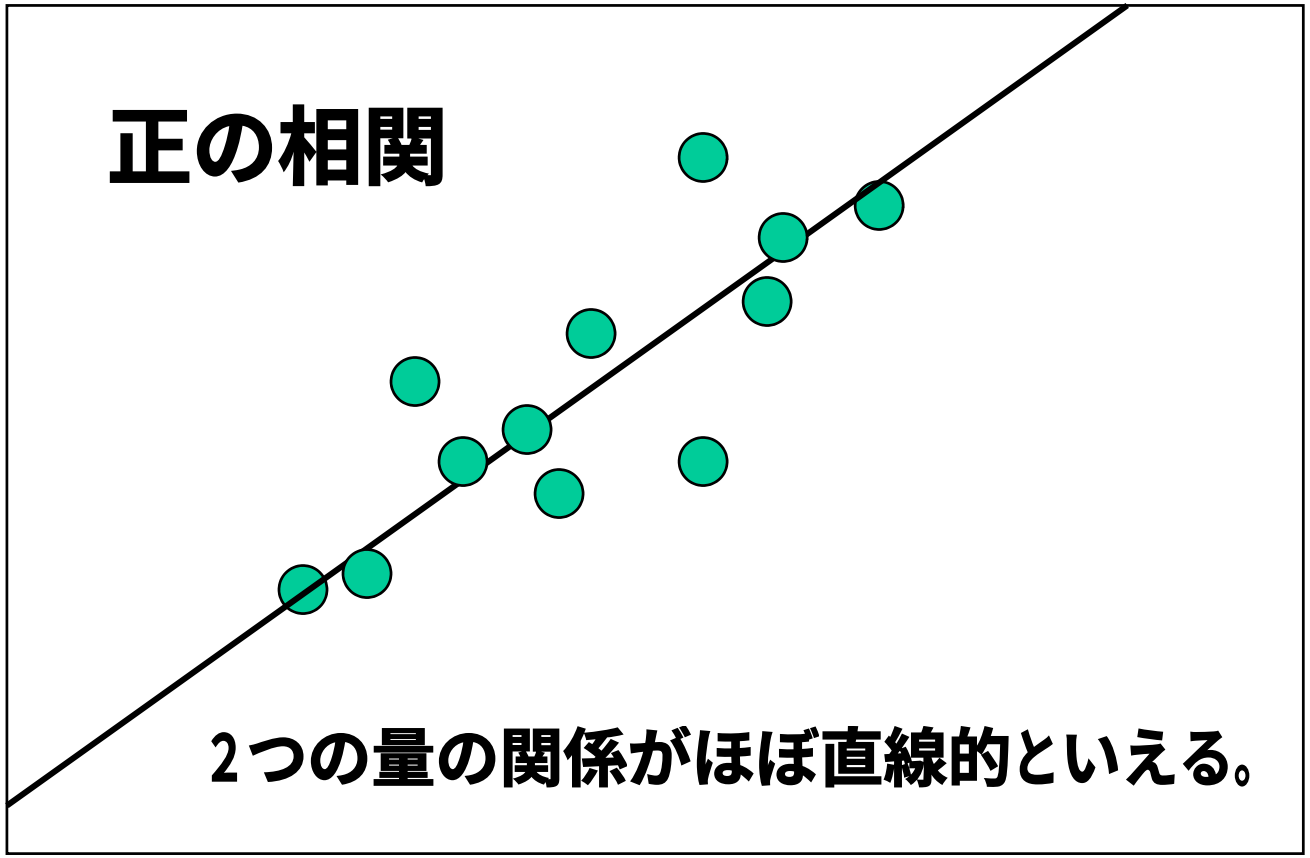
$R = -1$　**負の (完全な) 相関**
$R = 0$　**無相関 (関連性なし)**
$R = 1$　**正の (完全な) 相関**

- **決定係数 (直線当てはめの信頼性) は、相関係数の二乗 R^2**

$$0 \leqq R^2 \leqq 1$$

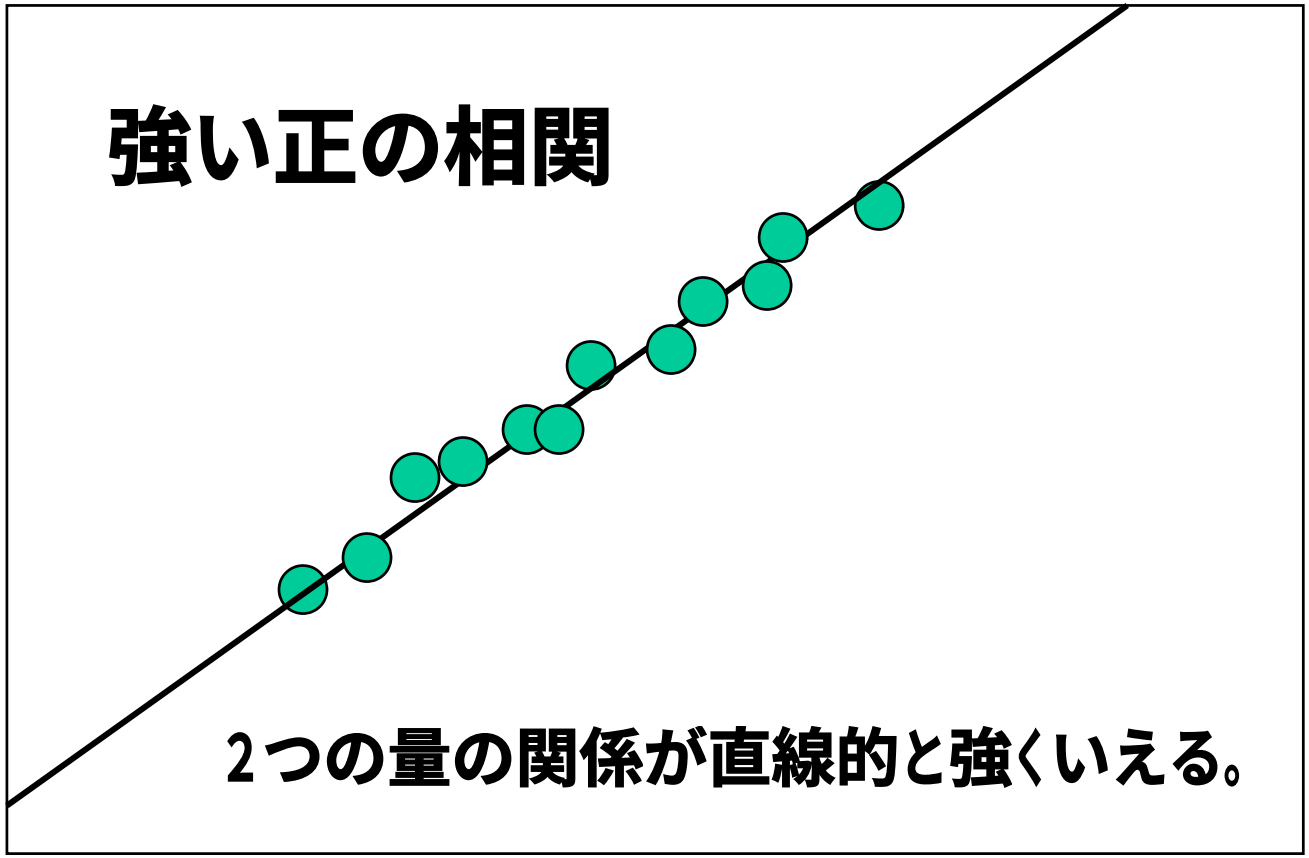
$R = 0$　**全く信頼できない。**　$R = 1$　**完全に信頼できる。**

相関係数 R　決定係数 R^2 値



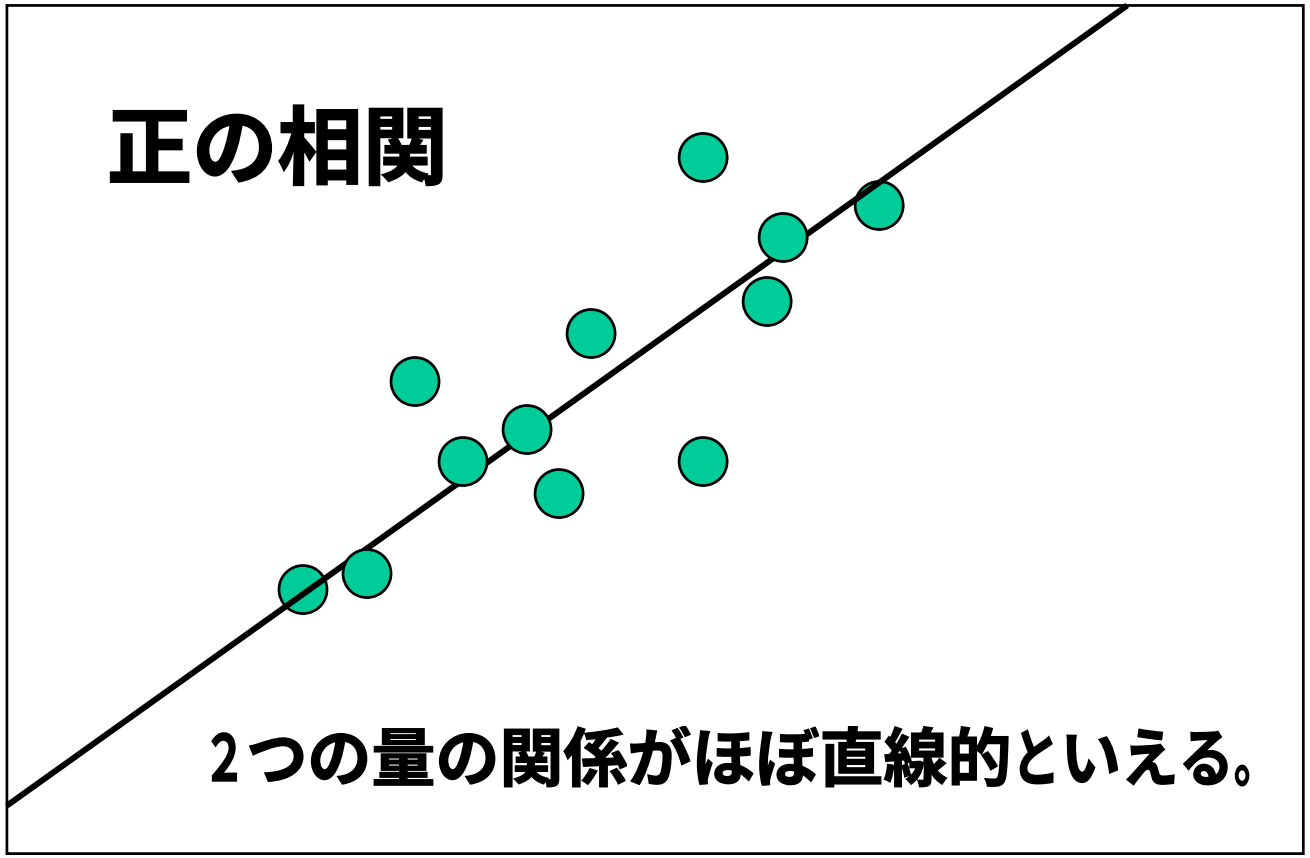
$R = 0.8, R^2 = 0.64$ くらい

相関係数 R　決定係数 R^2値



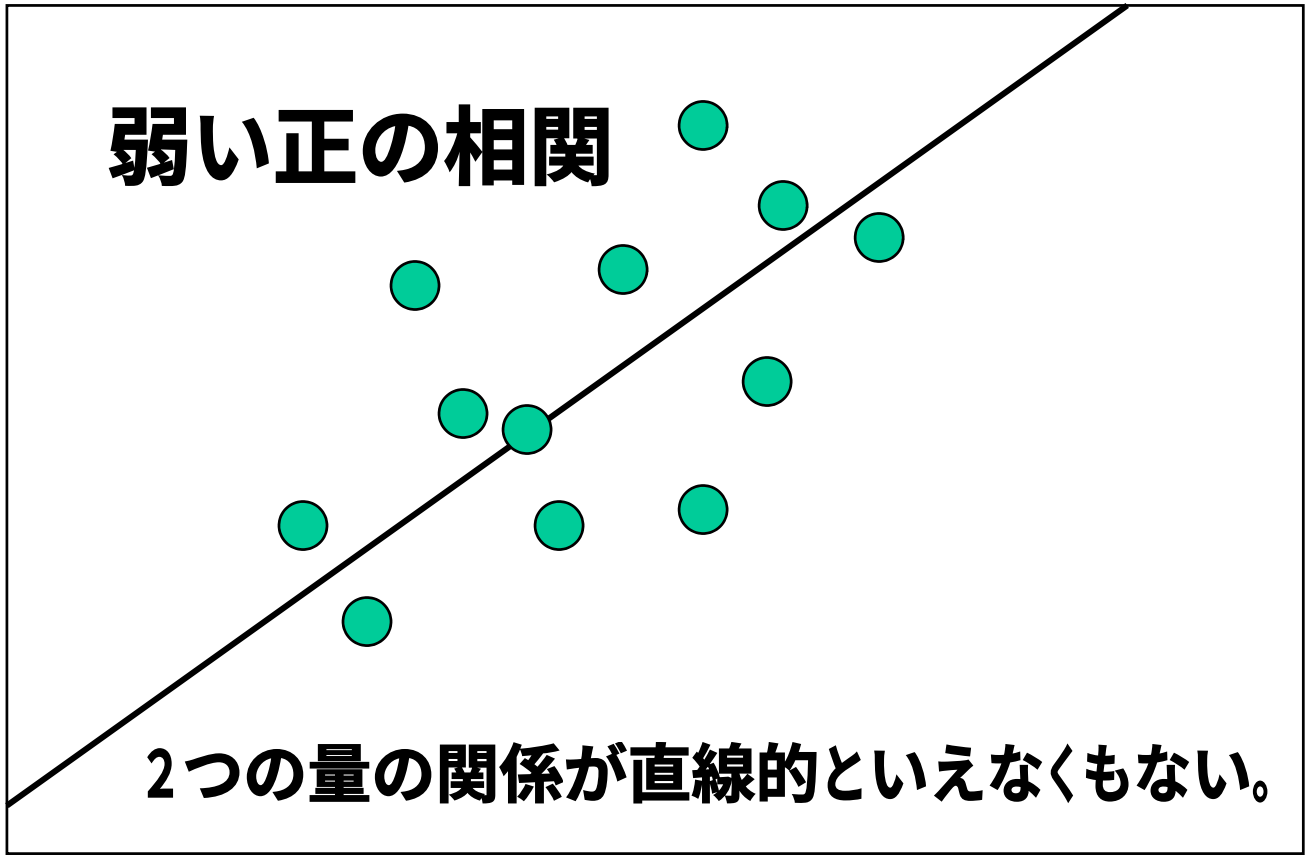
$R = 0.95, R^2 = 0.9$ くらい

相関係数 R　決定係数 R^2 値



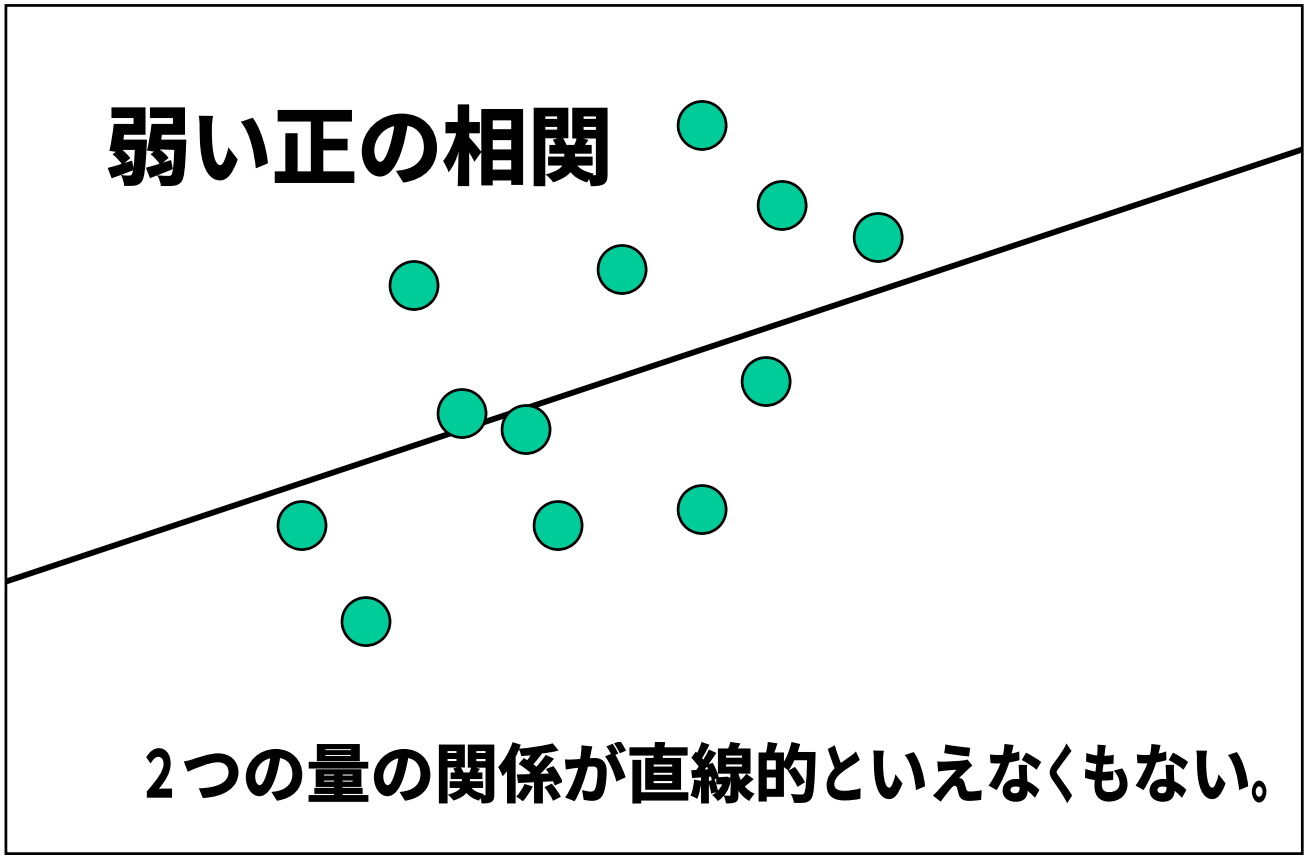
$R = 0.8, R^2 = 0.64$ くらい

相関係数 R　決定係数 R^2値



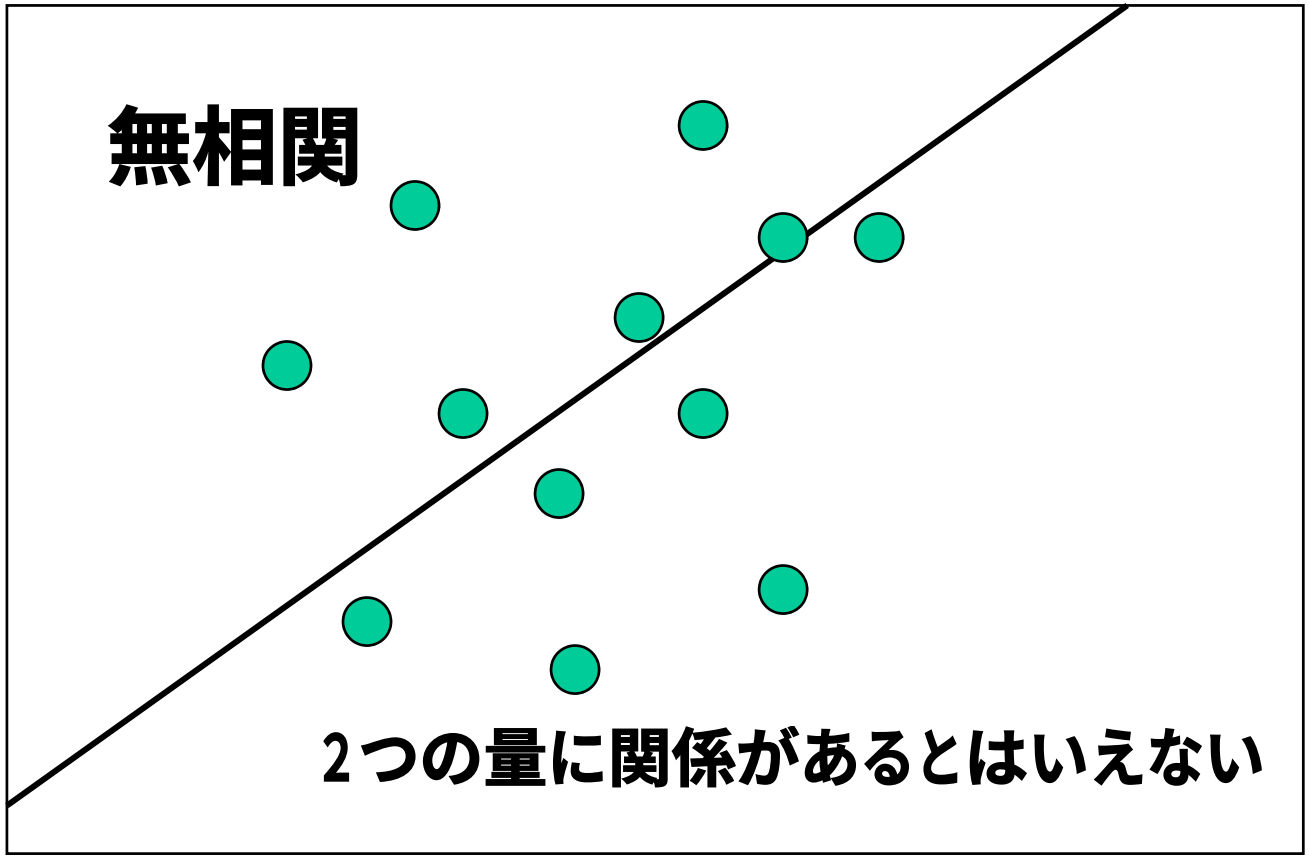
$R = 0.5, R^2 = 0.25$ くらい

相関係数 R　決定係数 R^2 値



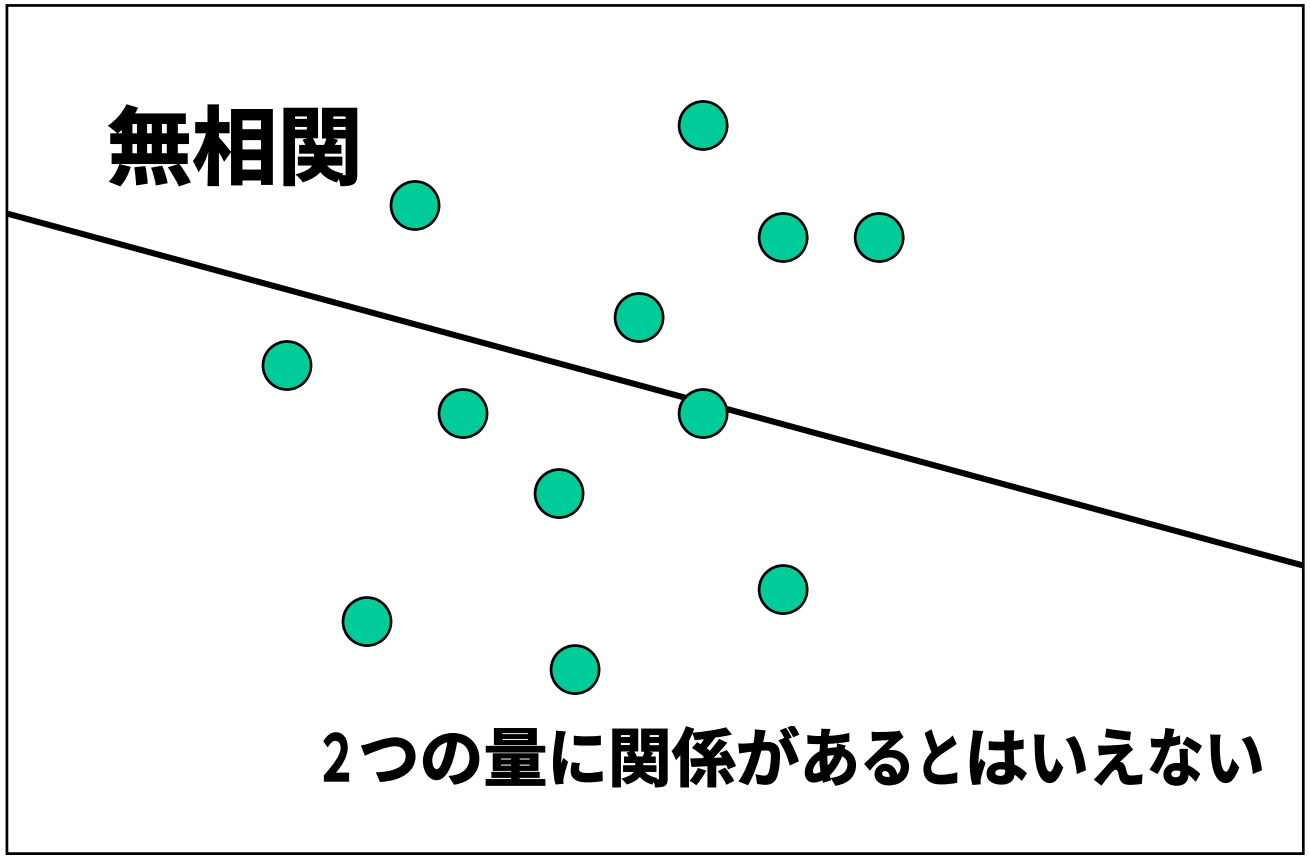
$R = 0.5, R^2 = 0.25$ くらい

相関係数 R　決定係数 R^2値



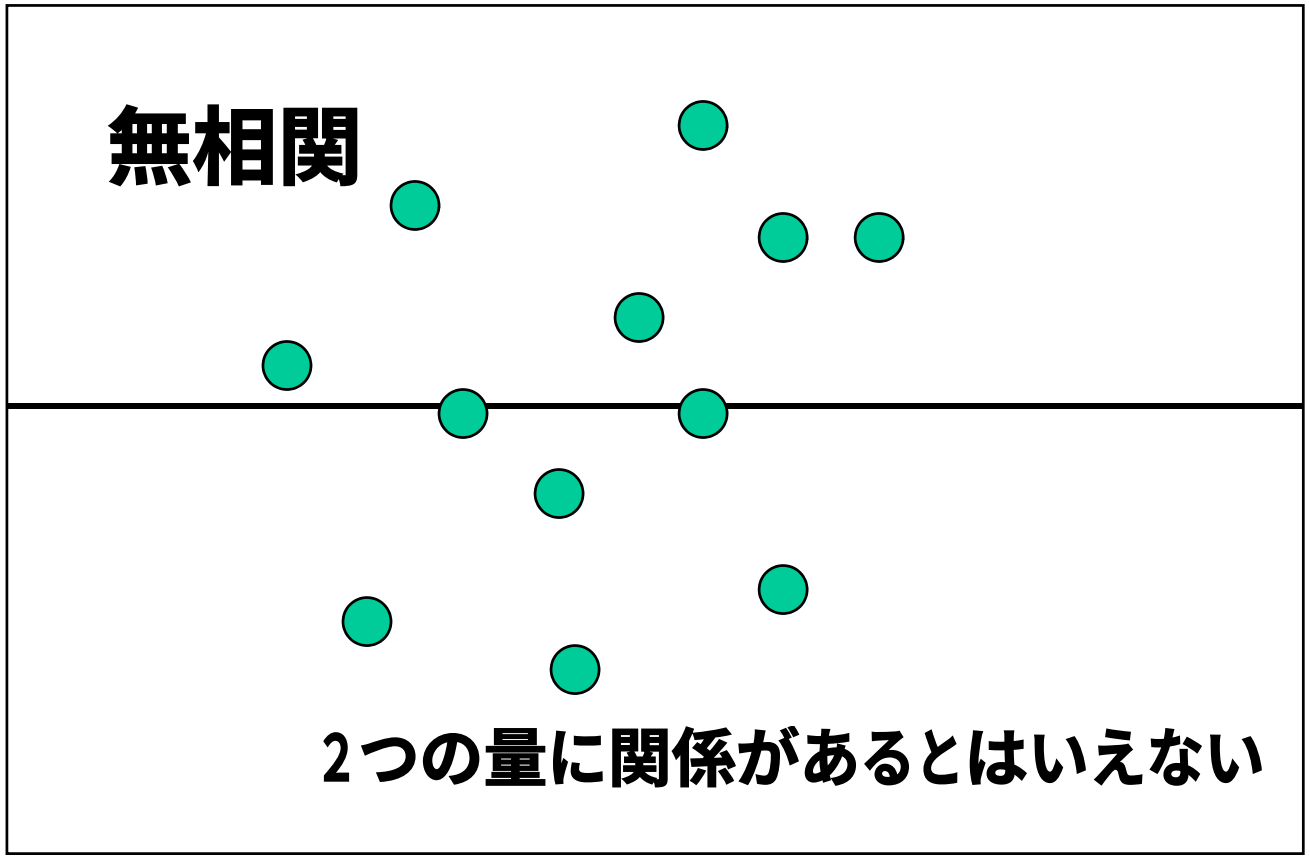
$$R = 0, R^2 = 0$$

相関係数 R　決定係数 R^2 値



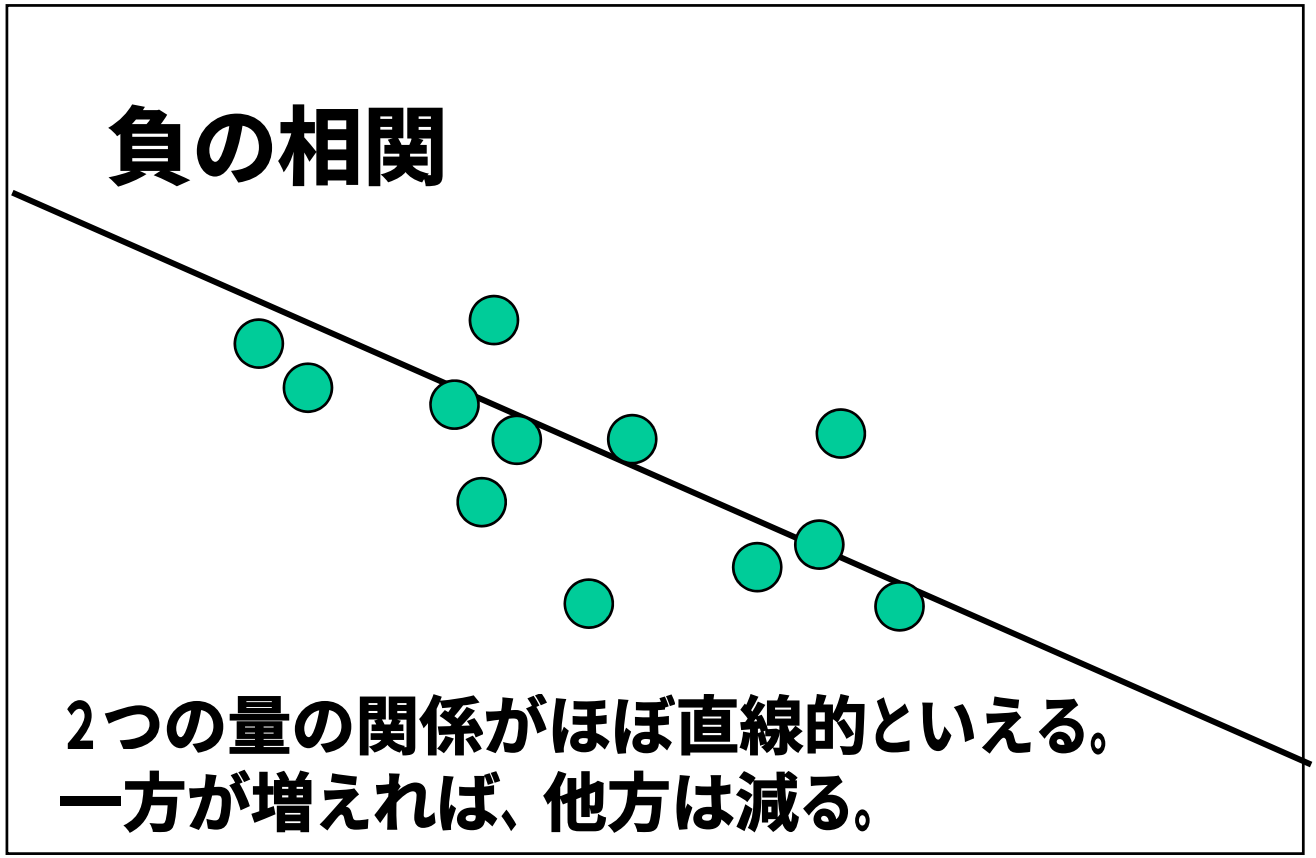
$$R = 0, R^2 = 0$$

相関係数 R　決定係数 R^2 値



$$R = 0, R^2 = 0$$

相関係数 R　決定係数 R^2 値



$R = -0.8, R^2 = 0.64$ くらい

正の相関、負の相関、無相関の例

- 正の相関があると思われる2つの量は？
  - 車の排気量と値段
  - 靴のサイズと身長
- 負の相関があると思われる2つの量は？
  - 車の燃費と値段
  - 気温とスキー場の積雪量
  - (先発) ピッチャーの防御率と勝利数
  - プロゴルファーの平均ストローク数と獲得賞金
- 無相関と思われる2つの量は？
  - 靴のサイズと学業成績
  - ＣＤシングルの曲の長さ (演奏時間) と販売枚数

演習：打点の多い打者の傾向分析

(ＥＳＰＮ　ホームページより)　　2004年シーズン

打者	打率 (ヒット数／打数)	長打率 (総累打数／打数)	打点
シェフィールド	．２９０	．５３４	１２１
松井秀	．２９８	．５２２	１０８
ロドリゲス	．２８６	．５１２	１０６
ポサーダ	．２７２	．４８１	８１
ジーター	．２９２	．４７１	７８
ウィリアムス	．２６２	．４３５	７０

- 以下のことがいえるか？
  - 打率が高いほど打点が多い。(打率と打点の関係の直線の式とR^2値をかけ)
  - 長打率が高いほど打点が多い。(長打率と打点の関係の直線の式とR^2値をかけ)

決定係数R^2値により、上の傾向がどの程度強くいえるかを論ぜよ！

演習：打点の多い打者の傾向分析

（ＥＳＰＮ　ホームページより）　2004年シーズン

打者	打率 （ヒット数／打数）	長打率 （総累打数／打数）	打点
シェフィールド	．２９０	．５３４	１２１
松井秀	．２９８	．５２２	１０８
ロドリゲス	．２８６	．５１２	１０６
ポサーダ	．２７２	．４８１	８１
ジーター	．２９２	．４７１	７８
ウィリアムス	．２６２	．４３５	７０

- 以下のことがいえるか？
  - 打率が高いほど打点が多い。　→　決定係数 R^2 の値に基づき、そういえることはいえるが、強くはいえない。
  - 長打率が高いほど打点が多い。→　決定係数 R^2 の値に基づき、そう強くいえる。

演習：イチロー選手の打撃傾向の分析

(ESPN ホームページより)

年	ゴロ／フライ率 (ゴロ数／フライ数)	打率 (ヒット数／打数)	本塁打率 (本塁打数／打数)
2001	2.63	0.350	0.0116 (8／692)
2002	2.48	0.321	0.0124 (8／647)
2003	1.77	0.312	0.0191 (13／679)
2004	3.29	0.372	0.0114 (8／704)
2005	2.06	0.303	0.0221 (15／679)
2006	1.84	0.322	0.0129 (9／695)

- 問題A. ゴロ／フライ率を横軸、打率を縦軸にとり、直線当てはめを行え。(直線の式をかけ)
- 問題B. この直線当てはめの決定係数R^2はどの程度か？ (信頼できそうか？)
- 問題C. ゴロ／フライ率を横軸、本塁打率を縦軸にとり、直線当てはめを行え。(直線の式をかけ)
- 問題D. ゴロ／フライ率がどのくらいになれば、打率4割が可能か？ (計算式もかけ)
- 問題E. そのときの本塁打数はどのくらいと予想されるか？ (700打数とする) (計算式もかけ)
- 問題F. ここで直線当てはめ (回帰分析) により得られた結果は野球のシミュレーションゲーム設計にどのように応用可能か？

演習問題の解答

- ゴロフライ率と打率の関係
  - Y = 45.4 X +221
  - Y = 400 = 45.4 X + 221
  - 45.4 X = 400 – 221 = 179
- 打率4割のときのゴロフライ率
  - X = 179/45.4 = 3.943 = 3.9
- 打率4割のときのホームラン数（700打数とする。）
  - 0.004 x 700 = 3 本

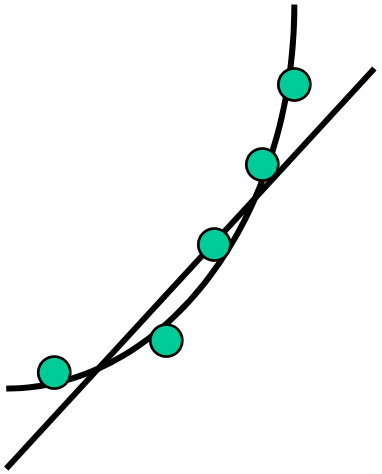
- 打率　4割打つためには、ゴロフライ率が 3.9
- そのときの本塁打数（700打数）は、3.
- 2次多項式予測の場合は、
  - 0.012 x 700 = 8 本

直線当てはめから多項式当てはめへ

- 2つの量 x, y の関係がいつでも直線的になるとは限らない。
- 2つの量 x, y の関係を表現する数式として、様々な関係が考えられる。

$$y = ax + b$$

- 誤差 $d_i = y_i - (a\,x_i + b)$

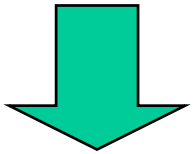


$$y = ax^2 + bx + c$$

- 誤差 $d_i = y_i - (a\,x_i^2 + b\,x_i + c)$

最小二乗基準

- すべての点が、2次多項式 $y = ax^2 + bx + c$ の上になるべく のっている。



$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \{y_i - (a\,x_i^2 + b\,x_i + c)\}^2$ を最小にする a, b, c を求める。

n 個の点があり、i 番目の点の座標値を (x_i, y_i) とする。

最小二乗基準

$f(a,b,c)=\sum_{i=1}^{n}\{y_i-(a\,x_i^2+b\,x_i+c)\}^2$ を最小にする $a,\ b,\ c$ を求める。

$f(a,b,c)$ を最小にする $a,\ b,c$ を求める。

$$\frac{\partial}{\partial a}f(a,b,c)=0 \quad \frac{\partial}{\partial b}f(a,b,c)=0 \quad \frac{\partial}{\partial c}f(a,b,c)=0$$

$a,\ b,\ c$ に関する偏微分が 0 になる。

Excelによる最小二乗法演習
2次多項式当てはめ

- 目的：現実のデータに、2次多項式当てはめを行い、その結果を考察して、最小二乗法による2次多項式当てはめの意義を確認する。
- 例題：プロゴルファーの平均ストローク数と獲得賞金額
  - 直線当てはめ
  - 2次多項式当てはめ
- 例題：イチロー選手の打撃傾向：直線 vs. 2次多項式
  - ゴロ／フライ率
  - 打率
  - 本塁打率

演習：プロゴルファーの平均ストローク数と獲得賞金額
（2004年　日本女子プロツアー　SANSPO　ホームページより）

選手名	平均ストローク	獲得賞金／ラウンド（万円）	優勝回数
不動	70.64	14277／72 ＝ 198.3	7
宮里	70.85	12297／82 ＝ 150.0	5
福嶋	71.55	4715／55 ＝ 85.7	1
肥後	72.00	5197／76 ＝ 68.4	1
米山	72.14	3855／81 ＝ 47.6	0
大山	72.18	3959／89 ＝ 44.5	0
木村	72.20	5264／84 ＝ 62.7	1
高	72.32	4140／88 ＝ 47.0	0
表	72.36	4824／90 ＝ 53.6	0
服部	72.37	4258／79 ＝ 53.9	1

- 平均ストロークを横軸、1ラウンドあたりの獲得賞金を縦軸にとり
  - 直線当てはめ
  - 2次多項式当てはめ

の両方を行い、比較せよ。R^2値も調べること。

演習：イチロー選手の打撃傾向の分析
直線当てはめ vs. 2次多項式当てはめ

(ESPN ホームページより)

年	ゴロ／フライ率 (ゴロ数／フライ数)	打率 (ヒット数／打数)	本塁打率 (本塁打数／打数)
2001	2.63	.350	.0116 (8／692)
2002	2.48	.321	.0124 (8／647)
2003	1.77	.312	.0191 (13／679)
2004	3.29	.372	.0114 (8／704)
2005	2.06	.303	.0221 (15／679)
2006	1.84	.322	.0129 (9／695)

- 問題A. ゴロ／フライ率を横軸、打率を縦軸にとり、2次多項式当てはめを行え。R^2値も調べて、当てはめの信頼性をチェックせよ。
- 問題B. ゴロ／フライ率を横軸、本塁打率を縦軸にとり、2次多項式当てはめを行え。R^2値も調べて、当てはめの信頼性をチェックせよ。
- 問題C. 2次多項式に基づく予測によると，ゴロ／フライ率がどのくらいになれば、打率4割が可能か？
- 問題D. そのときの本塁打数はどのくらいと予想されるか？ (700打数とする)
- 問題E. この予測は、直線当てはめによる予測に比べてどんな問題点があるか？

最小二乗法：演習課題

1. なんらかの2つの量を考え、それらの2つの量の関係を最小二乗法を用いて解析せよ。
2. 数値データの収集に関しては、インターネットを活用せよ。用いたデータのリストを示すこと。（最低でも、１０組以上のデータを集めること。身長と体重の関係を調べる場合を例にとると、最程でも１０人分の身長と体重のデータを集めるということである。）
3. エクセルで、線形近似、2次多項式近似などを試みよ。得られた式をかけ。また、決定係数（R^2値）も確認せよ。
4. 結果を以下の観点から考察せよ。
   - 解析結果によりどのようなことが言えるか？
   - 解析結果がどのように役立つか？
   - 解析結果を用いて、どんな予測が行えるか？

最小二乗法：演習課題の発表について

- 3人（どうしても無理な場合は、2人または4人）組をつくり、3人で相談して、演習課題を行い、代表者が、パワーポイントで、前に出て発表を行う（マイク使用可）。
- 発表は、一組につき１０分程度とする。また、代表者として発表した人以外は、必ず、発表に対して質問を行う。代表者として発表した人も質問はしてもよい。質問をする前に、学生番号と名前を述べること。
- 発表後に、発表に用いたパワーポイントファイルを提出すること。それに加えて、メンバー3人のそれぞれがどんな貢献をしたかを記述したページを、末尾に追加すること。
- パワーポイントの最初のページには、選んだ課題名と各組のメンバーの学生番号と名前を明記すること。
- 発表会は、6月中旬に行う。

グループ演習・発表

- 最小二乗法グループ演習の発表会を行います。3人程度で班をつくり、1つの班につき、発表時間10分、質疑応答時間5分で発表をしてもらいます。なお、パワーポイントの作成においては、できるだけ、以下の順にスライドをつくり説明していくこと（括弧内は標準的なスライド枚数）。ムービーや図やグラフを多用すること。ウェブからダウンロードした図については、アドレス等を示すこと。
  - 発表タイトル、班メンバー全員の学生番号と名前（発表者にしるしをつける）（1枚）
  - 背景：取り扱う題材の基本的な説明とそれをとりあげた動機（1枚程度）
  - 目的：どんなデータを使って何を解析するか？（1枚程度）
  - データ収集方法（2～3枚程度）
  - データ解析方法（2～3枚程度）
  - 結果：Excelの結果などをはりつけ、推定した式に基づいた予測結果などを示す。（3～4枚程度）
  - 考察：結果において、おもしろい点、役に立つ点、予想外（予想通り）の点などをまとめる。（2枚程度）
  - まとめ：発表の最初（背景）から最後（考察）までのまとめ（1枚）

最小2乗法グループプレゼン

- パワーポイントは、以下の構成に（できるだけ）従うこと。
  - 目的
    - どのような目的で、どのようなデータを解析するか？
  - データ収集方法
  - データ解析方法
    - データの値をそのまま使うのではなく、正規化（例えば、本塁打数ではなく、本塁打率にする）をする必要がある場合がある。（正規化とは、各データ値の条件をそろえるため、たとえば、0から1の範囲をとる値にすることをいう）
  - 結果（エクセルの出力結果などを貼り付ける）
  - 考察：
    - 結果からどのようなことが言えるか？
    - 結果において、おもしろい点はどこか？
    - 何の役に立ちそうか（どのような予測が行えるか）？