先週までの「データ解析」

Prologue: データを読み解くとは何なのか
(多変量解析とデータサイエンスと統計学とパターン認識と機械学習とデータマイニング)

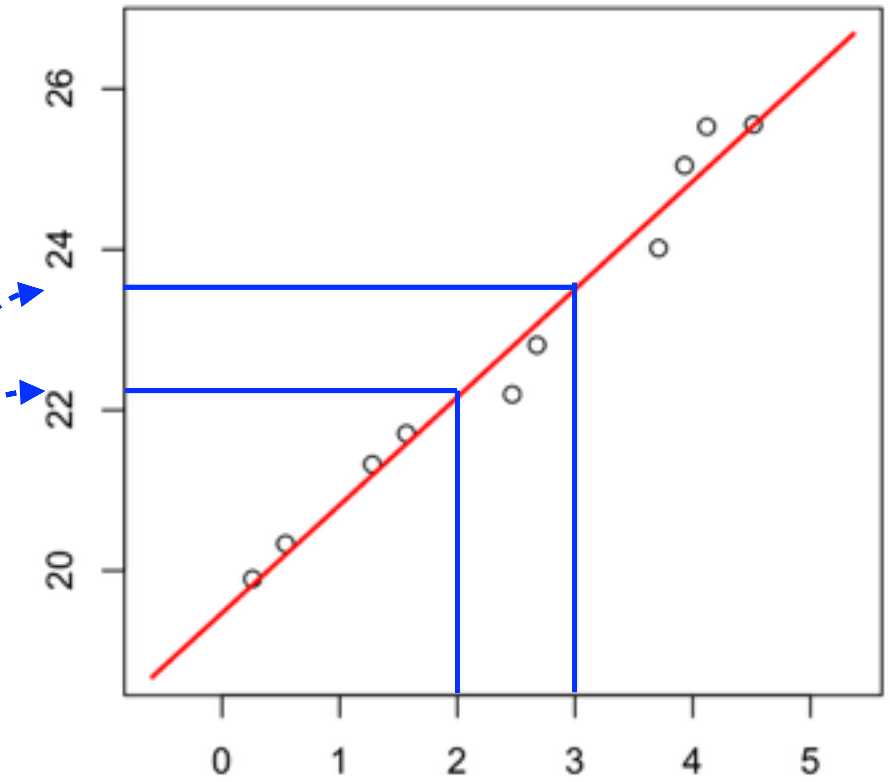
DAY-1 6/16 (01)(02) **単回帰: 点群への直線当てはめを”真剣に”考える**
(見えない世界へようこそ)

今日の内容

DAY-2 6/23 (03)(04) **重回帰と線形代数: 回帰の行列計算とその意味**
(データの計算とデータの解釈)

$$\hat{y} = 1.343664\, x + 19.474212$$

	x	y	$\hat{y}$
手元にあるデータ	0.541	20.336	20.20113442
	1.277	21.326	21.19007119
	0.26	19.898	19.82356481
	3.93	25.043	24.75481204
	2.676	22.81	23.06985726
	2.466	22.195	22.7876878
	1.566	21.708	21.57839012
	4.119	25.528	25.00876455
	4.515	25.554	25.54085554
	3.709	24.015	24.45786227
手元にないデータ	2.0	NA	22.16154
	3.0	NA	23.5052



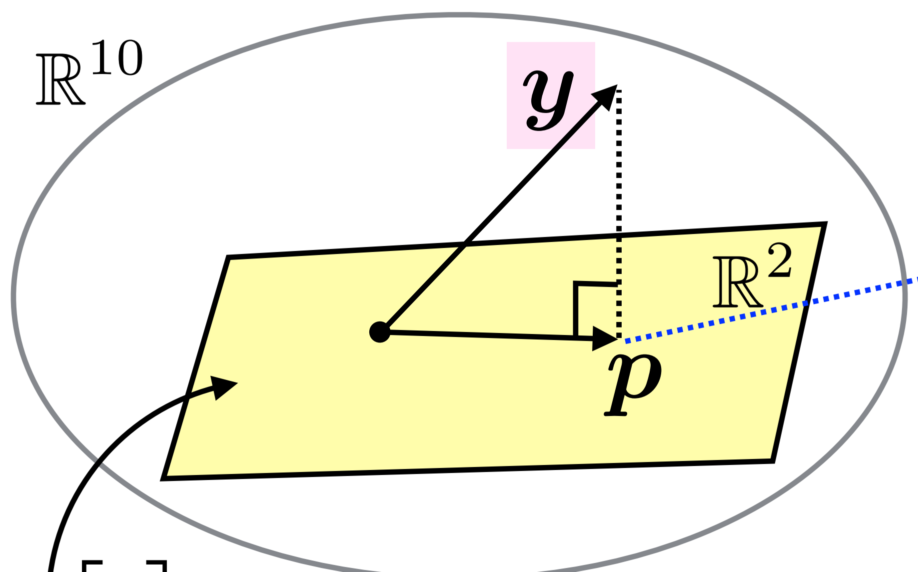
データ点にモデル式を当てはめることでデータにない点について何か語る

$$\boldsymbol{X} = \begin{bmatrix} 0.541 & 1 \\ 1.277 & 1 \\ 0.26 & 1 \\ 3.93 & 1 \\ 2.676 & 1 \\ 2.466 & 1 \\ 1.566 & 1 \\ 4.119 & 1 \\ 4.515 & 1 \\ 3.709 & 1 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 20.336 \\ 21.326 \\ 19.898 \\ 25.043 \\ 22.81 \\ 22.195 \\ 21.708 \\ 25.528 \\ 25.554 \\ 24.015 \end{bmatrix}$$

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \begin{bmatrix} 1.343664 \\ 19.474212 \end{bmatrix}$$

先週の理解(復習): 直交射影

$\begin{bmatrix} a \\ b \end{bmatrix}$ を色々変えて、予測値 $\hat{\boldsymbol{y}} = a \begin{bmatrix} 0.541 \\ 1.277 \\ 0.26 \\ 3.93 \\ 2.676 \\ 2.466 \\ 1.566 \\ 4.119 \\ 4.515 \\ 3.709 \end{bmatrix} + b \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$ を $\boldsymbol{y} = \begin{bmatrix} 20.336 \\ 21.326 \\ 19.898 \\ 25.043 \\ 22.81 \\ 22.195 \\ 21.708 \\ 25.528 \\ 25.554 \\ 24.015 \end{bmatrix}$ に近づける



$\hat{\boldsymbol{y}} = \boldsymbol{X} \begin{bmatrix} a \\ b \end{bmatrix}$ の可動域

$\boldsymbol{p} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$

のとき $\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2$ が最小に

つまり $\begin{bmatrix} a \\ b \end{bmatrix} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$

このテクニックを利用して例えば以下も解ける！

(1.5,2) および (3,3.5)の2点を通る直線の式を求めよ

① 表を作る　② 行列を作る　③計算！

x	y
1.5	2
3	3.5

$$\boldsymbol{X} = \begin{bmatrix} 1.5 & 1 \\ 3 & 1 \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} 2 \\ 3.5 \end{bmatrix}$$

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$

つまり傾き1,切片0.5

$$y = 1 \cdot x + 0.5$$

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{bmatrix} 45/4 & 9/2 \\ 9/2 & 2 \end{bmatrix}$$

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \begin{bmatrix} 8/9 & -2 \\ -2 & 5 \end{bmatrix}$$

$$\boldsymbol{X}^\top \boldsymbol{y} = \begin{bmatrix} 27/2 \\ 11/2 \end{bmatrix}$$

…でも、これって結局
何の計算してるの??

…転置行列をかける
意味は何?

単回帰を考えて得られたこと①

- 単変量の場合でも**行列やベクトル**を使えば問題が美しく解けるのを見ることができる！（線形代数との接点）
- 単にもっとも当てはまりの良い回帰の係数(傾きと切片)を計算すれば良いだけなら「線形代数」があればよくて「確率統計」の道具は要らない。

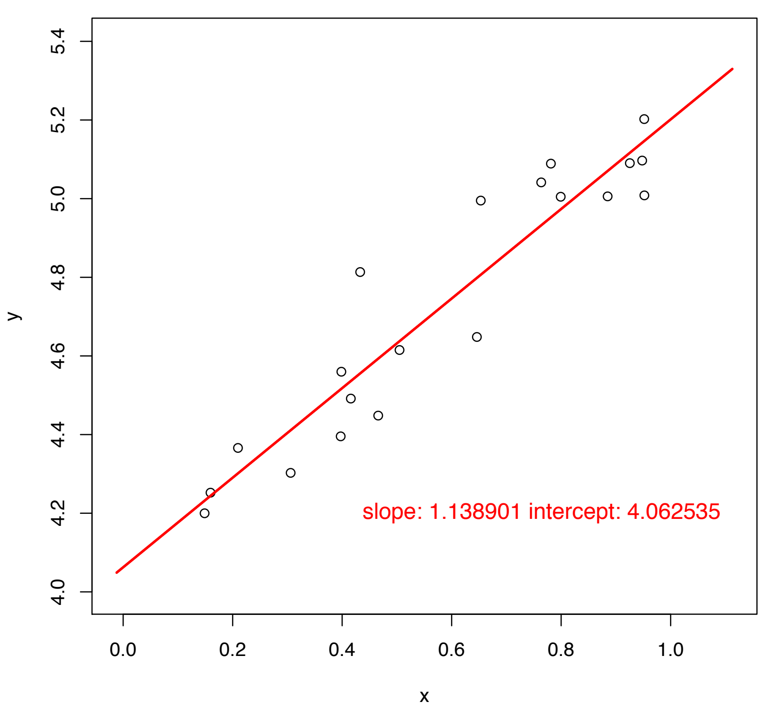
$\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ や $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ の意味は？

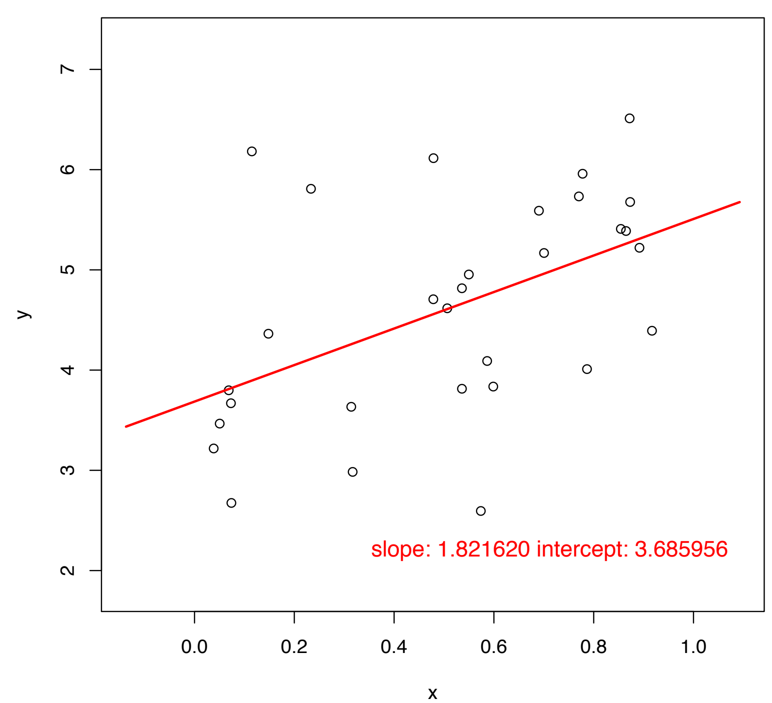
→ 回帰計算の意味を今日の後半でもう少し掘り下げます。
ベクトル・行列表記が活きる多変量解析計算の真骨頂！

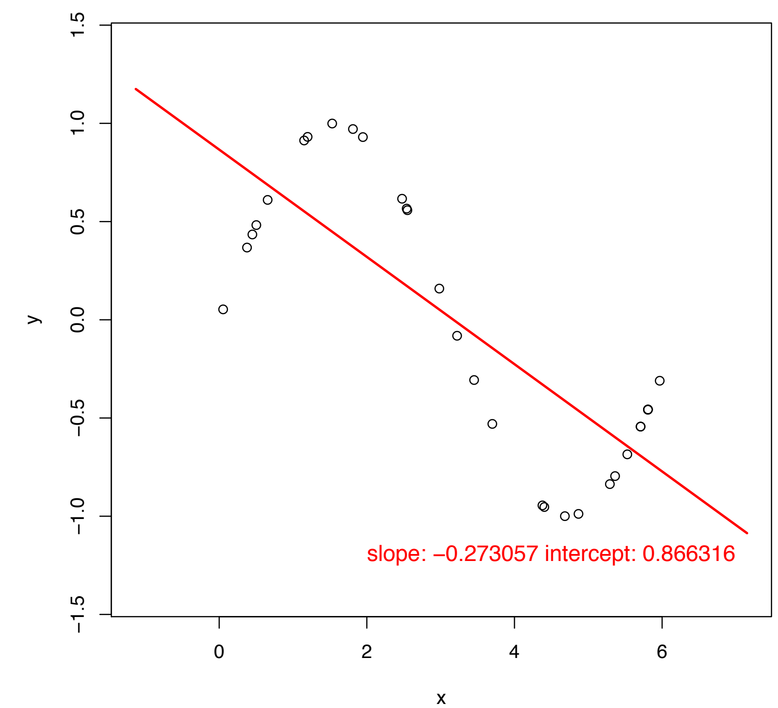
単回帰を考えて得られたこと②

$(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ が存在するならどんなに変なデータでも

線形回帰の「計算」は常に可能であることに注意





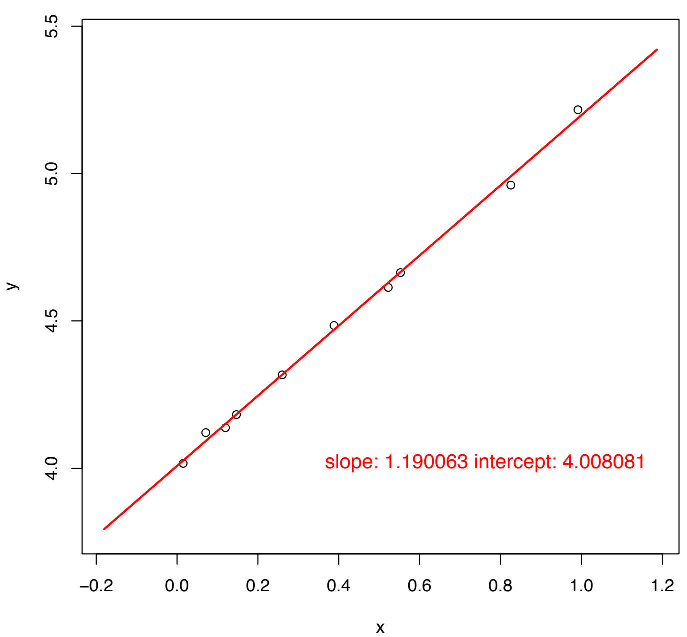


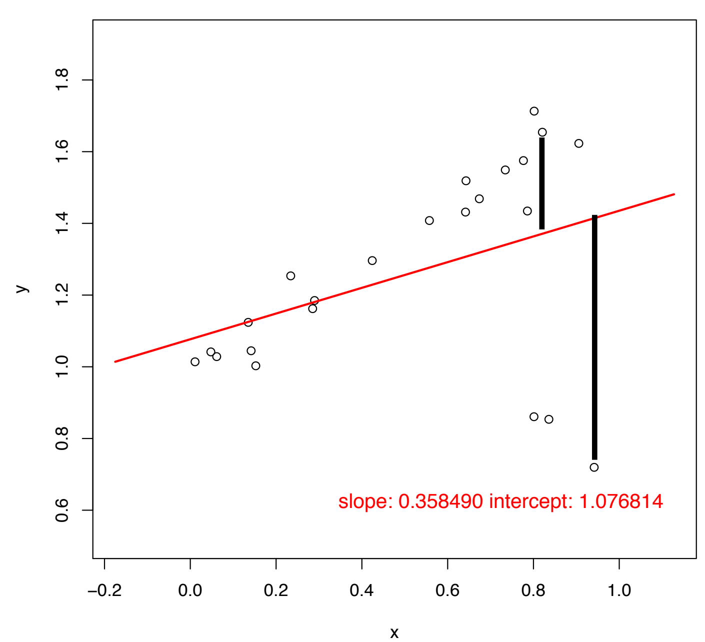
じゃあ良し悪しはどうやってわかるの？

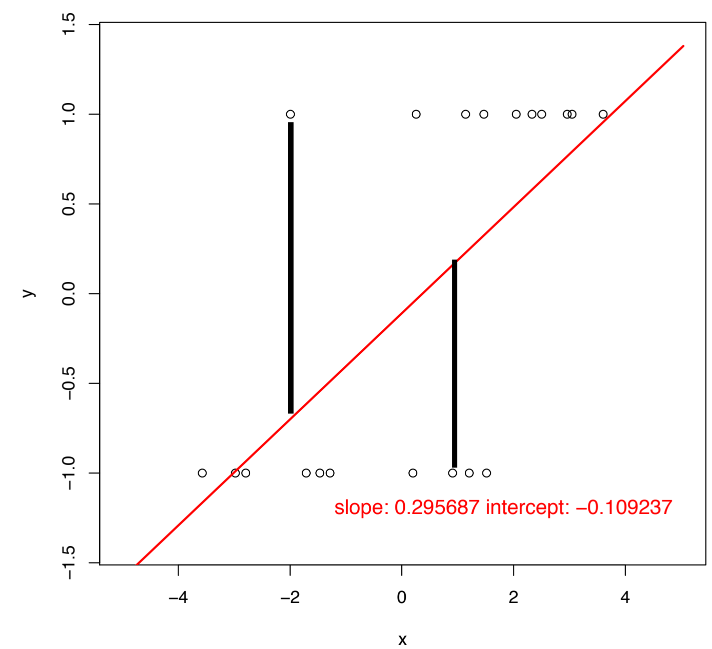
→ ここで確率統計が登場！来週のメインテーマ

単回帰を考えて得られたこと③

二乗誤差の「総和」を小さくするので、手元に既にあるデータであっても予測値がかなり悪くなり得る。







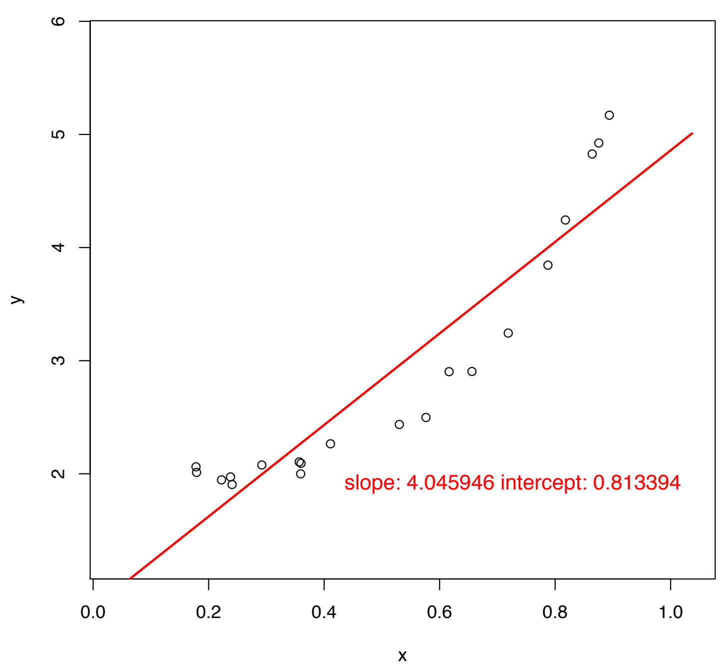
$\varepsilon_i = \hat{y}_i - y_i$ に対して、なぜ $|\varepsilon_1| + |\varepsilon_2| + \cdots + |\varepsilon_n|$

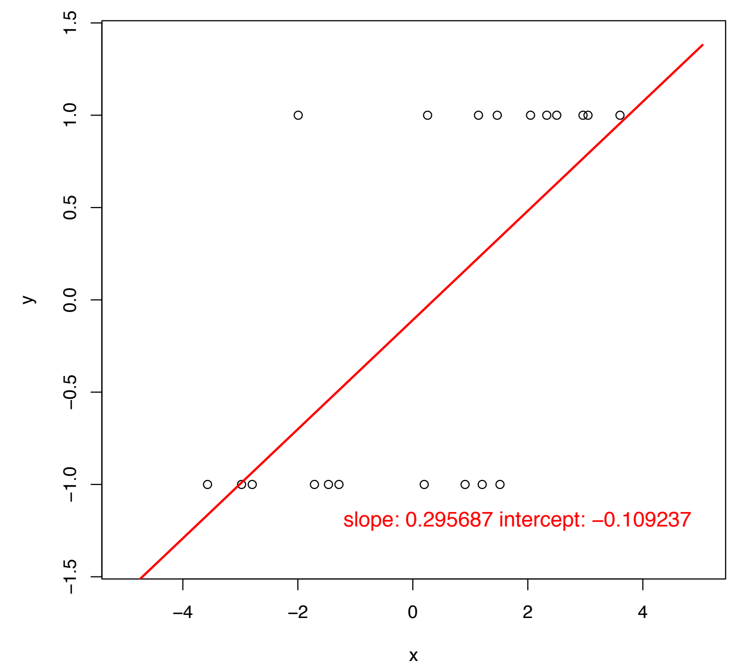
ではなく、二乗した $\varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_n^2$ を最小に？

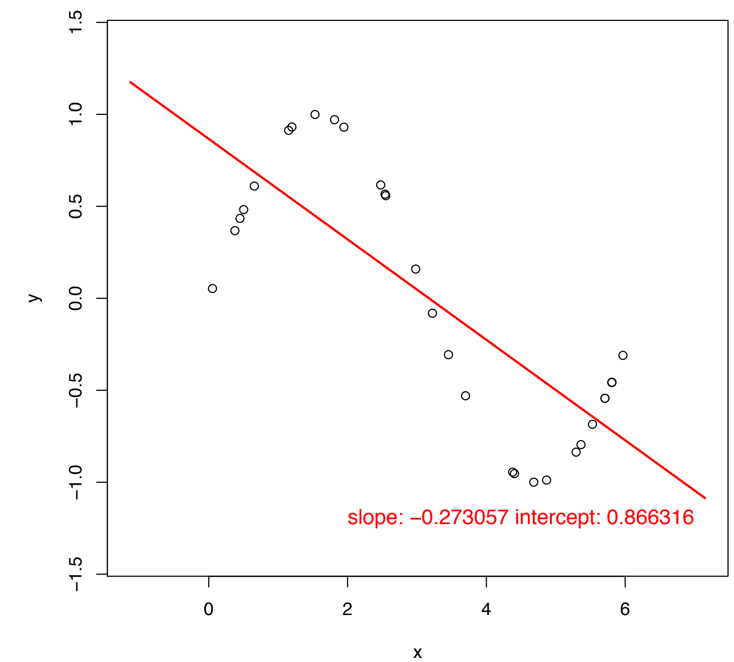
$\rightarrow$ これにも確率統計が必要。来週のテーマの一つ

単回帰を考えて得られたこと④

- 手法の良し悪しは**モデル式がデータに適合しているかどうか**で決まる！データに依らず、ある方法が良いとか悪いとか言うことはできない。







直線ではなくて曲線を当てはめるにはどうする？
→ 今日取り上げます。一般には機械学習の領域へ。

多変量の線形回帰：単回帰から重回帰へ

単回帰

説明変数 目的変数

	x	y
p_1	x_1	y_1
p_2	x_2	y_2
p_3	x_3	y_3
p_4	x_4	y_4
p_5	x_5	y_5
p_6	x_6	y_6

説明変数が複数に

重回帰

説明変数 **目的変数**

	x_1	x_2	$\cdots$	x_d	y
p_1	x_{11}	x_{12}		x_{1d}	y_1
p_2	x_{21}	x_{22}		x_{2d}	y_2
p_3	x_{31}	x_{32}	$\cdots$	x_{3d}	y_3
p_4	x_{41}	x_{42}		x_{4d}	y_4
p_5	x_{51}	x_{52}		x_{5d}	y_5
p_6	x_{61}	x_{62}		x_{6d}	y_6

1.2　重回帰分析とは

表 1.3 は東京のある駅の徒歩圏内の中古マンションに関するデータである.

表 1.3　中古マンションのデータ

サンプル No.	広さ x_1 (m^2)	築年数 x_2 (年数)	価格 y (千万円)
1	51	16	3.0
2	38	4	3.2
3	57	16	3.3
4	51	11	3.9
5	53	4	4.4
6	77	22	4.5
7	63	5	4.5
8	69	5	5.4
9	72	2	5.4
10	73	1	6.0

このデータに基づいて知りたいことは次の通りである.

(1) 価格は広さと築年数とによって予測できるだろうか.

(2) 予測できるとすればその精度はどのくらいか.

(3) 同じ地区で $x_1 = 70$, $x_2 = 10$, $y = 5.8$ を提示された. 価格は妥当か.

x	z	y
51	16	3
38	4	3.2
57	16	3.3
51	11	3.9
53	4	4.4
77	22	4.5
63	5	4.5
69	5	5.4
72	2	5.4
73	1	6

ポイント

- 説明変数が複数に（ x と z ）
- それに伴い回帰の予測式が

$$\hat{y} = a\,x + b\,z + c$$

- 決めたい係数の個数は「説明変数+1」個になる

$$\boldsymbol{X} = \begin{bmatrix} 51 & 16 & 1 \\ 38 & 4 & 1 \\ 57 & 16 & 1 \\ 51 & 11 & 1 \\ 53 & 4 & 1 \\ 77 & 22 & 1 \\ 63 & 5 & 1 \\ 69 & 5 & 1 \\ 72 & 2 & 1 \\ 73 & 1 & 1 \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} 3 \\ 3.2 \\ 3.3 \\ 3.9 \\ 4.4 \\ 4.5 \\ 4.5 \\ 5.4 \\ 5.4 \\ 6 \end{bmatrix}$$

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

同じ形に！

重回帰の計算

説明変数　　**目的変数**

	x_1	x_2	$\cdots$	x_d	y
p_1	x_{11}	x_{12}		x_{1d}	y_1
p_2	x_{21}	x_{22}		x_{2d}	y_2
p_3	x_{31}	x_{32}	$\cdots$	x_{3d}	y_3
p_4	x_{41}	x_{42}		x_{4d}	y_4
$\vdots$					$\vdots$
p_n	x_{n1}	x_{n2}		x_{nd}	y_n
	β_1	β_2	$\cdots$	β_d	**回帰係数**

回帰式

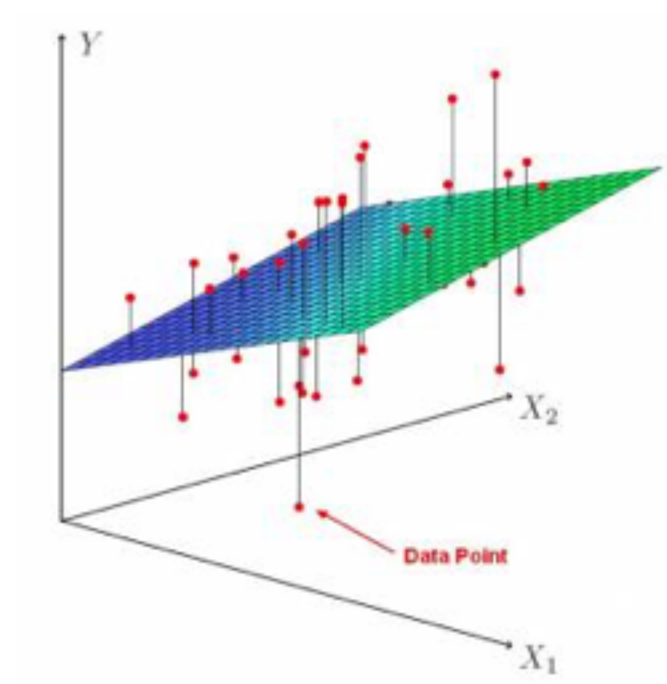
$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots \beta_d x_d$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{1d} \\ 1 & x_{21} & x_{22} & x_{2d} \\ 1 & x_{31} & x_{32} & x_{3d} \\ \vdots & & & \vdots \\ 1 & x_{n1} & x_{n2} & x_{nd} \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

二変数回帰のイメージ

回帰式は「平面」を表す



もっとも教科書的な導出：正規方程式

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ に直線を当てはめよ

$$J = \sum_{i=1}^{n} \bigl(y_i - (ax_i + b)\bigr)^2 \to \min \quad \text{(まず単回帰で)}$$

$$\iff \quad \frac{\partial J}{\partial a} = 0 \quad \text{かつ} \quad \frac{\partial J}{\partial b} = 0$$

$$\frac{\partial f(g(x))}{\partial x} = f'(g(x)) \cdot g'(x)$$

$$\frac{\partial J}{\partial a} = \sum_{i=1}^{n} 2 \cdot \bigl(y_i - (ax_i + b)\bigr) \cdot (-x_i) = 2a \sum_{i=1}^{n} x_i^2 + 2b \sum_{i=1}^{n} x_i - 2 \sum_{i=1}^{n} x_i y_i$$

$$\frac{\partial J}{\partial b} = \sum_{i=1}^{n} 2 \cdot \bigl(y_i - (ax_i + b)\bigr) \cdot (-1) = 2a \sum_{i=1}^{n} x_i + 2b \sum_{i=1}^{n} 1 - 2 \sum_{i=1}^{n} y_i$$

$$\iff \begin{cases} a\sum_{i=1}^n x_i^2 + b\sum_{i=1}^n x_i - \sum_{i=1}^n x_i y_i = 0 \\ a\sum_{i=1}^n x_i + b\sum_{i=1}^n 1 - \sum_{i=1}^n y_i = 0 \end{cases}$$

$$\iff \begin{cases} a\sum_{i=1}^n x_i^2 + b\sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i \\ a\sum_{i=1}^n x_i + b\sum_{i=1}^n 1 = \sum_{i=1}^n y_i \end{cases}$$

$$\iff \begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{bmatrix}$$

これは実は以下と同じになる

$$\boldsymbol{X}^\top \boldsymbol{X} \begin{bmatrix} a \\ b \end{bmatrix} = \boldsymbol{X}^\top \boldsymbol{y} \quad \Rightarrow \quad \begin{bmatrix} a \\ b \end{bmatrix} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

もっとも教科書的な導出：正規方程式

$(x_1, y_1, z_1), (x_2, y_2, z_2), \cdots, (x_n, y_n, z_n)$ に平面を当てはめよ

$$J = \sum_{i=1}^{n} \bigl(z_i - (ax_i + by_i + c)\bigr)^2 \to \min$$ (重回帰)

$$\iff \quad \frac{\partial J}{\partial a} = 0 \quad \text{かつ} \quad \frac{\partial J}{\partial b} = 0 \quad \text{かつ} \quad \frac{\partial J}{\partial c} = 0$$

$$\begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i y_i & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i & \sum_{i=1}^n y_i^2 & \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n y_i & \sum_{i=1}^n 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i z_i \\ \sum_{i=1}^n y_i z_i \\ \sum_{i=1}^n z_i \end{bmatrix}$$

ベクトルによる微分

$$\frac{\partial J}{\partial a} = 0 \quad \text{かつ} \quad \frac{\partial J}{\partial b} = 0 \quad \text{かつ} \quad \frac{\partial J}{\partial c} = 0$$

$$\Rightarrow \quad \begin{bmatrix} \partial J/\partial a \\ \partial J/\partial b \\ \partial J/\partial c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

ベクトルによる微分

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)^\top$$

$$\frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left(\frac{\partial f(\boldsymbol{\beta})}{\partial \beta_1}, \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_2}, \ldots \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_n} \right)^\top$$

ベクトルによる微分 (公式)

成り立つかを確認してみること

$$f(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x} + b \quad \Rightarrow \quad \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{a}$$

$$f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 + b \quad \Rightarrow \quad \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = 2\boldsymbol{x}$$

$$f(\boldsymbol{x}) = (\boldsymbol{A}\boldsymbol{x} + \boldsymbol{c})^\top (\boldsymbol{B}\boldsymbol{x} + \boldsymbol{d})$$

$$\Rightarrow \quad \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = 2\boldsymbol{A}^\top \boldsymbol{B}\boldsymbol{x} + \boldsymbol{A}^\top \boldsymbol{d} + \boldsymbol{B}^\top \boldsymbol{c}$$

重回帰 (二回目)

$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^3$ に平面を当てはめよ
$\|\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \to \min$

$$\boldsymbol{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ \vdots & & \vdots \\ x_n & y_n & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} \|\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}\|^2 &= \frac{\partial}{\partial \boldsymbol{\beta}} (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \frac{\partial}{\partial \boldsymbol{\beta}} \left(\boldsymbol{z}^\top \boldsymbol{z} - 2\boldsymbol{z}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \right) \\ &= -2\boldsymbol{X}^\top \boldsymbol{z} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \end{aligned}$$

バリエーションその１：多項式回帰

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ に2次曲線を当てはめよ

$$J = \sum_{i=1}^{n} \Bigl(y_i - (ax_i^2 + bx_i + c)\Bigr)^2 \to \min$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & & \vdots \\ x_n^2 & x_n & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} a \\ b \\ c \end{bmatrix} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

$\boldsymbol{\beta}$ に関して線形な曲線や

k次曲線でも要領は同じ

バリエーションその２：リッジ回帰

$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^3$ に平面を当てはめよ

$\|\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \to \min$

$$\boldsymbol{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ \vdots & & \vdots \\ x_n & y_n & 1 \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left[\|\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \right]$$

$$= -2\boldsymbol{X}^\top \boldsymbol{z} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta} = \boldsymbol{0}$$

$$\Rightarrow \quad \boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{z}$$

$\lambda \neq 0$ なら
逆行列が常に存在！

やってみる？

式を計算するのが面倒なら正規方程式作るとこまで

練習問題１
とある２次関数が３点 (1,3), (-1,7),(3,7) を通るとき、
この関数の式を求めよ

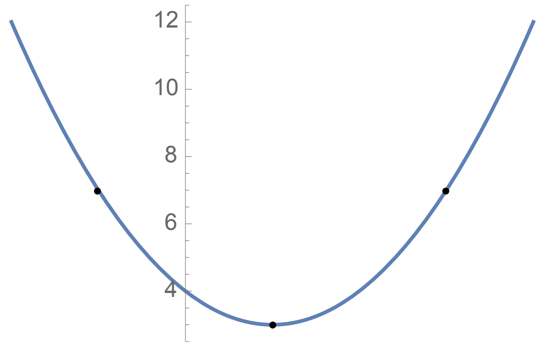
練習問題2
３点 (1,2,3), (1,-1,1),(3,1,2) を通る平面の式を求めよ

やってみる？

式を計算するのが面倒なら正規方程式作るところまで

練習問題１
とある２次関数が３点 (1,3), (-1,7),(3,7) を通るとき、
この関数の式を求めよ

x^2	x	y
1	1	3
1	-1	7
9	3	7



$$X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 9 & 3 & 1 \end{pmatrix}; \; y = \{3, 7, 7\};$$

```
Inverse[Transpose[X].X].Transpose[X].y
{1, -2, 4}
```

$$y = x^2 - 2x + 4$$

やってみる？

式を計算するのが面倒なら正規方程式作るとこまで

練習問題2

３点 (1,2,3), (1,-1,1),(3,1,2) を通る平面の式を求めよ

x	y	z
1	2	3
1	-1	1
3	1	2

$$X = \begin{pmatrix} 1 & 2 & 1 \\ 1 & -1 & 1 \\ 3 & 1 & 1 \end{pmatrix}; \; y = \{3, 1, 2\};$$

```
Inverse[Transpose[X].X].Transpose[X].y
```

$$\left\{-\frac{1}{6}, \frac{2}{3}, \frac{11}{6}\right\}$$

$$z = -\frac{1}{6}x + \frac{2}{3}y + \frac{11}{6}$$