http://www.math.sci.ehime-u.ac.jp/~matsuura/

重回帰分析

Ver.2022.09.04

愛媛大学
データサイエンスセンター（CDSE）
理工学研究科／理学部

まつうら まさや
松浦　真也

2020年4月設立

CDSE
Center for Data Science, Ehime University

重回帰分析（目的）

【例】「体脂肪率」は「体重」と「身長」からどの程度推定可能か？

データ出典：The Data And Story LibraryのBodyfat
https://dasl.datadescription.com/datafile/bodyfat/

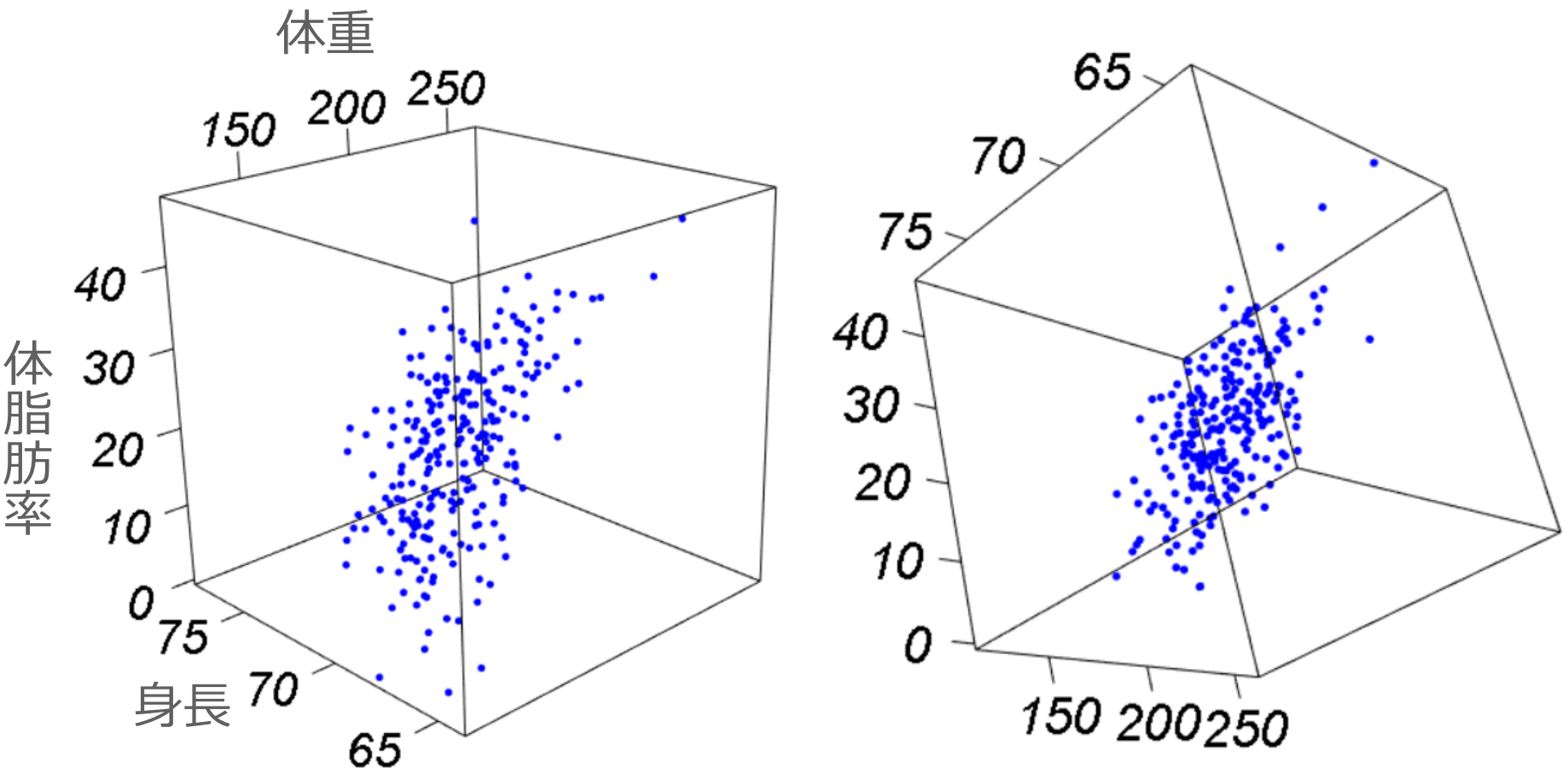
人物（男）	体重（ポンド）	身長（インチ）	体脂肪率（%）
1	154.25	67.75	12.3
2	173.25	72.25	6.1
3	154	66.25	25.3
4	184.75	72.25	10.4
5	184.25	71.25	28.7
6	210.25	74.75	20.9
7	181	69.75	19.2
8	176	72.5	12.4
⋮	⋮	⋮	⋮
250	207.5	70	31.9

重回帰分析（目的）

【例】「体脂肪率」は「体重」と「身長」からどの程度推定可能か？

データ出典：The Data And Story LibraryのBodyfat

https://dasl.datadescription.com/datafile/bodyfat/



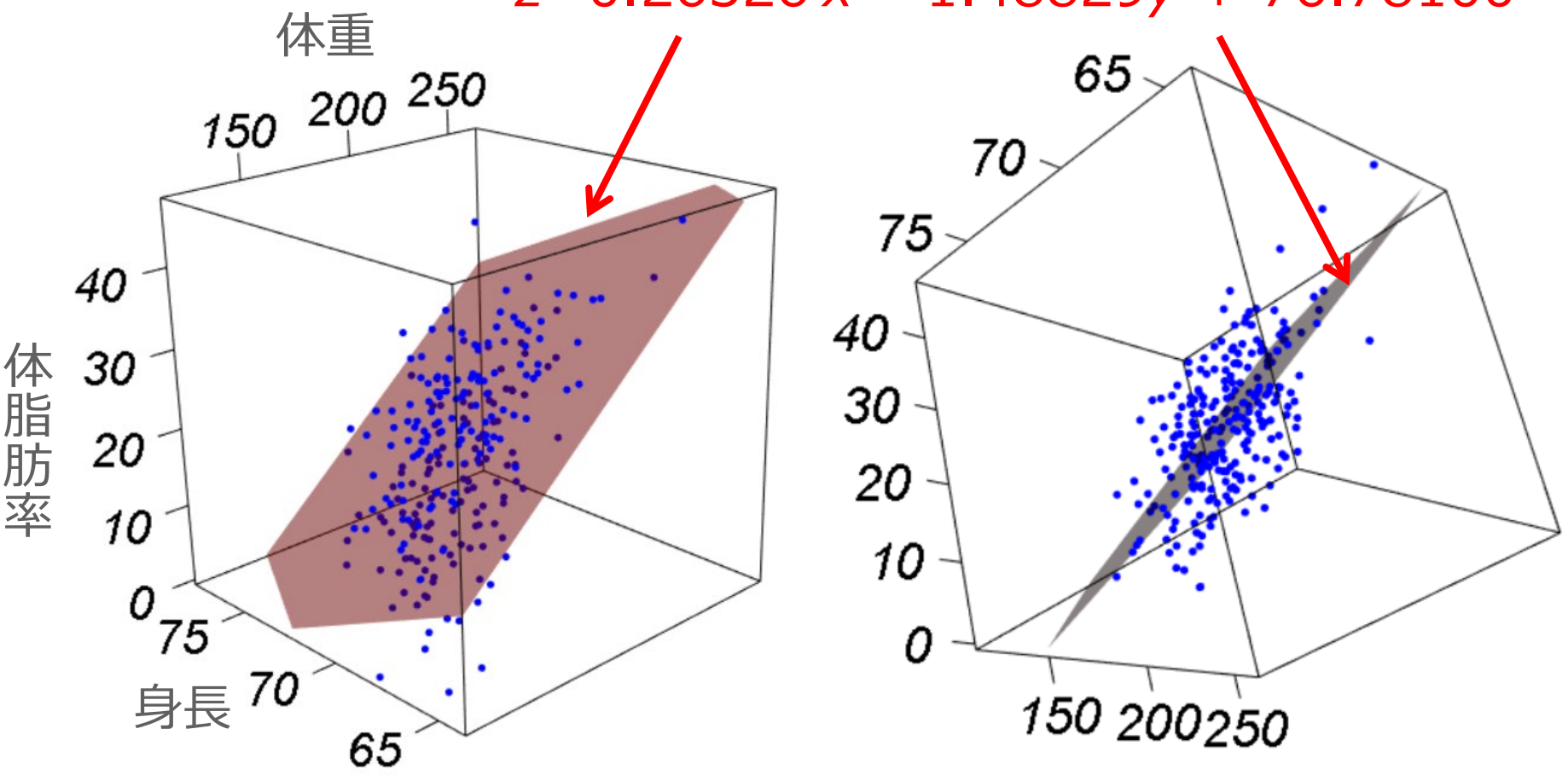
重回帰分析（目的）

【例】「体脂肪率」は「体重」と「身長」からどの程度推定可能か？

データ出典：The Data And Story LibraryのBodyfat

https://dasl.datadescription.com/datafile/bodyfat/

$z = 0.26326\,x - 1.48829y + 76.78100$



重回帰分析（目的）

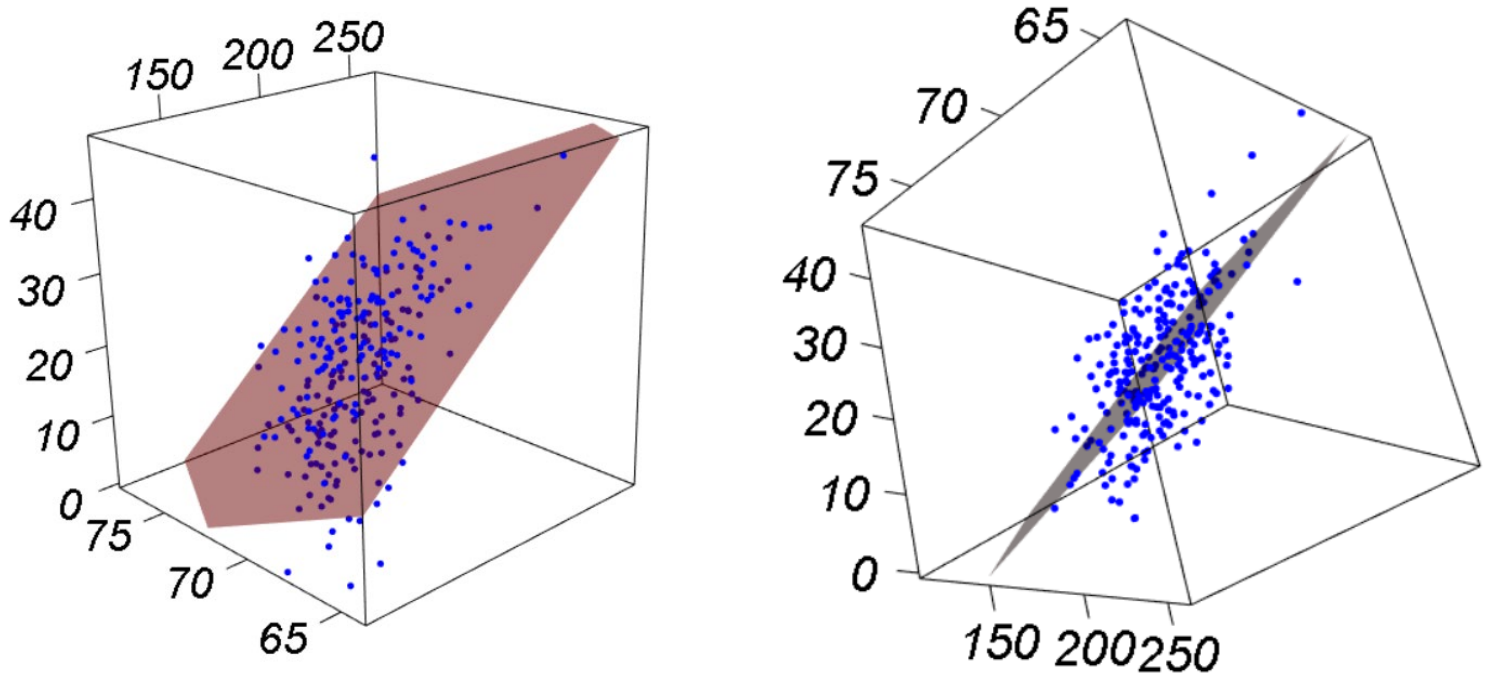
【例】「体脂肪率」は「体重」と「身長」からどの程度推定可能か？

回帰式

偏回帰係数

$$z = 0.26326x - 1.48829y + 76.78100$$

目的変数　説明変数　説明変数



z=0.26326 x − 1.48829y + 76.78100

重回帰分析（最小２乗法）

人物 i	体重 x_{i1}	身長 x_{i2}	体脂肪率 y_i
1	154.25	67.75	12.3
2	173.25	72.25	6.1
⋮	⋮	⋮	⋮
n	207.5	70	31.9

※一般化を見越して、x、y、z ではなく、x_1、x_2、y とする

人物に依存しない未知の定数（これを推定）

重回帰モデル

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$$

y_i：人物 i の体脂肪率　x_{i1}：人物 i の体重　x_{i2}：人物 i の身長　e_i：残差

$(1 \leq i \leq n)$

残差平方和 $S_e = \sum_{i=1}^{n} e_i^2$ が最小になるように $\beta_0, \beta_1, \beta_2$ を定める

重回帰分析（ベクトル表記）

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \quad (1 \leq i \leq n)$$

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ \vec{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \ \vec{x}_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix}, \ \vec{x}_2 = \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix}, \ \vec{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

$$\vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x}_1 + \beta_2 \vec{x}_2 + \vec{e}$$

（ $\vec{y}$ を $\vec{1}, \ \vec{x}_1, \ \vec{x}_2$ の一次結合で近似）

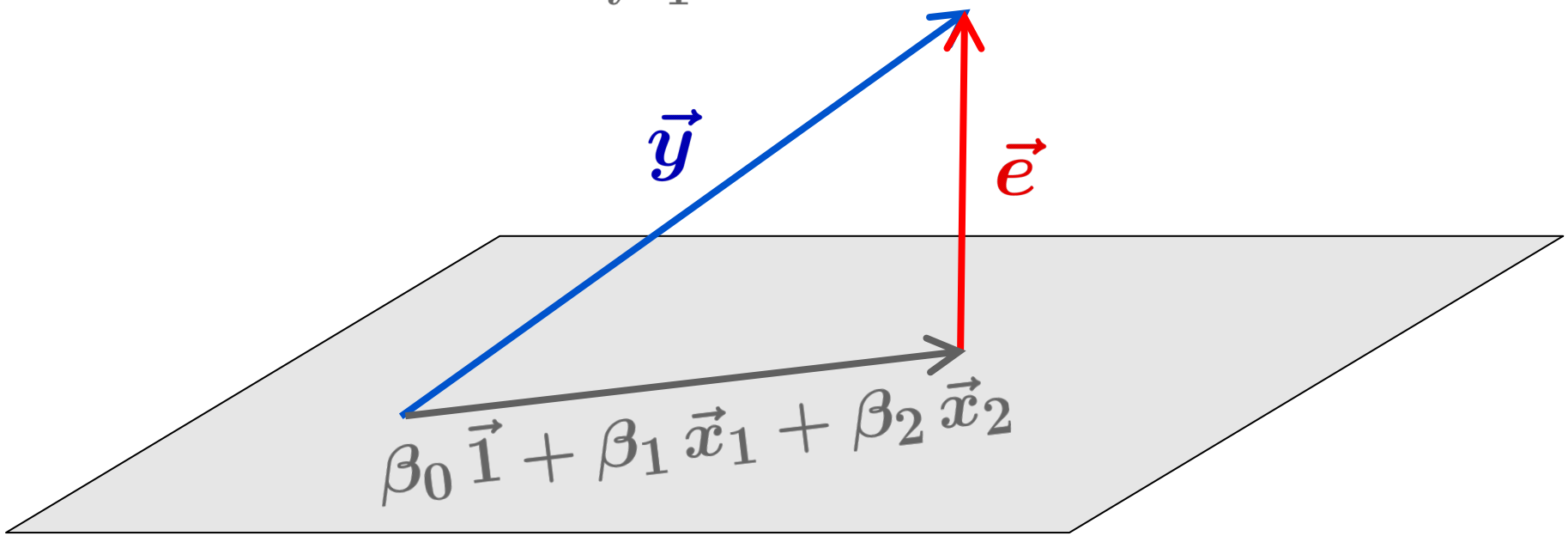
残差平方和 $S_e = \sum_{i=1}^{n} e_i^2 = |\vec{e}|^2$ を最小化

重回帰分析（幾何学的考察）

$$\vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x}_1 + \beta_2 \vec{x}_2 + \vec{e}$$

（$\vec{y}$ を $\vec{1}$, $\vec{x}_1$, $\vec{x}_2$ の一次結合で近似）

残差平方和 $S_e = \sum_{i=1}^{n} e_i^2 = |\vec{e}|^2$ を最小化

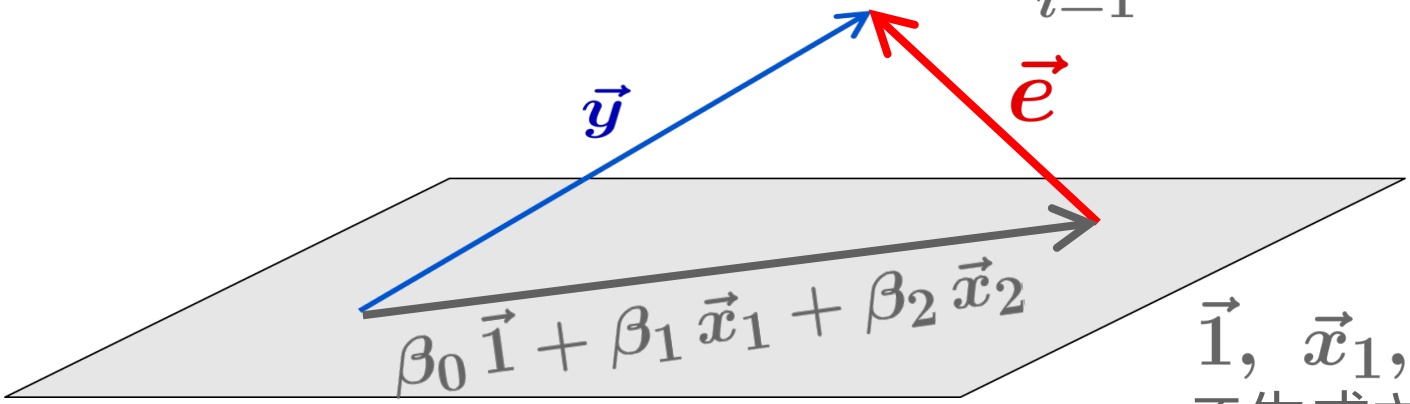


$\vec{1}$, $\vec{x}_1$, $\vec{x}_2$ で生成される空間

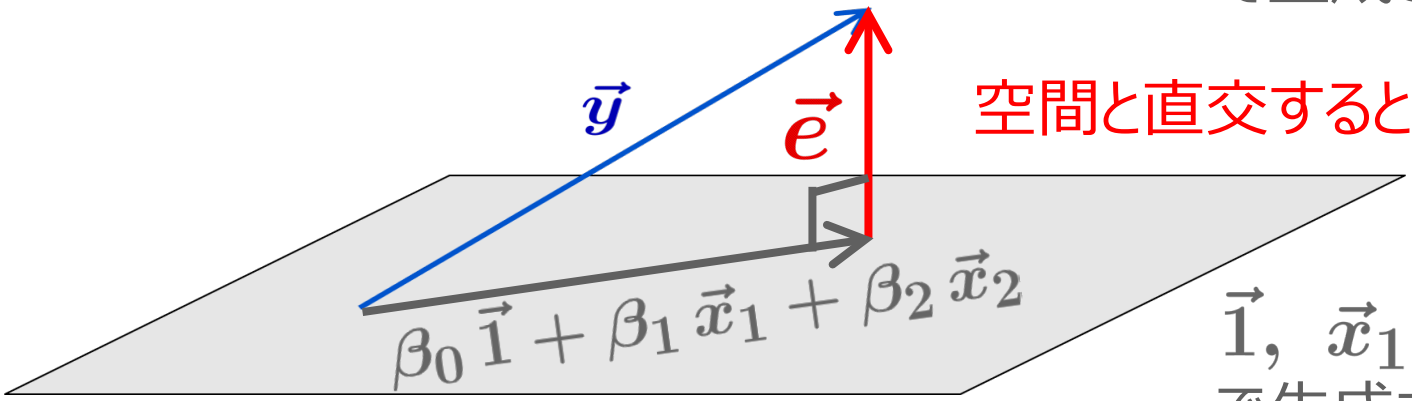
重回帰分析（幾何学的考察）

$$\vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x}_1 + \beta_2 \vec{x}_2 + \vec{e}$$

残差平方和 $S_e = \sum_{i=1}^{n} e_i^2 = |\vec{e}|^2$ を最小化



$\vec{1}$, $\vec{x}_1$, $\vec{x}_2$
で生成される空間



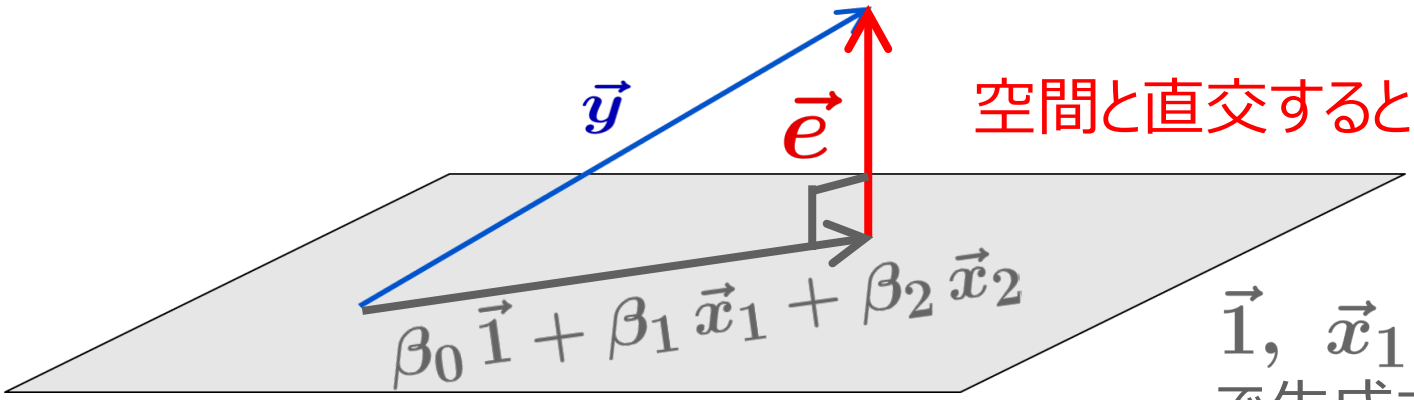
空間と直交するとき $\vec{e}$ 最小

$\vec{1}$, $\vec{x}_1$, $\vec{x}_2$
で生成される空間

重回帰分析（幾何学的考察）

$$\vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x}_1 + \beta_2 \vec{x}_2 + \vec{e}$$

残差平方和 $S_e = \sum_{i=1}^{n} e_i^2 = |\vec{e}|^2$ を最小化



空間と直交するとき $\vec{e}$ 最小

$\vec{1}$, $\vec{x}_1$, $\vec{x}_2$
で生成される空間

内積：$(\vec{1}, \vec{e}) = 0$, $(\vec{x}_1, \vec{e}) = 0$, $(\vec{x}_2, \vec{e}) = 0$

重回帰分析（代表値のベクトル内積表記）

平均 $\overline{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{i1},\ \overline{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{i2},\ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

偏差平方和・偏差積和 $S_{11} = \sum_{i=1}^{n} (x_{i1} - \overline{x}_1)^2,\ S_{22} = \sum_{i=1}^{n} (x_{i2} - \overline{x}_2)^2,$

$S_{12} = S_{21} = \sum_{i=1}^{n} (x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2),\ S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2,$

$S_{1y} = \sum_{i=1}^{n} (x_{i1} - \overline{x}_1)(y_i - \overline{y}),\quad S_{2y} = \sum_{i=1}^{n} (x_{i2} - \overline{x}_2)(y_i - \overline{y})$

内積

$$\frac{(\vec{x}_1, \vec{1})}{n} = \overline{x}_1, \frac{(\vec{x}_2, \vec{1})}{n} = \overline{x}_2, \frac{(\vec{y}, \vec{1})}{n} = \overline{y}, \left(\vec{x}_1 - \overline{x}_1\vec{1}, \vec{x}_1 - \overline{x}_1\vec{1}\right) = S_{11},$$

$$\left(\vec{x}_2 - \overline{x}_2\vec{1}, \vec{x}_2 - \overline{x}_2\vec{1}\right) = S_{22},\ \left(\vec{x}_1 - \overline{x}_1\vec{1}, \vec{x}_2 - \overline{x}_2\vec{1}\right) = S_{12} = S_{21},$$

$$\left(\vec{x}_1 - \overline{x}_1\vec{1}, \vec{y} - \overline{y}\vec{1}\right) = S_{1y},\quad \left(\vec{x}_2 - \overline{x}_2\vec{1}, \vec{y} - \overline{y}\vec{1}\right) = S_{2y}$$

重回帰分析（回帰係数の計算）

重回帰モデル $\vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x}_1 + \beta_2 \vec{x}_2 + \vec{e}$

$$\Downarrow$$

$$\overline{y} = \frac{(\vec{y}, \vec{1})}{n} = \beta_0 + \beta_1 \overline{x}_1 + \beta_2 \overline{x}_2$$

$$\Downarrow$$

$$\vec{y} - \overline{y}\vec{1} = \beta_1 (\vec{x}_1 - \overline{x}_1 \vec{1}) + \beta_2 (\vec{x}_2 - \overline{x}_2 \vec{1}) + \vec{e}$$

$$\Downarrow$$

$$\begin{cases} S_{1y} = (\vec{x}_1 - \overline{x}_1 \vec{1}, \vec{y} - \overline{y}\vec{1}) = \beta_1 S_{11} + \beta_2 S_{12} \\ S_{2y} = (\vec{x}_2 - \overline{x}_2 \vec{1}, \vec{y} - \overline{y}\vec{1}) = \beta_1 S_{21} + \beta_2 S_{22} \end{cases}$$

$$\Downarrow$$

$$\begin{bmatrix} S_{1y} \\ S_{2y} \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\Downarrow$$

$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}^{-1} \begin{bmatrix} S_{1y} \\ S_{2y} \end{bmatrix}, \quad \beta_0 = \overline{y} - \beta_1 \overline{x}_1 - \beta_2 \overline{x}_2$$

重回帰分析（結論）

回帰モデル $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \quad (1 \leq i \leq n)$

残差平方和 $S_e = \sum_{i=1}^{n} e_i^2$ が最小になるように $\beta_0, \beta_1, \beta_2$ を定める。

このときの $\beta_0, \beta_1, \beta_2$ を $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ とすると、

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}^{-1} \begin{bmatrix} S_{1y} \\ S_{2y} \end{bmatrix}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

ただし、

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{i1},\ \bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{i2},\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,\ S_{11} = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2,$$

$$S_{22} = \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2,\ S_{12} = S_{21} = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2),$$

$$S_{1y} = \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(y_i - \bar{y}),\ S_{2y} = \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)(y_i - \bar{y})$$

重回帰分析（決定係数）

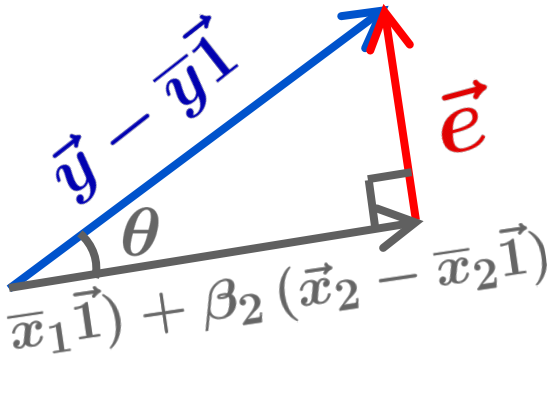
回帰式（の変形版）　$(\vec{1}, \vec{e}) = 0,\ (\vec{x}_1, \vec{e}) = 0,\ (\vec{x}_2, \vec{e}) = 0$

$$\vec{y} - \overline{y}\vec{1} = \beta_1 (\vec{x}_1 - \overline{x}_1\vec{1}) + \beta_2 (\vec{x}_2 - \overline{x}_2\vec{1}) + \vec{e}$$

$\vec{y}$ の変動　　$\vec{x}_1, \vec{x}_2$ の変動で決まる部分　　残差

決定係数（寄与率）の定義

$$R^2 = \frac{(\vec{x}_1, \vec{x}_2\text{の変動で決まる部分の大きさ})^2}{(\vec{y}\text{の変動の大きさ})^2}$$



$$R^2 = \frac{1}{S_{yy}} \begin{bmatrix} \beta_1 & \beta_2 \end{bmatrix} \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{12} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

$$= \frac{1}{S_{yy}} \begin{bmatrix} S_{1y} & S_{2y} \end{bmatrix} \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{12} \end{bmatrix}^{-1} \begin{bmatrix} S_{1y} \\ S_{2y} \end{bmatrix} = \cos^2\theta$$

※ $0 \leq R^2 \leq 1$

重回帰分析（一般の場合）

【例】「体脂肪率」は「体重」と「身長」などからどの程度推定可能か？

データ出典：The Data And Story LibraryのBodyfat
https://dasl.datadescription.com/datafile/bodyfat/

人物	体重	身長	胸囲	…	手首	年齢	体脂肪率
1	154.25	67.75	93.1	…	17.1	23	12.3
2	173.25	72.25	93.6	…	18.2	22	6.1
3	154	66.25	95.8	…	16.6	22	25.3
4	184.75	72.25	101.8	…	18.2	26	10.4
5	184.25	71.25	97.3	…	17.7	24	28.7
6	210.25	74.75	104.5	…	18.8	24	20.9
7	181	69.75	105.1	…	17.7	26	19.2
8	176	72.5	99.6	…	18.8	25	12.4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
250	207.5	70	112.4	…	20.9	74	31.9

重回帰分析（一般の場合）

回帰モデル　$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + e_i \quad (1 \le i \le n)$

残差平方和 $S_e = \sum_{i=1}^{n} e_i^2$ が最小になるように β_j $(0 \le j \le m)$ を定める。

このときのβ_jを$\hat{\beta}_j$とすると、

偏回帰係数：$\begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{bmatrix} = \begin{bmatrix} S_{11} & \cdots & S_{1m} \\ \vdots & \ddots & \vdots \\ S_{m1} & \cdots & S_{mm} \end{bmatrix}^{-1} \begin{bmatrix} S_{1y} \\ \vdots \\ S_{my} \end{bmatrix}$, $\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{m} \hat{\beta}_j \bar{x}_j$

ただし、$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, $S_{jk} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$,

$S_{jy} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(y_i - \bar{y})$

決定係数：$R^2 = \frac{1}{S_{yy}} \begin{bmatrix} S_{1y} & \cdots & S_{my} \end{bmatrix} \begin{bmatrix} S_{11} & \cdots & S_{1m} \\ \vdots & \ddots & \vdots \\ S_{m1} & \cdots & S_{mm} \end{bmatrix}^{-1} \begin{bmatrix} S_{1y} \\ \vdots \\ S_{my} \end{bmatrix}$

重回帰分析（留意事項）

・回帰分析は、目的変数の値を説明変数から推測するために用いられる。
しかし、回帰分析で、因果関係が示せるわけではない。

例：街中で半袖の人の割合から、冷麺の売上高が推測できたとしても、
服装が原因で冷麺の売上高が決まるわけではない（気温が原因）。

・（線形の）回帰分析では、線形な関係しか反映されない。
例えば、$y=x^2$ などの関係を調べるには、非線形回帰分析が必要。

・高度な解析の前に、散布図を描くなど、直感的な把握も大切。

・説明変数の数は、多ければ多い方が良いというものではない。

・せっかく数学を専門的に学んだのなら、単に統計的視点だけでなく、
線形代数や幾何学的な視点での理解も深めて欲しい。
（それができるのが、数学を学んだ強み。）

重回帰分析（参考文献）

参考文献

・体脂肪率のデータ

The Data And Story Library, Bodyfat,
https://dasl.datadescription.com/datafile/bodyfat/
（最終閲覧日 2022年9月4日）