最小二乗法による線形重回帰分析

明治大学 理工学部 応用化学科
データ化学工学研究室　金子 弘昌

最小二乗法による線形重回帰分析

✓Multiple Linear Regression (MLR)

✓Ordinary Least Squares (OLS)

✓Classical Linear Regression (CLS)

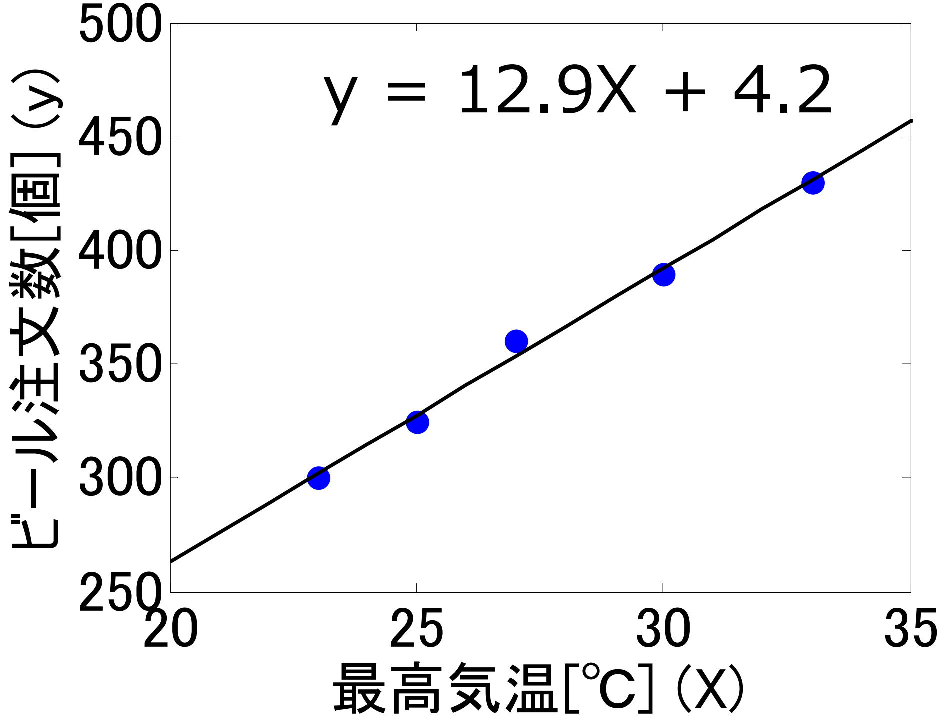
などと呼ばれます

回帰分析ってなに？

目的変数（y）と説明変数（X）の関係をモデル化し、
Xによってyがどれだけ説明できるのかを定量的に分析すること

✓例

- 目的変数 (y)
  - ビール注文数[個]
- 説明変数 (X)
  - 最高気温[℃]



どうやってモデル化する（式を作る）のか？

説明変数が２つのときの線形重回帰分析

$$
\begin{aligned}
\mathrm{y} &= b_0 + \mathrm{x}_1 b_1 + \mathrm{x}_2 b_2 + \mathrm{f} \\
&= \mathrm{y}_\mathrm{C} + \mathrm{f}
\end{aligned}
$$

$$\left(\mathrm{y}_\mathrm{C} = b_0 + \mathrm{x}_1 b_1 + \mathrm{x}_2 b_2\right)$$

y： 目的変数

$\mathrm{x}_1, \mathrm{x}_2$： 説明変数 (記述子)

b_0： 定数項

b_1, b_2： 回帰係数

y_C： yの、xで表すことができる部分

f： yの、xで表すことができない部分
(誤差、残差)

オートスケーリング(標準化)のメリット

$$\mathrm{y} = b_0 + \mathrm{x}_1 b_1 + \mathrm{x}_2 b_2 + \mathrm{f}$$
$$= \mathrm{y}_\mathrm{C} + \mathrm{f}$$

$$\left(\mathrm{y}_\mathrm{C} = b_0 + \mathrm{x}_1 b_1 + \mathrm{x}_2 b_2 \right)$$

y, x_1, x_2 にオートスケーリングを行えば、$b_0 = 0$

よって、 $\mathrm{y} = \mathrm{x}_1 b_1 + \mathrm{x}_2 b_2 + \mathrm{f}$

サンプルが n 個のとき

$$\mathrm{y} = \mathrm{x}_1 b_1 + \mathrm{x}_2 b_2 + \mathrm{f}$$

サンプル n 個のとき、

$$y^{(1)} = x_1^{(1)} b_1 + x_2^{(1)} b_2 + f^{(1)}$$
$$y^{(2)} = x_1^{(2)} b_1 + x_2^{(2)} b_2 + f^{(2)}$$
$$\vdots$$
$$y^{(n)} = x_1^{(n)} b_1 + x_2^{(n)} b_2 + f^{(n)}$$

$y^{(i)}$: i 番目のサンプルにおける目的変数の値

$x_j^{(i)}$: i 番目のサンプルにおける j 番目の説明変数の値

$f^{(i)}$: i 番目のサンプルにおける誤差の値

行列で表す

$$y^{(1)} = x_1^{(1)} b_1 + x_2^{(1)} b_2 + f^{(1)}$$

$$y^{(2)} = x_1^{(2)} b_1 + x_2^{(2)} b_2 + f^{(2)}$$

$$\vdots$$

$$y^{(n)} = x_1^{(n)} b_1 + x_2^{(n)} b_2 + f^{(n)}$$

$$\Rightarrow \quad \begin{aligned} \mathbf{y} &= \mathbf{x}_1 b_1 + \mathbf{x}_2 b_2 + \mathbf{f} \\ &= \mathbf{X}\mathbf{b} + \mathbf{f} \end{aligned}$$

$$\mathbf{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix}, \ \mathbf{x}_1 = \begin{pmatrix} x_1^{(1)} \\ x_1^{(2)} \\ \vdots \\ x_1^{(n)} \end{pmatrix}, \ \mathbf{x}_2 = \begin{pmatrix} x_2^{(1)} \\ x_2^{(2)} \\ \vdots \\ x_2^{(n)} \end{pmatrix}, \ \mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2] = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \\ \vdots & \vdots \\ x_1^{(n)} & x_2^{(n)} \end{pmatrix},$$

$$\mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

回帰係数を求めたい

$$\mathbf{y} = \mathbf{x}_1 b_1 + \mathbf{x}_2 b_2 + \mathbf{f}$$
$$= \mathbf{X}\mathbf{b} + \mathbf{f}$$

b_1, b_2、つまり $\mathbf{b}$ を求めたい

最小二乗法

残差 $f^{(i)}$ の二乗和 (G) が最小という条件で $\mathbf{b}$ を求める方法

$$G = \sum_{i=1}^{n} f_i^2 = \sum_{i=1}^{n} \left(y^{(i)} - b_1 x_1^{(i)} - b_2 x_2^{(i)} \right)^2$$

最小値を取る

↓

極小値を取る

↓

G を b_1, b_2 で偏微分したものが 0

誤差の二乗和を回帰係数で偏微分して 0

$$\frac{\partial G}{\partial b_1} = -2\sum_{i=1}^{n} x_1^{(i)}\left(y_i - b_1 x_1^{(i)} - b_2 x_2^{(i)}\right) = 0$$

$$\frac{\partial G}{\partial b_2} = -2\sum_{i=1}^{n} x_2^{(i)}\left(y_i - b_1 x_1^{(i)} - b_2 x_2^{(i)}\right) = 0$$

まとめて行列で表すと、

$$\begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(n)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(n)} \end{pmatrix} \begin{pmatrix} x_1^{(1)} & x_2^{(1)} \\ x_1^{(2)} & x_2^{(2)} \\ \vdots & \vdots \\ x_1^{(n)} & x_2^{(n)} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(n)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(n)} \end{pmatrix} \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

回帰係数、ついに求まる

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

両辺に左から $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ の逆行列 $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ を掛ける

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

回帰モデルの精度の指標 r^2

✓r^2 (決定係数、説明分散)

- 1に近いほど精度の高い回帰モデル
- 相関係数 r を二乗したものとは異なる
- 異なるデータセットの間で r^2 を比較してはいけない

$$r^2 = 1 - \frac{\sum_{i=1}^{n} \left(y^{(i)} - y_{\mathrm{C}}{}^{(i)} \right)^2}{\sum_{i=1}^{n} \left(y^{(i)} - y_{\mathrm{A}} \right)^2}$$

$y^{(i)}$: i 番目のサンプルにおける目的変数の値

$y_{\mathrm{C}}{}^{(i)}$: i 番目のサンプルにおける目的変数の計算値

y_{A} : 目的変数の平均値

n : サンプル数

回帰モデルの精度の指標 RMSE

✓RMSE (Root Mean Square Error)

- 回帰モデルの誤差の指標
- 0に近いほど精度の高い回帰モデル
- 異なるデータセットの間で RMSE を比較してはいけない

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y^{(i)} - {y_{\mathrm{C}}}^{(i)}\right)^2}{n}}$$

$y^{(i)}$: i 番目のサンプルにおける目的変数の値

${y_{\mathrm{C}}}^{(i)}$: i 番目のサンプルにおける目的変数の計算値

n : サンプル数

回帰モデルの精度の指標 MAE

✓MAE (Mean Absolute Error)

- 回帰モデルの誤差の平均
- 0に近いほど精度の高い回帰モデル

$$MAE = \frac{\sum_{i=1}^{n} \left| y^{(i)} - y_{\mathrm{C}}{}^{(i)} \right|}{n}$$

$y^{(i)}$: i 番目のサンプルにおける目的変数の値

$y_{\mathrm{C}}{}^{(i)}$: i 番目のサンプルにおける目的変数の計算値

n : サンプル数