多変量解析
第 4 回: 回帰分析 (3) 重回帰分析

藤本 衡

2018 年 5 月 2 日

本日の内容

- 2 以上の説明変数を扱う—重回帰分析
- 重回帰分析の定式化: ベクトル・行列表現
- 重回帰係数の解を求める

N 個の説明変数がある回帰モデル

$$
\begin{aligned}
Y &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_N X_N + \varepsilon \\
&= \alpha + \sum_{j=1}^{N} \beta_j X_j + \varepsilon
\end{aligned} \tag{1}
$$

n 個体からデータが得られたとき

$(x_{i1}, x_{i2}, \ldots, x_{iN}, y_i)$: **個体 i から得られる観測値**
ε_i: **個体 i に関する観測誤差**

$$
\begin{aligned}
y_1 &= \alpha + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_N x_{1N} + \varepsilon_1 \\
y_2 &= \alpha + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_N x_{2N} + \varepsilon_2 \\
&\vdots \\
y_n &= \alpha + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_N x_{nN} + \varepsilon_n
\end{aligned}
$$

ベクトル・行列を使って表現できそうな？

残差平方和の最小化

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \alpha - \sum_{j=1}^{N} \beta_j x_{ij} \right)^2 \tag{2}$$

は α, β の凸 2 次関数である。$\dfrac{\partial RSS}{\partial \alpha} = 0$ を満たす α は

$$\alpha = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{N} \beta_j x_{ij} \right) = \bar{y} - \sum_{j=1}^{N} \beta_j \bar{x}_j \tag{3}$$

(3) を (2) に代入すると、

$$\mathrm{RSS} = \sum_{i=1}^{n} \left\{ (y_i - \bar{y}) - \sum_{j=1}^{N} \beta_j (x_{ij} - \bar{x}_j) \right\}^2 \tag{4}$$

残差平方和の最小化 (2)

$v_i = y_i - \bar{y},\ \chi_{ij} = x_{ij} - \bar{x}_j$ とおき、以下のように記号を定義する。

$$\boldsymbol{X} = \begin{pmatrix} \chi_{11} & \chi_{12} & \cdots & \chi_{1N} \\ \chi_{21} & \chi_{22} & \cdots & \chi_{2N} \\ \vdots & \vdots & & \vdots \\ \chi_{n1} & \chi_{n2} & \cdots & \chi_{nN} \end{pmatrix}, \quad \boldsymbol{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{pmatrix}$$

このとき

$$\mathrm{RSS} = (\boldsymbol{v} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{v} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{v}^\top\boldsymbol{v} - 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{v} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} \qquad (5)$$

残差平方和の最小化 (3)

(5) 式を β について偏微分し極小点を求めると、

$$\frac{\partial \mathrm{RSS}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^\top \boldsymbol{v} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = 0$$

すなわち、以下の一次方程式を満たす β が、求める解である。

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{v}$$

$\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{C}, \quad \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{v} = \boldsymbol{c}_y$ とおくと以下のように書き換えられる。

$$\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{c}_y \tag{6}$$

よって、(6) 式の解 β^* は次式で得られる。

$$\boldsymbol{\beta}^* = \boldsymbol{C}^{-1}\boldsymbol{c}_y \tag{7}$$

分散共分散行列 (covariance matrix)

C の (k, ℓ) 要素は X_k と X_ℓ の共分散を与える
($k = \ell$ のときは X_k の分散)。C を分散共分散行列と呼ぶ。
C は実対称行列である点に注意。

$$\frac{1}{n}\sum_{i=1}^{n} \chi_{ik}\chi_{i\ell} = \frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)(x_{i\ell} - \bar{x}_\ell) = s_{k,\ell}$$

同様に、$\boldsymbol{c}_y$ の第 k 要素は x_k と y の共分散を与える。

$$\frac{1}{n}\sum_{i=1}^{n} \chi_{ik}\upsilon_k = \frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)(y_k - \bar{y}) = s_{k,y}$$

方程式の解法

- (7) 式は逆行列の計算が必要になる
- C が小さければ大したことはないが、大きいと大変
- 対称行列（エルミート行列）を用いた方程式の例は多い
- 解法もいろいろ提案されている（主に行列分解を用いた方法）

例: LU 分解

C を下三角行列 L と上三角行列 U に分解する $(C = LU)$
$\gamma = U\beta$ とおくと、(6) 式は以下の 2 つの方程式と等価であり、それらは単なる代入の繰り返しで解を得ることができる。

$$L\gamma = c_y, \quad U\beta = \gamma$$

LU 分解の求め方

C の対角・右上三角成分は U の、左下三角成分は L の非零成分になる。なお、L の対角成分は 1 とする（Doolittle 法）。

```
for (k = 0; k < m; k++) {
  for (i = k + 1; i < m; i++) {
    C[i][k] /= C[k][k];
    for (j = k + 1; j < m; j++) {
      C[i][j] -= C[i][k] * C[k][j];
    }
  }
}
```

C の対角成分 C[k][k] が非零でなければならない点に注意。

QR 分解

C を直交行列 Q と上三角行列 R に分解する $(C = QR)$

$$QR\boldsymbol{\beta} = \boldsymbol{c}_y \longrightarrow R\boldsymbol{\beta} = Q^{-1}\boldsymbol{c}_y$$

$Q^{-1} = Q^{\intercal}$ であり（直交行列の定義）、R は上三角行列なので LU 分解と同様に代入によって β の解が得られる。
Q, R を得るにはハウスホルダー変換やグラム・シュミット分解などを用いる。

重回帰分析小テスト

以下のデータが与えられたときに、最小二乗法により $\alpha,\ \beta_1,\ \beta_2$ を求めよ。

日	最高気温 y	最低気温 x_1	日照時間 x_2
1	11.4	-2.0	9.5
2	7.6	-0.8	9.1
3	7.5	-2.4	8.2
4	9.0	-3.1	9.8
5	11.5	-0.7	9.8