最小二乗法に関するメモ*

千葉豪

平成 30 年 1 月 12 日

1 基本的な考え方

1.1 基本中の基本

あるパラメータ x に依存して決まるパラメータ y があるとする[1]。I 個の異なる x の値に対して y の値が得られたものとし、そのセットを (x_1, y_1)、(x_2, y_2)、...、(x_I, y_I) と記述することとする。このふたつのパラメータについて $y = ax + b$ なる関係があると仮定し、この I 個のデータセットから係数 a、b を決めることを考える。

最小二乗法では、以下で定義される G が最小化されるように、これらの係数を決める。

$$G = \sum_{i=1}^{I} \{(ax_i + b) - y_i\}^2 \tag{1}$$

G が最小値をとる場合には、以下の式が満足される。

$$\frac{\partial G}{\partial a} = 0, \tag{2}$$

$$\frac{\partial G}{\partial b} = 0 \tag{3}$$

それではこれらの式に式 (1) を実際に代入してみよう。このとき、式 (2) は以下のように書ける。

$$\frac{\partial G}{\partial a} = \sum_{i=1}^{I} 2\left(ax_i + b - y_i\right) \cdot x_i = 2\left\{a\sum_{i=1}^{I} x_i^2 + b\sum_{i=1}^{I} x_i - \sum_{i=1}^{I} x_i y_i\right\} = 0 \tag{4}$$

同様に、式 (3) は以下のように書ける。

$$\frac{\partial G}{\partial b} = \sum_{i=1}^{I} 2\left(ax_i + b - y_i\right) = 2\left\{a\sum_{i=1}^{I} x_i + b\sum_{i=1}^{I} 1 - \sum_{i=1}^{I} y_i\right\} = 0 \tag{5}$$

式 (4)、(5) は以下のようにまとめられる。

$$a\sum_{i=1}^{I} x_i^2 + b\sum_{i=1}^{I} x_i = \sum_{i=1}^{I} x_i y_i, \tag{6}$$

$$a\sum_{i=1}^{I} x_i + b\sum_{i=1}^{I} 1 = \sum_{i=1}^{I} y_i \tag{7}$$

*/Document/Education/LeastSquare/

[1]例としては、GM 計数管の印加電圧（x に対応）と放射線の計数率（y に対応）が挙げられるであろう。

2 個の未知数 a、b に対して 2 本の方程式が立っているため、これらの未知数を一意的に決めることが出来ることが分かるであろう [2]。

式 (6)、(7) を行列形式で記述すると以下のように書ける。

$$\begin{pmatrix} \sum_{i=1}^{I} x_i^2 & \sum_{i=1}^{I} x_i \\ \sum_{i=1}^{I} x_i & \sum_{i=1}^{I} 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{I} x_i y_i \\ \sum_{i=1}^{I} y_i \end{pmatrix} \tag{8}$$

これを行列 $\boldsymbol{M}$、ベクトル $\boldsymbol{\beta}$、$\boldsymbol{n}$ を用いて以下のように記述する。

$$\boldsymbol{M}\boldsymbol{\beta} = \boldsymbol{n} \tag{9}$$

未知数 a、b を要素に持つベクトル $\boldsymbol{\beta}$ は、行列 $\boldsymbol{M}$ に逆行列が存在すれば（行列 $\boldsymbol{M}$ が non-singular であれば)、以下のように一意的に決めることが出来る [3]。

$$\boldsymbol{\beta} = \boldsymbol{M}^{-1}\boldsymbol{n} \tag{11}$$

もし、$y = ax + b$ ではなく、$y = b$ のように y が x に依存しない振る舞いをすると仮定するならばどうなるであろうか。この場合は、式 (5) において $a = 0$ とすればよく、以下の式が得られる。

$$b\sum_{i=1}^{I} 1 - \sum_{i=1}^{I} y_i = 0 \tag{12}$$

従って、未知数 b は以下のように決まる。

$$b = \frac{\sum_{i=1}^{I} y_i}{\sum_{i=1}^{I} 1} = \frac{\sum_{i=1}^{I} y_i}{I} \tag{13}$$

これは、I 個の y_i の平均をとっていることに他ならない。

1.2 基本

これまでの議論に引き続いて、以下のような行列、ベクトルを定義しよう。

$$\boldsymbol{X} = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_I & 1 \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_I \end{pmatrix} \tag{14}$$

[2] そうとはならない場合もあるがここでは省略。

[3] なお、2×2 の行列の逆行列は以下のように求められる。

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{10}$$

これらの行列、ベクトルを用いて、式 (1) は以下のように書くことが出来る。

$$G = \sum_{i=1}^{I} \{(ax_i + b) - y_i\}^2 = (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}) \tag{15}$$

実は、式 (9) で定義した $\boldsymbol{M}$、$\boldsymbol{n}$ は $\boldsymbol{X}$、$\boldsymbol{y}$ を用いて以下のように記述できる。

$$\boldsymbol{M} = \boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} x_1 & x_2 & \dots & x_I \\ 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_I & 1 \end{pmatrix}, \tag{16}$$

$$\boldsymbol{n} = \boldsymbol{X}^T\boldsymbol{y} = \begin{pmatrix} x_1 & x_2 & \dots & x_I \\ 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_I \end{pmatrix} \tag{17}$$

従って、最小二乗法で与えられた式 (9) は、$\boldsymbol{X}$、$\boldsymbol{y}$ を用いて以下のように書き換えられる。

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y} \tag{18}$$

ここで、データセット (x_1, y_1)、(x_2, y_2)、...、(x_I, y_I) を用いて、以下の方程式を考える。

$$\begin{aligned} ax_1 + b &= y_1, && (19)\\ ax_2 + b &= y_2, && (20)\\ &\dots && (21)\\ ax_I + b &= y_I && (22) \end{aligned}$$

データセットの個々の値は誤差を含んでばらつくため、この方程式を満足する解は基本的にはあり得ない。さて、これらの式は行列形式では以下のように書ける。

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y} \tag{23}$$

前掲の式 (18) は、この方程式に対する正規方程式（Normal equation）と呼ばれる。

2 Meyer の教科書から

Meyer の教科書 [1] における最小二乗法についての記述を以下にまとめる。

観測パラメータ Y が、複数のパラメータ X_n（$n = 1, 2, ..., N$）の線形結合で表せるとすると、Y は以下の式で記述される。

$$Y = \beta_1 X_1 + \cdots + \beta_N X_N \tag{24}$$

ここで我々は、X_n と Y の値を観測から得ることにより、未知の定数（パラメータ）である β_n を求めることを考える [4]。

[4] 前節の例（$y = ax + b$）のように定数項（b）が存在する場合は、X_n のうちひとつが 1.0 で固定されていると考えればよいだろう。

さて、X_n の値が誤差なく求められる（測定される）一方、測定される Y の値は誤差を含んだ形で求められるとする。この場合、式 (24) の代わりに以下の式が書けるであろう。

$$Y = \beta_1 X_1 + \cdots \beta_N X_N + \epsilon \tag{25}$$

ϵ は測定誤差に対応するランダム変数を示す。従って、Y はある確率分布に従う確率変数となる。

ここで、I 回の観測を行ったものとし、$I > N$ とする。y_i が i 番目の観測の結果得られた Y の値であるとするならば、y_i は式 (25) に従い以下のように書くことができるであろう。

$$y_i = \beta_1 x_{1i} + \cdots + \beta_N x_{Ni} + \epsilon_i, \qquad i = 1, 2, \cdots, I \tag{26}$$

ϵ_i は i 番目の観測における測定誤差である。一般的に、測定誤差は、観測回に依存せずある一定の確率分布に従い、その確率分布の期待値はゼロ、また、観測毎に独立であると見做せる。従って、

$$E[\epsilon_i] = 0, \qquad \mathrm{Cov}[\epsilon_i, \epsilon_j] = \begin{cases} \sigma^2 & (i = j), \\ 0 & (i \neq J). \end{cases} \tag{27}$$

と書ける。これらを用いると、観測値 y_i の期待値は

$$E[y_i] = E\left[\sum_{n=1}^{N} \beta_n x_{ni}\right] + E[\epsilon_i] = \sum_{n=1}^{N} \beta_n x_{ni} \tag{28}$$

と書ける。

さて、式 (26) は行列形式では $\boldsymbol{y} = \boldsymbol{X}_{I\times N}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ と書ける。このとき、式 (28) は $E[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta}$ と書ける。一般的に、測定では、異なるパラメータ X_i と X_j の寄与が全ての測定において同一の比率となるようには行わない[5]。従って、行列 $\boldsymbol{X}_{I\times N}$ の全ての列は互いに独立であると言え、$I > N$ を仮定しているので $rank(\boldsymbol{X}_{I\times N}) = N$ となる。

今、我々は、観測値 $\boldsymbol{y}$ からパラメータ $\boldsymbol{\beta}$ を推定することを目的としている。そこで、観測値 $\boldsymbol{y}$ に対して線形である $\boldsymbol{\beta}$ の予測子 $\bar{\boldsymbol{\beta}} = \boldsymbol{M}_{I\times M}\boldsymbol{y}$ を考える。すると、我々が考えるべき問題は、**$\boldsymbol{\beta}$ に対する線形予測子のうち、（分散が最小となる意味で）最良で、不偏な（その期待値が β となる）ものが何であるか**について考えることに置き換わる。

まずは、線形予測子が不偏となる条件、すなわち $E[\bar{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ となる条件について考えよう。線形予測子 $\bar{\boldsymbol{\beta}}$ の期待値は

$$E[\bar{\boldsymbol{\beta}}] = E[\boldsymbol{M}\boldsymbol{y}] = \boldsymbol{M}E[\boldsymbol{y}] = \boldsymbol{M}\boldsymbol{X}\boldsymbol{\beta} \tag{29}$$

と書ける。ここで、$rank(\boldsymbol{X}_{I\times N}) = N$ なので、$\boldsymbol{X}$ の擬似逆行列[6] を $\boldsymbol{X}^\dagger$ とすると $\boldsymbol{X}^\dagger\boldsymbol{X} = \boldsymbol{I}_{N\times N}$ が成り立つ。従って、$\boldsymbol{M} = \boldsymbol{X}^\dagger$ とすれば $E[\bar{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ となることから、線形予測子 $\bar{\boldsymbol{\beta}} = \boldsymbol{X}^\dagger\boldsymbol{y}$ が不偏であることが分かる。

次は $\bar{\boldsymbol{\beta}} = \boldsymbol{X}^\dagger\boldsymbol{y}$ が最良の線形予測子になることを示そう。不偏の線形予測子として $\boldsymbol{\beta}^* = \boldsymbol{L}\boldsymbol{y}$ を仮定すると、前述のように不偏であることから

$$E[\boldsymbol{\beta}^*] = \boldsymbol{L}E[\boldsymbol{y}] = \boldsymbol{L}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta} \tag{32}$$

[5]例えば、全ての測定で $X_j/X_i = \alpha$ であったとするならば、測定データから X_i と X_j の各々に対する β を評価できないことは直感的に理解できるであろう。

[6]行列 $\boldsymbol{A}_{I\times N}$ の URV 分解を

$$\boldsymbol{A}_{I\times N} = \boldsymbol{U}\begin{pmatrix} \boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}_{I\times N}\boldsymbol{V}^T \tag{30}$$

と書いたとき、擬似逆行列 $\boldsymbol{A}^\dagger$ は、

$$\boldsymbol{A}^\dagger_{N\times I} = \boldsymbol{V}\begin{pmatrix} \boldsymbol{C}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}_{N\times I}\boldsymbol{U}^T \tag{31}$$

と書ける。ここで、$rank(\boldsymbol{C}) = rank(\boldsymbol{A})$ である。

が得られる。$\boldsymbol{\beta}$ に依存せずこの式が成り立つためには $\boldsymbol{LX} = \boldsymbol{I}_{N\times N}$ が満足されなければならない。さて、不偏の線形予測子 $\boldsymbol{\beta}^*$ の n 番目の要素の分散は以下のように計算できる。

$$\mathrm{Var}[\beta_n^*] = \mathrm{Var}[\boldsymbol{L}_{n*}\boldsymbol{y}] = \mathrm{Var}\left[\sum_{i=1}^{I} l_{ni}y_i\right] == \sum_{i=1}^{I} l_{ni}^2 \mathrm{Var}(y_i) = \sigma^2 \sum_{i=1}^{I} l_{ni}^2 = \sigma^2 ||\boldsymbol{L}_{n*}||_2^2 \tag{33}$$

$\boldsymbol{L}_{n*}\boldsymbol{X} = \boldsymbol{e}_n^T$ より、β_n^* の分散を最小にする $\boldsymbol{L}_{n*}$ は、$\boldsymbol{z}^T\boldsymbol{X} = \boldsymbol{e}_n^T$ を満足する $\boldsymbol{z}$ のうちそのノルムが最小のものであることが分かる。擬似逆行列に関する理論より、それは $\boldsymbol{z}^T = \boldsymbol{e}_n^T\boldsymbol{X}^\dagger = \boldsymbol{X}_{n*}^\dagger$ と与えられることが分かっている。従って、全ての n について β_n^* の分散を最小にするのは $\boldsymbol{L}_{n*} = \boldsymbol{X}_{n*}^\dagger$ となるため、分散を最小にする不偏線形予測子は $\boldsymbol{\beta} = \boldsymbol{X}^\dagger\boldsymbol{y}$ と与えられることが分かる。

以上をまとめると、以下のようになる。

式 (24) で記述されるように N 個のパラメータ X_n に依存して決まるパラメータ Y があり、これらのパラメータについての I 個の測定データを用いて $\boldsymbol{\beta}$ を推定する場合（ただし $rank(\boldsymbol{X}_{I\times N})$）、不偏かつ分散を最小にするものは以下で与えられる。

$$\boldsymbol{\beta} = \boldsymbol{X}^\dagger\boldsymbol{y} \tag{34}$$

$rank(\boldsymbol{X}_{I\times N}) = N$ のとき、

$$\boldsymbol{X}^\dagger = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T \tag{35}$$

なのでこれを式 (34) に代入すると以下を得る。

$$\boldsymbol{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{36}$$

これは正規方程式 (18) と対応するものであることが分かるであろう。

参考文献

[1] C. Meyer, *Matrix analysis and applied linear algebra,* Society for industrial and Applied Mathematics, Philadelphia (2000).