# 7.　最小2乗法

田中雅博

最適化プログラミング

## 1　1次関数の当てはめ

### 1.1　変数が1つの場合

いま $x$ とそれに関係する $y$ のデータが $N$ 組ある。すなわち、$\{(x_1, y_1), \ldots, (x_N, y_N)\}$ とする。

$x, y$ の組の例

- $x$ が最高気温、$y$ がその日のアイスクリームの売り上げ
- $x$ が身長, $y$ が体重
- $x$ が勉強時間, $y$ がテストの点（$x$ は**決定変数**、$y$ は $x$ に依存する**従属変数**)

これらが直線の上に乗っていれば、$y = ax + b$ と書ける。しかし、一般にはばらつく。そこで、$y_i$ と $ax_i + b$ との誤差を $e_i$ とすると、

$$y_i = ax_i + b + e_i, \quad i = 1, \ldots, N$$

となる。$e_i$ の値はわからない。

そこで、$y$ を $x$ の1次式で近似するという意味で、誤差の2乗を最小にすることを考える。説明変数 $x$ の値 $x_i$ に対して直線上の値は $ax_i + b$ だから、$y_i$ との誤差は

$$e_i = y_i - (ax_i + b), \quad i = 1, \ldots, N$$

したがって、$N$ 組のデータに対する誤差の2乗和は

$$J = \frac{1}{2}\sum_{i=1}^{N} e_i^2 = \frac{1}{2}\sum_{i=1}^{N} (y_i - (ax_i + b))^2 \tag{1}$$

であり、これを $a, b$ の値を変えて最小化する。

この問題は、$x_i, y_i$ のそれぞれの値が既に与えられている定数と見なせば、$a$ と $b$ **を変数として扱うことによる、2変数関数の最小化問題**であるすなわち、停留点（極小値を与える点) を求めるには、

$$\frac{\partial J}{\partial a} = 0, \quad \frac{\partial J}{\partial b} = 0$$

とすればよい。

(1) 式を直接そのまま $a$ で微分して 0 とおくことにより、

$$\sum_{i=1}^{N} (y_i - (ax_i + b)) (-x_i) = 0$$

これより

$$-\sum_{i=1}^{N} y_i x_i + \sum_{i=1}^{N} a x_i x_i + \sum_{i=1}^{N} b x_i = 0$$

すなわち

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

を得る。同様に、(1) 式を直接そのまま $b$ で微分して 0 とおくことにより、

$$\sum_{i=1}^{N} (y_i - (ax_i + b)) = 0$$

よって

$$a \sum_{i=1}^{N} x_i + bN = \sum_{i=1}^{N} y_i \tag{3}$$

を得る。(2) 式と (3) 式を連立させることにより

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & N \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} y_i \end{pmatrix} \tag{4}$$

となり、これ（正規方程式と呼ばれている）を解くことにより、$a$ と $b$ が求まる。

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & N \end{pmatrix}$$

が逆行列をもつとき

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & N \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} y_i \end{pmatrix} \tag{5}$$

となる。

実は、行列とベクトルを使った表現の方が簡潔でわかりやすい。

$$y_i = ax_i + b + e_i, \quad i = 1, \ldots, N$$

を縦に $N$ 個並べてベクトルを作ると、

$$\boldsymbol{y} = M\boldsymbol{\theta} + \boldsymbol{e} \tag{6}$$

と書ける。ここに、

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad M = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} a \\ b \end{pmatrix}, \quad \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{pmatrix}$$

である。また、誤差の 2 乗和は

$$\sum_{i=1}^{N} e_i^2 = \boldsymbol{e}^T \boldsymbol{e}$$

であるから、

$$J = \frac{1}{2}\boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{y} - M\boldsymbol{\theta})^T(\boldsymbol{y} - M\boldsymbol{\theta})$$

となる。一般的に $\boldsymbol{a}$ と $\boldsymbol{b}$ の内積を $(\boldsymbol{a}, \boldsymbol{b})$ と書くと、

$$J = \frac{1}{2}(\boldsymbol{e}, \boldsymbol{e}) = \frac{1}{2}((\boldsymbol{y} - M\boldsymbol{\theta}), (\boldsymbol{y} - M\boldsymbol{\theta}))$$

$$= \frac{1}{2}(\boldsymbol{y}, \boldsymbol{y}) - \frac{1}{2}(\boldsymbol{y}, M\boldsymbol{\theta}) - \frac{1}{2}(M\boldsymbol{\theta}, \boldsymbol{y}) + \frac{1}{2}(M\boldsymbol{\theta}, M\boldsymbol{\theta})$$

$$= \frac{1}{2}(\boldsymbol{y}, \boldsymbol{y}) - (M\boldsymbol{\theta}, \boldsymbol{y}) + \frac{1}{2}(M\boldsymbol{\theta}, M\boldsymbol{\theta})$$

$$= \frac{1}{2}(\boldsymbol{y}, \boldsymbol{y}) - (\boldsymbol{\theta}, M^T\boldsymbol{y}) + \frac{1}{2}(\boldsymbol{\theta}, M^TM\boldsymbol{\theta}) \tag{7}$$

となる。ここで、パラメータに関して最小化を行うために、$\dfrac{\partial J}{\partial \boldsymbol{\theta}} = \boldsymbol{0}$ とおく。$\theta$（変数ベクトル）が現れているのは第 2 項目（1 次式)、第 3 項目（2 次式）である。

1 次式、2 次式の微分公式

定理 1：一般的に変数ベクトル $\boldsymbol{x}$ と定数ベクトル $\boldsymbol{a}$ の内積を $\boldsymbol{x}$ により微分すると

$$\frac{\partial(\boldsymbol{a}, \boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{a}$$

となる。

定理 2：一般的に変数ベクトル $\boldsymbol{x}$ と定数の対称行列 $\boldsymbol{A}$ からなる以下の二次式を $\boldsymbol{x}$ により微分すると

$$\frac{\partial(\boldsymbol{x}, A\boldsymbol{x})}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}$$

これらを (7) 式に適用すると

$$\frac{\partial J}{\partial \boldsymbol{\theta}} = -M^T\boldsymbol{y} + M^TM\boldsymbol{\theta} = \boldsymbol{0}$$

となるので、$M^TM$ が逆行列を持てば

$$\boldsymbol{\theta} = (M^TM)^{-1}M^T\boldsymbol{y} \tag{8}$$

となる。**この式は (5) 式と同等である。**

この式の誘導がわからない人は

上記の (8) 式が最小 2 乗法による未知変数の最適値ということだけ理解しても、使うことはできる。

```
’例題 4.2’
X=[4 15 30 100 200]’;
M=[X ones(5,1)];
Y=[ -17 -4 -7 50 70]’;
```

```
th=inv(M'*M)*M'*Y
plot(X',Y','o')
hold on
X=(0:250)';
M=[X ones(251,1)];
Yest=M*th;
figure(1)
plot(X',Yest','r')
xlabel('x')
ylabel('y')
hold off
```

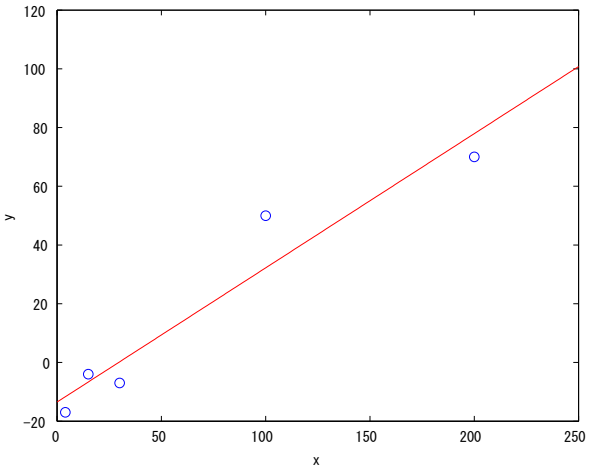


図 1: 変数が 1 つの場合の最小二乗法による推定値

## 1.2 変数が 2 つ以上の場合

いま，変数 $y$ の値を，$l$ 種類の説明変数 $x_1, \ldots, x_l$ を用いて推定したいとする．$x_1, \ldots, x_l$ の値は容易に取得できるものとする．このとき，$y$ はどのようにして推定できるのであろうか．もっとも簡単なモデルは，線形回帰モデル

$$y = a_0 + a_1 x_1 + \cdots + a_l x_l \tag{9}$$

とするものである．係数 $a_0, a_1, \ldots, a_l$ は各変数がそれぞれ $y$ に及ぼす影響の強さを示す．この場合，$y$ は誤差を含まないが，一般的に,「真」の $y$ は $x_1, \ldots, x_l$ で正確に決まることはほとんどなく，誤差を含むことから，(9) 式を改めて

$$y = a_0 + a_1 x_1 + \cdots + a_l x_l + e \tag{10}$$

と書く．ここに，$e$ は誤差項である．これを、重回帰モデルともいう。

重回帰モデルを推定するには，まず，個々のデータに対する (10) 式を縦に並べて改めてベクトルを太字，行列を大文字で表現した式

$$\boldsymbol{y} = M\boldsymbol{\theta} + \boldsymbol{e} \tag{11}$$

を定義しよう．ただし，上付き $y^{(j)}$ は，$j$ 番目のデータに対する $y$ を示す．他の変数の上付きも同様である．また

$$\boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}, \quad M = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_l^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_l^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^{(N)} & x_2^{(N)} & \cdots & x_l^{(N)} \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} a_0 \\ \vdots \\ a_l \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} e^{(1)} \\ e^{(2)} \\ \vdots \\ e^{(N)} \end{bmatrix}$$

とおいた．ここで，誤差の 2 乗和 $\sum_{j=1}^{N}(e^{(j)})^2 = (\boldsymbol{y} - M\boldsymbol{\theta})^\top(\boldsymbol{y} - M\boldsymbol{\theta})$ を最小にすることを考えよう．2 乗誤差の極値を与えるパラメータを、1 変数の場合と同じ手順で求めると、

$$\hat{\boldsymbol{\theta}} = (M^\top M)^{-1} M^\top \boldsymbol{y} \tag{12}$$

で与えられる．

この場合に，注意が必要なのは，データの個数と独立性である．逆行列を構成する行列 $M$ のサイズは $N \times (l+1)$ なので，$N < l+1$ の場合は，$M$ の階数が高々 $N$ であり，$M^\top M$ は決して逆行列をもたない．つまり，データの個数が，データの次元に満たないときは，回帰モデルは最小 2 乗法では係数を一意的に決定できない．

では，データ個数 $N$ が $N \geq l+1$ のときは，いつも $\theta$ が求まるのであろうか．実はここで，多重共線性といわれる問題がある．つまり，説明変数 $1, x_1, \ldots, x_l$ のうちで一次従属な変数があると，いくらデータをたくさん集めても（すなわち，$N$ を大きくしても），$M$ の階数は $l$ 以下になり，$M^\top M$ は逆行列をもたない．また，たとえ逆行列をもつ場合でも，行列式が 0 に近ければ，得られた結果は信頼がおけない．

## 2　多項式の当てはめ

問題

$N$ 個のデータ $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ がある。$y$ を $x$ の 2 次式で近似せよ。
（例えば、ロケットを打ち上げた時、$x_i$ を平面上の距離、$y_i$ が高さ）

当てはめる式を $y = ax^2 + bx + c$ とすると、今までの議論同様、

$$J = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - (ax_i^2 + bx_i + c)\right)^2 \to \min$$

これを最小化するために、

$$y = ax_i^2 + bx_i + c + e$$

とし、全部のデータを縦に並べると

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \quad M = \begin{pmatrix} x_1^2 & x_1 & 1 \\ \vdots & \vdots & \vdots \\ x_N^2 & x_N & 1 \end{pmatrix}$$

$$\boldsymbol{\theta} = (M^T M)^{-1} M^T Y$$

```
X=[-1 0 0 1]';
Y=[0 -2 -1 0]';
M=[X.^2 X ones(4,1)];
th=inv(M'*M)*M'*Y
figure(3)
plot(X',Y','o')
X=(-1.5:0.1:1.5)';
M=[X.^2 X ones(31,1)];
hold on
Yest=M*th;
plot(X',Yest','r')
xlabel('x')
ylabel('y')
hold off
```

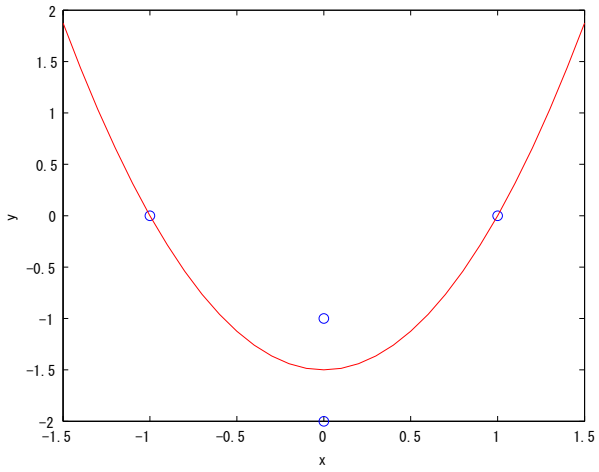


図 2: 二次曲線近似