最小２乗法のはなし

2014.6.12.

内容

最小２乗法のはなし 1
★最小２乗法の考え方 1
★最小２乗法によるパラメータの決定 2
◇パラメータの信頼区間 3
◇重みの異なるデータの取扱い 4
★相関係数・決定係数（最小２乗法を語るもう一つの立場） 5
★実験条件の誤差の影響 5
問題 6

★最小２乗法の考え方

飲料水中のカルシウム濃度をＡさんとＢくんが測定し、Ａさんは 0.67 mmol/L という結果を、Ｂくんは 0.72 mmol/L という結果を出したとしよう。ＡさんとＢくんの技量が同じなら、カルシウム濃度として平均値 0.70 mmol/L を採用することになるだろう。でもＡさんの技量の方がＢくんより高く、Ａさんの分析値の標準偏差が 0.02 mmol/L、Ｂくんの標準偏差が 0.04 mmol/L であることが分かっているとしたら、カルシウム濃度はいくらと推定するのがよいだろうか？

こうした問題を扱うのに、カルシウム濃度としてもっともありそうな値を採用するという考え方がある。正規分布を仮定すると、カルシウム濃度を t とした時、Ａさんが x_A、Ｂくんが x_B を与える確率 $P(x_A, x_B)$は次式で与えられる：

$$P(x_A, x_B) = \frac{1}{2\pi\sigma_A\sigma_B}\exp\left[-\frac{(x_A - t)^2}{2\sigma_A^2} - \frac{(x_B - t)^2}{2\sigma_B^2}\right]$$

ここでＡさんＢくんの測定値の標準偏差を、それぞれσ_A、σ_B とした。カルシウム濃度 t が分からないのだが、この $P(x_A, x_B)$がもっとも大きくなるように t を推定するというのがこの立場である。確率 $P(x_A, x_B)$がもっとも大きくなるのは

$$S = \frac{(x_A - t)^2}{\sigma_A^2} + \frac{(x_B - t)^2}{\sigma_B^2}$$

がもっとも小さくなる時、つまり t の推定値 t_e は、x_A、x_B にそれぞれ分散の逆数だけの重みを付けた平均

$$t_e = \frac{1}{\sigma_A^{-2} + \sigma_B^{-2}}\left(\sigma_A^{-2}x_A + \sigma_B^{-2}x_B\right)$$

であり、その分散（信頼区間の分散）は

$$\left\langle\left\langle t_e^2 \right\rangle\right\rangle = \frac{\sigma_A^2\sigma_B^2}{\sigma_A^2 + \sigma_B^2}$$

で与えられる。先の例ではＡさんの標準偏差 0.02 mmol/L はＢくんの標準偏差 0.04

mmol/L の半分だったから、Ａさんの測定値はＢくんの測定値の４倍の重みがあり、推定値としては 0.68 mmol/L、その標準偏差は 0.018 mmol/L 程度ということになる。測定値を特徴づけるパラメータ t を推定するこの手法を､多変数のパラメーターに拡張したのが最小２乗法とみることができる。

★最小２乗法によるパラメータの決定

実験条件 x を変えて物性値 y の測定を N 回行ったデータ (x_i, y_i)（$i = 1, 2, \ldots, N$ ）があり、y の測定値 y_i にはかたよりがなく精密さが一定で（分散 σ^2）、実験条件 x_i には誤差がないものとしよう。ここで y には x に対し次の線形の関係が成立しているものとし、パラメータ a と b を定めることを考える。

$$y = ax + b$$

最小２乗法は、先のカルシウム濃度の推定で考えたように、測定値のばらつきが正規分布に従うと仮定し、もっとも確率密度が高くなるように a、b を決める手法と考えてよい。ここでは測定値 y_i の分散がすべて等しいとしているので、残差２乗和

$$S = \Sigma(y_i - ax_i - b)^2$$

が最小にすればよい *。それには次の方程式（正規方程式）を解けばよい（S_q は q についての総和 $S_q = \Sigma q_i$）。

$$S_{xx}a + S_x b = S_{xy}$$

$$S_x a + Nb = S_y$$

ここからパラメータは次のように定まる。

$$a = \frac{NS_{xy} - S_x S_y}{NS_{xx} - S_x^2} \quad (= \langle\langle xy\rangle\rangle / \langle\langle x^2\rangle\rangle)$$

$$b = \frac{S_y - aS_x}{N} = \frac{S_{xx}S_y - S_{xy}S_x}{NS_{xx} - S_x^2}$$

最小２乗法で定めた直線は点 $(\bar{x},\ \bar{y})$ を通る †（ここで $\bar{x},\ \bar{y}$ はそれぞれ x、y のデータセットの平均 S_x/N、S_y/N）。また残差２乗和の最小値 $S_{\min}$ は

$$S_{\min} = S_{yy} - (bS_y + aS_{xy})$$

で与えられる。

もっと一般的に実験条件が温度・圧力・濃度等の m 個の要素で与えられる場合には m 成分の実験条件を与えるベクトル $\boldsymbol{x}$ と m 個のパラメータのベクトル $\boldsymbol{a}$ を用い $y = {}^t\boldsymbol{x}\,\boldsymbol{a}$ という関係に、N 個の実験条件 X（X は N 行 m 列の行列）に対する N 個の実験結果 Y を当てはめる問題と見ることができる。残差２乗和は $S = (Y - X\boldsymbol{a})^2$ で表され、正規方程式は次のコンパクトな形で表示できる：

---

* この文脈では、与えられたデータセットについて分散を最小にするようにパラメーターを決めていると考えてもよい（不偏最小分散推定。ガウス-マルコフの定理）。

† かつて計算機等の利用が不便だった時代、簡易な推定法として実験データを２グループに分け、それぞれのデータセットの平均値となる２点を結んで、直線関係をえる手法も行われた

$${}^tXX\boldsymbol{a} = {}^tXY$$

パラメータ $\boldsymbol{a}$ はこの連立方程式の解なので、形式的に次のように書ける：

$$\boldsymbol{a} = ({}^tXX)^{-1}\ {}^tXY$$

また残差２乗和の最小値は次式で与えられる：

$$S_{\min} = {}^tY\,(Y - X\boldsymbol{a})$$

ここでは X として m 個の要素からなる実験条件を考えたが、1 個の要素 x について測定値 y を x の $m-1$ 次の多項式に当てはめる問題は、$\{1, x, x^2, \ldots, x^{m-1}\}$という基底ベクトルの線形結合で関数 $y(x)$への当てはめを行うことと考えれば、同様の扱いが可能であることがわかる。少し具体的に 2 次方程式

$$y = a_0 + a_1x + a_2x^2$$

への当てはめの問題だとすると、正規方程式は次の形に整理できる：

$$\{1\}\{1\}\ a_0 + \{1\}\{x\}\ a_1 + \{1\}\{x^2\}a_2 = \{1\}\{y\}$$

$$\{x\}\{1\}\ a_0 + \{x\}\{x\}\ a_1 + \{x\}\{x^2\}a_2 = \{x\}\{y\}$$

$$\{x^2\}\{x\}\ a_0 + \{x^2\}\{x\}\ a_1 + \{x^2\}\{x^2\}a_2 = \{x^2\}\{y\}$$

ここで$\{x^n\}$は N 個の要素からなるベクトル$(x_1^n, x_2^n, x_3^n, \ldots, x_N^n)$で、$\{u\}\{v\}$は内積を表す。

もとの x の値が$(0, q, 2q, 3q, \ldots, (N-1)q)$ のように等間隔に取られている場合などには、基底を変換して互いに直交する n 次の多項式 $p_n(x)$を用いて$\{1, p_1(x), p_2(x), \ldots, p_{m-1}(x)\}$という基底を用い、正規方程式を対角化して、より見通しのよい形にすることも可能である（等間隔のデータの場合にはチェビシェフの多項式*が用いられる）。

◇パラメータの信頼区間

測定データ y_i の分散をσ^2 とすると、最小 2 乗法で定めたパラメータ a の分散は次のように求めることができる。まず a の表現を整理して

$$a = \frac{NS_{xy} - S_xS_y}{NS_{xx} - S_x^2} = \sum\left(\frac{Nx_i - S_x}{NS_{xx} - S_x^2}\right)y_i$$

とすれば、y_iは互いに独立なので

$$\langle\langle a^2\rangle\rangle = \sum\left(\frac{Nx_i - S_x}{NS_{xx} - S_x^2}\right)^2\sigma^2 = \frac{\sigma^2}{\left(NS_{xx} - S_x^2\right)^2}\sum\left(Nx_i - S_x\right)^2$$

ここで

$$\sum\left(Nx_i - S_x\right)^2 = N^2\sum x_i^2 - 2N\sum x_i\,S_x + \sum S_x^2 = N^2S_{xx} - 2NS_x^2 + NS_x^2 = N\left(NS_{xx} - S_x^2\right)$$

なので

$$\langle\langle a^2\rangle\rangle = \frac{N}{NS_{xx} - S_x^2}\sigma^2 \quad (= \sigma^2/[N\langle\langle x^2\rangle\rangle])$$

---

* 近似理論で登場する有名なチェビシェフの多項式 $\cos(k\cos^{-1}x)$とは別物なので注意。

同様にして

$$\langle\langle b^2\rangle\rangle = \frac{S_{xx}}{NS_{xx} - S_x^2}\sigma^2$$

$$\langle\langle ab\rangle\rangle = -\frac{S_x}{NS_{xx} - S_x^2}\sigma^2$$

を得ることができる。推定パラメータの信頼区間の大きさは、測定値の標準偏差σ に比例し、データ点の数の平方根$\sqrt{N}$ と実験条件の幅（$\sqrt{NS_{xx} - S_x^2}/N$ ）に反比例する。測定データの分散σ^2 が分かっていないときには次式で推定することになる：

$$\sigma^2 = \frac{S_{\min}}{N-2} = \frac{S_{yy} - (bS_y + aS_{xy})}{N-2}$$

測定データを $y = ax + b$ に当てはめてパラメータ a、b を推定する時、できるだけ x の範囲を広くとって測定する（$\sqrt{NS_{xx} - S_x^2}/N$ を大きくする）のが望ましく、切片 b の評価には $x = 0$ 付近の値を取る（S_{xx}を小さくする）のが望ましい。

なお得られるパラメータ a, b の信頼区間は一般に独立でない（$\langle\langle ab\rangle\rangle \neq 0$）。特に原点から x 方向に遠く離れたデータ（$|S_x|$ が大きい）を用いてパラメーターを推定するときには、信頼区間の精密な評価には注意が必要である。

実験条件が m 個の要素で与えられる場合に、パラメータ $\boldsymbol{a}$ の共分散行列$\langle\langle \boldsymbol{a}\ {}^t\boldsymbol{a}\rangle\rangle$は次式で与えられる

$$\langle\langle \boldsymbol{a}\ {}^t\boldsymbol{a}\rangle\rangle = ({}^tXX)^{-1}\ {}^tX\langle\langle Y\ {}^tY\rangle\rangle\ X\ ({}^tXX)^{-1} = ({}^tXX)^{-1}\ \sigma^2$$

◇重みの異なるデータの取扱い

異なる精度を持った測定値 y_i を取り扱う場合、それぞれの測定値の分散をσ_i^2 とすると、$1/\sigma_i^2$ の重みを付けた残差２乗和 S^*を最小にすることを考えればよい：

$$S^* = \sum_i \frac{(y_i - ax_i - b)^2}{\sigma_i^2}$$

化学で出会う典型的な例としては、たとえば有効数字３ケタのデータの線形の式へのあてはめ、相対誤差が一定と見なせるデータのあてはめの問題がある。ただし相対誤差が一定の場合でも、その対数を取ったものへの線形の式のあてはめは、先の分散一定とした取り扱いで十分であることに注意する：

$$\ln [y_0(1 \pm \delta y)] = \ln y_0 \pm \delta y$$

したがって例えばサーミスターの抵抗 R を各温度で有効数字４ケタ程度で測定し、そこからサーミスターの温度依存性を与える関係式 $R_0\exp(B/T)$の Bパラメーターを決める場合には、$\ln (R/\Omega)$と $1/T$について、分散を一定と仮定して前節の扱いに従えばよい。

なお多次元のデータを扱う一般的な場合には、重み因子の行列 W を考え（Wは対角要素が $1/\sigma_i^2$ の対角行列）、残差２乗和は $S = {}^t(Y - X\boldsymbol{a})W(Y - X\boldsymbol{a})$で表され、正規方程式は次の形で表示できる：

$${}^tXWX\boldsymbol{a} = {}^tXWY$$

したがってパラメータは次式で与えられることになる：

$$\boldsymbol{a} = ({}^{t}XWX)^{-1}\,{}^{t}XWY$$

★相関係数・決定係数（最小２乗法を語るもう一つの立場）

観測データの変動を説明する立場からは、次の関係式を想定する（$\bar{x}$, $\bar{y}$ はそれぞれ x、y のデータセットの平均 S_x/N、S_y/N）：

$$y - \bar{y} = a(x - \bar{x})$$

観測データの変動を説明するために条件 x を持ち出すことの当否を問題とするには、この形がはっきりしている（こうした立場では回帰 regression 分析と呼ぶことが多い）。x と y の間の相関の強さを示す量として相関係数がある：

$$r = \frac{NS_{xy} - S_x S_y}{\sqrt{(NS_{xx} - S_x^2)(NS_{yy} - S_y^2)}} \quad (= \langle\langle xy\rangle\rangle / \sqrt{\langle\langle x^2\rangle\rangle\langle\langle y^2\rangle\rangle}\ )$$

x と y のデータ点$(x_i,\ y_i)$間に直線関係が正確に成立すれば相関係数は±1、はずれが大きくなるに従って 0 に近づく。相関係数 r は決定係数（寄与率）r^2 の形で扱われることも多い。

$$r^2 = 1 - \frac{NS_{\min}}{NS_{yy} - S_y^2}$$

決定係数は条件 x に対する依存性を考慮することで、どこまで観測データのゆらぎ（残差２乗和）を説明できたかを示すものといえ、完全に説明できれば 1、依存性を考慮してもゆらぎに変動がない場合には 0 になる。化学では多くの場合、相関があることが自明でパラメーターの推定に重きが置かれるが、そもそもの相関のあるなしについては、決定係数に基づいた検定が必要になる。

★実験条件の誤差の影響

実験条件 x には誤差がないとしてきたが、実際には実験条件にも測定値 y と同程度の誤差が見込まれる場合も多い。こうした場合の線形の関係式のパラメータ推定について考えてみよう。問題を明確にするために先のデータセット$(x_i,\ y_i)$（$i = 1, 2, \ldots, N$）について、

$$x = cy + d$$

という関係を仮定し、x を y の関数として処方箋通りに最小２乗法でパラメーターを決めてみたとしよう。すると勾配 c に注目すると

$$c = \frac{NS_{xy} - S_x S_y}{NS_{yy} - S_y^2}$$

である。さてここで $y = ax + b$ という関係を想定すれば、$a = 1/c$ になりそうである。しかし先に最小２乗法で得た a についての表式をみると

$$ac = r^2 \le 1$$

より逆数関係は成り立たない。いわば x に対し y をプロットするか、y に対し x をプロットするかで $\mathrm{d}y/\mathrm{d}x$ の値が異なるのは、実験条件の誤差をどのように考慮するかによっている。

問題

学生番号＿＿＿＿＿＿＿＿　　氏名＿＿＿＿＿＿＿＿

☆N個の一連のデータ$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$を、測定値$y_i$の分散が$x_i$によらず$\sigma^2$で一定であるとして、$y = ax$という関係式に最小２乗法であてはめることを考える。

$S_{xx} = \sum_i x_i^2$、$S_x = \sum_i x_i$、$S_{yy} = \sum_i y_i^2$、$S_y = \sum_i y_i$、$S_{xy} = \sum_i x_i y_i$を用いて推定される係数aを表せ。また推定される係数aの分散はどのようにあらわされるか？