最小二乗法のシートの作成

学籍番号: XXkcXXX
坂本 直志

平成 16 年 10 月 20 日

1 課題

与えられたデータ x_i, y_i に対して、最小二乗法を用いて x_i, y_i の関係を近似する直線の方程式 $y = ax + b$ の a, b を求める表計算シートを作りなさい。そして、与えられたデータ (A 参照) に対して実際に値を求め、点列と近似直線をグラフに書く手順を説明しなさい。

2 最小二乗法とは

二つ値の組となるデータ $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$ に対して、これらのデータの間に $y = ax + b$ という関係が成り立つと仮定して、a, b を求めることを考える。

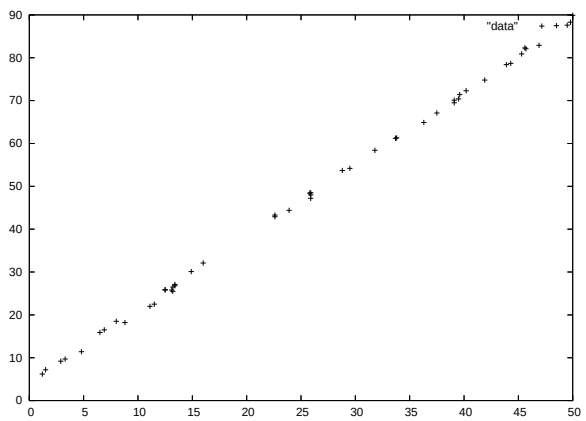

図 1: データ例

もし、得られたデータが正確に $y = ax + b$ の関係を満たしている時は、どのデータもその $y = ax + b$ で表される直線の上に乗っている。従って、データの中から任意に二点を選んで、その二点を通る直線を求めれば良い。

しかし、得られたデータが $y = ax + b$ を正確に満たしてない場合、つまり、得られたデータのどの二点を結ぶ直線に対して、他のデータ点がその直線に乗らない場合は、このような解法は使えない。このような状況では、なるべく各点に近い直線の式を求めることが必要となる。そのためには、一つ直線を決めた時にその直線が各点にどの程度近いか考える必要がある。

a, b に良い値が定まっている場合、各 $x_i.\ y_i$ に対して $ax_i + b$ と y_i は近い値になっているはずである。つまり、$y_i - ax_i - b$ は 0 にはならなくても 0 に近い値になるはずである。従って、各点に対してこの値の二乗を求め和をとると、各点に対する直線の近さを表すことができる。この二乗の和の値は a と b により定まるので、関数 $f(a,b)$ と考えることができる。

$$\begin{aligned} f(a,b) &= \sum_{i=1}^{n} (y_i - ax_i - b)^2 \\ &= \sum_{i=1}^{n} (y_i^2 + a^2 x_i^2 + b^2 - 2ax_i y_i - 2by_i + 2abx_i) \end{aligned} \tag{1}$$

この式を最小にするような a, b は、与えられた各点を良く近似する直線の式 $y = ax + b$ となる。以後はこの $f(a,b)$ の値を小さくする a, b の求め方を考える。$f(a,b)$ は偏微分可能なので、この値を最小にする a, b に対して、$f(a,b)$ は同時に極値をとる。従って、 a, b に対して、偏微分を行い極値になる a, b を求める。

$$\begin{aligned} \frac{\partial f}{\partial a} &= \sum_{i=1}^{n} (2ax_i^2 - 2x_i y_i + 2bx_i) \\ &= 2a \sum_{i=1}^{n} x_i^2 - 2 \sum_{i=1}^{n} x_i y_i + 2b \sum_{i=1}^{n} x_i \end{aligned} \tag{2}$$

$$\begin{aligned}\frac{\partial f}{\partial b} &= \sum_{i=1}^{n}(2b-2y_i+2ax_i)\\ &= 2nb-2\sum_{i=1}^{n}y_i+2a\sum_{i=1}^{n}x_i \end{aligned} \tag{3}$$

式 (2), (3) の値を 0 にする a, b が $f(a,b)$ の値を最小にする。つまり、次の連立方程式を解けば求める a, b が得られる。

$$\begin{cases} \left(\sum_{i=1}^{n}x_i^2\right) & a & + & \left(\sum_{i=1}^{n}x_i\right) & b & = & \sum_{i=1}^{n}x_iy_i \\ \left(\sum_{i=1}^{n}x_i\right) & a & + & n & b & = & \sum_{i=1}^{n}y_i \end{cases} \tag{4}$$

3 解法のあらまし

データをシート上に配置し、最小二乗法を適用するため、上記の連立方程式 (4) を解く。そのためにはまず各係数をデータから求める。求める係数は $n, \sum_{i=1}^{n}x_i, \sum_{i=1}^{n}y_i, \sum_{i=1}^{n}x_iy_i, \sum_{i=1}^{n}x_i^2$ である。n は COUNT 関数で、 $\sum_{i=1}^{n}x_i, \sum_{i=1}^{n}y_i$ は SUM 関数で求める。また $\sum_{i=1}^{n}x_iy_i, \sum_{i=1}^{n}x_i^2$ は x_iy_i, x_i^2 を各データに対して求めておいてから SUM 関数で求める。そしてこれらの値から連立方程式を解き、a と b を求める。

得られた a, b と各 x_i から ax_i+b の値を求める。

最終的に、与えられたデータと、作成した x と $ax+b$ の値でグラフを作成する。

4 シートのデザインとグラフ作成の手順

始めにワークシートでデータの配置を定める。求める a, b の値は A2, B2 に求めることにし、そのために必要な値である $n, \sum_{i=1}^{n}x_i, \sum_{i=1}^{n}y_i, \sum_{i=1}^{n}x_iy_i, \sum_{i=1}^{n}x_i^2$ はそれぞれ A4, B4, C4, D4, E4 に配置する。与えられたデータは x_i については A6 以降の A 列に、対応する y_i は B6 以降の B 列に配置する。そして、それぞれに見出しを付ける (図 2)。

	A	B	C	D	E
1	a	b			
2					
3	N	Sum X	Sum Y	Sum XY	Sum X^2
4					
5	X	Y			
6	8.8	18.2			
7	31.8	58.4			
8	13.1	25.8			
	⋮	⋮			
55	46.9	82.9			

図 2: 初期配置

次に、各データに対して x_iy_i, x_i^2 を計算する。まず、 C6 に =A6*B6 を入れ、ハンドルを使って以後の C 列にこの式をコピーし、各行に対して左側の二つの値の積 x_iy_i を計算する。次に D6 に =A6*A6 を入れ、ハンドルを使って以後の D 列にこの式をコピーし、各行に対して A 列の値の二乗 x_i^2 を計算する。

さて、これらのデータに対して、$n, \sum_{i=1}^{n}x_i, \sum_{i=1}^{n}y_i, \sum_{i=1}^{n}x_iy_i, \sum_{i=1}^{n}x_i^2$ を計算する。以後、データは 6 行目から 55 行目まで入っているとする。A4 にはデータの件数 n を入れるため、=COUNT(A6:A55) を入れる。B4 には $\sum_{i=1}^{n}x_i$ を入れるため =SUM(A6:A55) を入れる。この B4 の式を利用して、C4 には $\sum_{i=1}^{n}y_i$ 、 D4 には $\sum_{i=1}^{n}x_iy_i$ 、E4 には

$\sum_{i=1}^{n} x_i^2$ を入れる。そのために =SUM(A6:A55) の入った B4 のハンドルを横に引き、C4, D4, E4 にコピーする。その結果、 C4 に =SUM(B6:B55)、D4 に =SUM(C6:C55)、E4 に =SUM(D6:D55) が入る。

以上により求めた連立方程式 (4) の係数から a、 b を求める。連立方程式 (4) を解くと次のようになる。

$$
\begin{aligned}
a &= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\
b &= \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i\right)^2 - n\sum_{i=1}^{n} x_i^2}
\end{aligned}
$$

これに n = A4, $\sum_{i=1}^{n} x_i$=B4, $\sum_{i=1}^{n} y_i$=C4, $\sum_{i=1}^{n} x_i y_i$=D4, $\sum_{i=1}^{n} x_i^2$=E4 を適用して、A2 には =(A4*D4-B4*C4)/(A4*E4-B4*B4) を入れ、 B2 には=(B4*D4-C4*E4)/(B4*B4-A4*E4) を入れる。

最後に求めた a, b を使って、$ax_i + b$ の値を求める。E5 に ax+b とラベルを入れる。そして、E6 に =$A$2*A6+$B$2 を入れて、ハンドルを E 列方向に引くことで、各 x_i の値に対して $ax_i + b$ を計算する。完成したワークシートを図 3 に示す。

	A	B	C	D	E
1	a	b			
2	=(A4*D4-B4*C4)/(A4*E4-B4*B4)	=(B4*D4-C4*E4)/(B4*B4-A4*E4)			
3	N	Sum X	Sum Y	Sum XY	Sum X^2
4	=COUNT(A6:A55)	=SUM(A6:A55)	=SUM(B6:B55)	=SUM(C6:C55)	=SUM(D6:D55)
5	X	Y	XY	X$\hat{2}$	ax+b
6	8.8	18.2	=A6*B6	=A6*A6	=$A$2*A6+$B$2
7	31.8	58.4	=A7*B7	=A7*A7	=$A$2*A7+$B$2
8	13.1	25.8	=A8*B8	=A8*A8	=$A$2*A8+$B$2
	⋮	⋮	⋮	⋮	⋮
55	46.9	82.9	=A55*B55	=A55*A55	=$A$2*A55+$B$2

図 3: 表のレイアウト

求めた値に対してグラフを作成する。X の見出しの入っている A5 から Y の最後の値である B55 までの範囲を選択し、さらに ctrl キーを押しながら $ax_i + b$ の入っている E5 から E55 を選択する。そして、「挿入→グラフ」でグラフ作成を行う。グラフの種類は散布図のデータポイント付き折れ線を選び、それ以外は「次へ」をそのまま押し、「完了」を押す。完成したグラフに対して、次の手順により、Y については結んでいる線を消し、ax+b については各点のシンボルを消す。まず、グラフの「凡例」で、「Y」を右クリックし表示されたメニューから「オブジェクトの属性」を選ぶ。「データ系列」画面の「線」タブを選択し、「線の属性」の「スタイル」を「透明」にすることで線を消す。そして、「Ok」を押す。次に「ax+b」を右クリックし、表示されたメニューから「オブジェクトの属性」を選ぶ。そして「線」タブを選択し、「線の終点に使う印」で「選択」メニューから「シンボルなし」を選ぶことで点を消す。から「線形回帰」を選び「Ok」を押す。以上によりデータ点と、近似直線が表示される。

5 まとめ、検討

本レポートでは、与えられた x_i, y_i データに対して直線近似する式を求めるため、最小二乗法を使用した表計算のシートを作成した。今回はあらかじめ直線で近似する事を前提としたが、他の曲線を仮定しても同様の議論は可能と思われる。

また使用した表計算ソフトは OpenOffice.org 1.1.0 だった。グラフ作成において、点だけと線だけのグラフを共存させるのに、線や点を消していくという操作に手間取った。

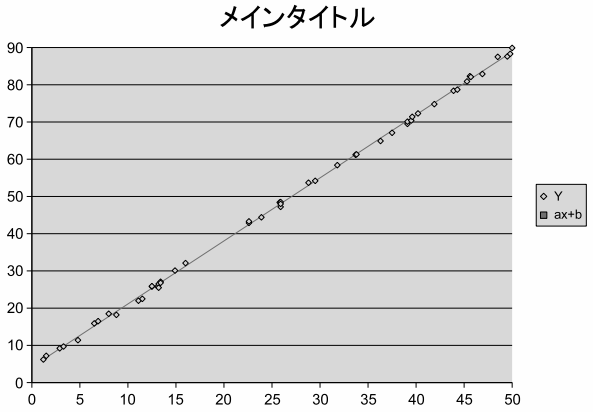

図 4: グラフの作成例

なお、今回はシートを作成するのに全て計算する式を与えたが、表計算ソフトによっては組み込み関数が用意されていたり、グラフの機能で近似直線をひく機能があるものもある。Openoffice.org では SLOPE 関数と INTERCEPT 関数でそれぞれ直線の傾きと Y 切片を求めることができ、また、グラフの作図において回帰曲線を引くオプションもある。

参考文献

[1] 森正武. 数値解析, 共立数学講座, 第 12 巻. 共立出版, 1973.

A 付録: 与えられたデータ

x	y	x	y
8.8	18.2	43.9	78.4
31.8	58.4	45.6	82.3
13.1	25.8	2.9	9.2
36.3	64.9	40.2	72.3
12.5	25.8	25.9	48.0
33.8	61.3	45.7	82.1
44.3	78.7	13.2	26.5
45.3	80.9	39.1	70.1
49.8	88.3	1.2	6.2
50.0	89.9	13.4	27.1
22.6	43.3	14.9	30.1
29.5	54.2	37.5	67.1
33.7	61.2	13.4	26.8
4.8	11.4	28.8	53.7
22.6	42.9	12.5	25.9
6.5	15.9	11.5	22.5
3.3	9.7	48.5	87.5

x	y
1.5	7.2
25.9	47.2
11.1	22.0
16.0	32.1
13.2	25.5
39.6	71.4
6.9	16.5
49.5	87.6
25.8	48.4
39.1	69.5
25.9	48.5
41.9	74.8
8.0	18.5
23.9	44.4
39.5	70.4
46.9	82.9