← PC版は別頁

◇２乗誤差の考え方◇

図１のような幾つかの測定値 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ の近似直線を求めたいとする．

近似直線との「誤差の最大値」を小さくするという考え方では，図２において黄色の●で示したような少数の例外的な値（外れ値）だけで決まってしまい適当でない．

各測定値と予測値の「誤差の総和」が最小になるような直線を求めると各測定値が対等に評価されてよいが，誤差の正負で相殺し合って消えてしまうので，「２乗誤差」が最小となるような直線を求めるのが普通である．すなわち，求める直線の方程式を

$$y=px+q$$

とすると，

$$E(p, q)=(y_1-px_1-q)^2+(y_2-px_2-q)^2+\ldots$$

が最小となるような係数 p,q を求める．

Σ記号で表わすと

$$E(p, q)=\sum_{k=1}^{n}(y_k-px_k-q)^2$$

が最小となるような係数 p,q を求めることになる．

２乗誤差が最小となる係数 p,q を求める方法を「最小２乗法」という．また，このようにして求められた直線 $y=px+q$ を「回帰直線」という．

図１

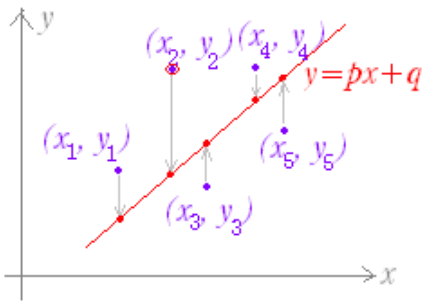
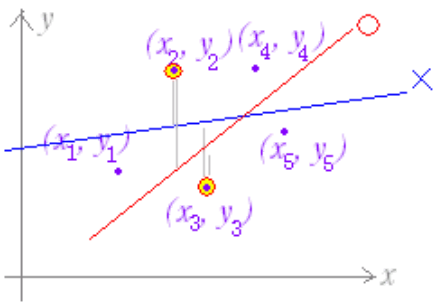


図２



◇最小２乗法◇

３個の測定値 $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ からなる観測データに対して，２乗誤差が最小となる直線 $y=px+q$ を求めてみよう．

$$E(p, q)=(y_1-px_1-q)^2+(y_2-px_2-q)^2+(y_3-px_3-q)^2$$

$$=y_1{}^2+p^2x_1{}^2+q^2-2py_1x_1+2pqx_1-2qy_1$$

$+{y_2}^2+p^2{x_2}^2+q^2-2py_2x_2+2pqx_2-2qy_2$

$+{y_3}^2+p^2{x_3}^2+q^2-2py_3x_3+2pqx_3-2qy_3$

$=p^2({x_1}^2+{x_2}^2+{x_3}^2)-2p(y_1x_1+y_2x_2+y_3x_3)$
$+2pq(x_1+x_2+x_3)$

$-2q(y_1+y_2+y_3)+({y_1}^2+{y_2}^2+{y_3}^2)+3q^2$

※のように考えると

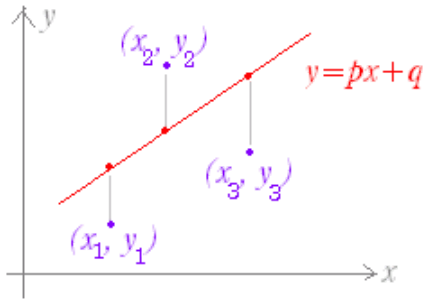
$2p({x_1}^2+{x_2}^2+{x_3}^2)-2(y_1x_1+y_2x_2+y_3x_3)$
$+2q(x_1+x_2+x_3)=0$
$2p(x_1+x_2+x_3)-2(y_1+y_2+y_3)+6q=0$

の解 p,q が，回帰直線 $y=px+q$ となる．

一般に，データが n 個の場合についてΣ記号で表わすと， p,q の連立方程式

$$p\sum_{k=1}^{n}(x_k)^2+q\sum_{k=1}^{n}(x_k)=\sum_{k=1}^{n}(x_ky_k)\cdots(1)$$
$$p\sum_{k=1}^{n}x_k+nq=\sum_{k=1}^{n}y_k\cdots(2)$$

の解が回帰直線 $y=px+q$ の係数 p,q を与える．



※　一般に $E=ap^2+bq^2+cpq+dp+eq+f$
（a,b,c,d,e,f は定数）で表わされる２変数 p,q の関数の極小値は

$$\frac{\partial E}{\partial p}=\frac{\partial E}{\partial q}=0\cdots(*)$$

すなわち，

連立方程式　$2ap+cq+d=0,\ 2bq+cp+e=0$

の解 $p,\ q$ から求まり，これにより２乗誤差が最小となる直線 $y=px+q$ が求まる．

（上記の式 (*) は極小となるための必要条件であるが，最小２乗法の計算においては十分条件も満たすことが分かっている．）

◇表計算ソフトでの数値計算◇

上では，最小２乗法や回帰直線の数学的な側面について述べたが，表計算ソフトや統計処理用のソフトでは，測定値から回帰直線を求める手続きが自動化されている．

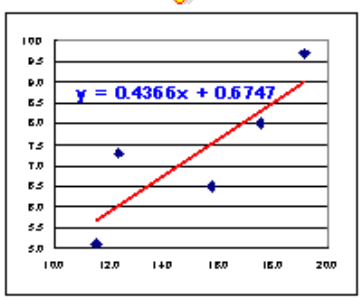
例　Excelにおいて，図３のような測定結果から「散布図」「回帰直線」「回帰直線の方程式」を求めるには，

(1)　A1:B6の範囲を選択し，グラフメニューから散布図を描く．
(2)　次に，描かれた点の１つを右クリックし，近似曲線の追加→線形近似を選ぶと回帰直線が描かれる．
(3)　さらに，描かれた直線を右クリックし，近似曲線の書式設定→オプション→グラフに数式を表示するを選ぶ．

※回帰直線の方程式や回帰直線を用いた予測値だけが必要なときは，
1) 関数 SLOPE(既知のy：図ではB2:B6, 既知のx：図ではA2:A6)　により傾き 0.4366 が得られる．
2) 関数 INTERCEPT(既知のy：図ではB2:B6, 既知のx：図ではA2:A6) により切片 0.6747 が得られる．
3) 新しいxの値（例えばx=18.5）に対するyの予測値（回帰直線上のyの値）は，TREND(既知のｙ，既知のｘ，新しいｘ，１) で得られ 8.752 となる．

図３

	A	B
1	x	y
2	11.5	5.1
3	12.3	7.3
4	15.7	6.5
5	17.5	8.0
6	19.1	9.7



問

次の測定値から，「散布図」「回帰直線」「回帰直線の方程式」を求めよ．[Check]

x	y
2.3	10.2
3.5	9.7
4.6	16.2
6.8	14.6
8.1	17.9

◇連続関数における最小２乗法◇

連続関数についても同様に考えることができ，区間 $a \leqq x \leqq b$ において

$$E(p,q)=\int_a^b \{f(x)-px-q\}^2 dx$$

が最小となる 係数 p,q を求めることができる.

これは，図４において黄色で示した部分について，誤差の２乗の定積分が最小となる直線を求めることに対応している.

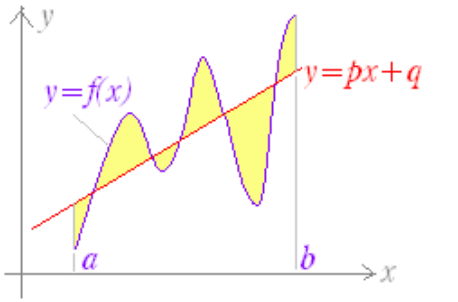
上で解説した離散的なデータに準じて求めると，次のような計算になる.

$$E(p,q)=\int_a^b \{f(x)-px-q\}^2 dx$$
$$=\int_a^b f(x)^2 dx+p^2\int_a^b x^2 dx+q^2\int_a^b dx$$
$$-2p\int_a^b xf(x)dx+2pq\int_a^b xdx$$
$$-2q\int_a^b f(x)dx$$

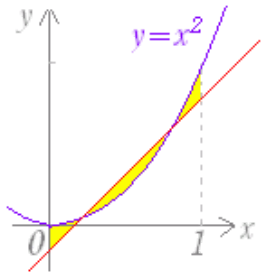
を p, q で偏微分することにより，

$$p\int_a^b x^2 dx+q\int_a^b xdx=\int_a^b xf(x)dx \cdots(1)$$
$$p\int_a^b xdx+q(b-a)=\int_a^b f(x)dx \cdots(2)$$

図４



例　区間 $0 \leqq x \leqq 1$ における関数 $y=x^2$ の回帰直線を求めると，



左の(1)より　$\frac{1}{3}p+\frac{1}{2}q=\frac{1}{4}$　…(1)

左の(2)より　$\frac{1}{2}p+q=\frac{1}{3}$　…(2)

これを解いて，$p=1, q=-\frac{1}{6}$ が得られ，回帰直線の方程式は

$$y=x-\frac{1}{6}$$

となる.

問

区間 $-1 \leqq x \leqq 1$ における関数 $y=x^3$ の回帰直線の方程

式を求めよ. [Check]