Curs 8:
Aproximarea în medie pătratică
Regresia liniară. Regresia polinomială

Octavia-Maria BOLOJAN

octavia.nica@math.ubbcluj.ro

21 Noiembrie 2016

Aproximarea prin metoda celor mai mici pătrate

- Interpolarea funcțiilor nu este potrivită pentru prelucrarea datelor experimentale, deoarece acestea sunt afectate de erori de măsurare
- În acest caz, este preferabil ca graficul funcției să treacă prin puncte determinate experimental
- Cea mai des utilizată metodă pentru determinarea unei funcții care aproximează dependența reprezentată printr-un set de date este **metoda celor mai mici pătrate**
- Aceasta urmărește minimizarea erorilor

$\longrightarrow$determinarea aproximării în sensul celor mai mici pătrate se reduce la rezolvarea unui sistem de ecuații algebrice liniare, cu un număr de ecuații mai mare decât numărul de necunoscute

Matematic, problema se poate formula astfel:

Într-un interval $[a, b]$ sunt specificate n puncte $x_1, x_2, ..., x_n$ și valorile corespunzătoare ale unei funcții $f(x)$

$$f(x_i) = y_i, \quad i = \overline{1, n}.$$

În cazul în care valorile tabelate $y_i = f(x_i)$ sunt afectate de erori, este firesc să se impună minimizarea dintre functiile f și F (funcție model), adică:

$$d(f, F) = \text{minim}$$

Trebuie să determinăm parametrii a_i ai funcției model $F(x; a_i)$ astfel încât aceasta să aproximeze cel mai bine funcția $f(x)$, adică să minimizeze funcționala

$$S = \sum_{i=1}^{n} \left[y_i - F(x_i; a_j) \right]^2 .$$

Minimul acestei funcționale în raport cu parametrii a_j este caracterizat prin relațiile

$$\frac{\partial S}{\partial a_j} = 0, j = 1, 2, ..., m$$

din care pot fi determinați parametrii a_j.

- Această metodă de aproximare poartă numele de **regresie**, **ajustare de model** sau **aproximare în medie prin metoda celor mai mici pătrate** (în literatura de specialitate, termenul corespunzător folosit în limba engleză este *fitting*)

- Regresia în sensul celor mai mici pătrate implică geometric determinarea parametrilor a_j pentru care curba $y = F(x_i; a_j)$ aproximează cel mai bine dependenta $y = f(x)$

Regresia liniară

$\longrightarrow$ este cel mai simplu exemplu de modelare prin metoda celor mai mici pătrate

Aproximarea unor date printr-o dependență liniară care minimizează suma pătratelor dintre punctele date și dreapta de aproximare se numește **regresie liniară**

- Erorile de aproximare a valorilor experimentale, prin folosirea estimatorului liniar, sunt date de:

$$e_i = |y_i - (ax_i + b)|,$$

unde $y = ax_i + b, i = 0, 1, 2, ..., n.$

- Suma pătratelor acestor erori este o funcție de a și b sub forma

$$S(a, b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

- Minimizarea erorilor de aproximare implică îndeplinirea condițiilor

$$\begin{aligned} \frac{\partial S}{\partial a} &= -2\sum_{i=1}^{n}(y_i - ax_i - b) = 0 \\ \frac{\partial S}{\partial a} &= -2\sum_{i=1}^{n}x_i(y_i - ax_i - b) = 0 \end{aligned}$$

- Putem determina coeficienții a și b introducând simbolul sumă în paranteze; obținem astfel relațiile

$$\begin{aligned} nb + a\sum_{i=1}^{n}x_i &= \sum_{i=1}^{n}y_i \\ b\sum_{i=1}^{n}x_i + a\sum_{i=1}^{n}x_i^2 &= \sum_{i=1}^{n}x_iy_i \end{aligned}$$

Scris sub formă matricealăă, avem

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}.$$

Notând

$$\begin{aligned} h_1 &= \sum_{i=1}^{n} x_i, \quad h_2 = \sum_{i=1}^{n} x_i^2, \\ g_1 &= \sum_{i=1}^{n} y_i, \quad g_2 = \sum_{i=1}^{n} x_i y_i, \end{aligned}$$

avem

$$\begin{aligned} nb + ah_1 &= g_1 \\ bh_1 + ah_2 &= g_2, \end{aligned}$$

unde

$$a = \frac{h_1 g_1 - n g_2}{h_1^2 - n h_2}, \quad b = \frac{h_1 g_2 - h_2 g_1}{h_1^2 - n h_2}$$

Exemplu

Avem baza de date obținută din experimentele de laborator:

x_i	1	3	5	7	9	11
y_i	6.12	7.36	8.01	9.18	10.15	11.05

Să se obțină estimatorul liniar (linia de regresie), în sensul celor mai mici pătrate.

Soluție.
Necunoscutele a și b din ecuația $y = ax + b$ se obțin prin rezolvarea sistemului

$$\begin{bmatrix} 6 & \sum_{i=1}^{6} x_i \\ \sum_{i=1}^{6} x_i & \sum_{i=1}^{6} x_i^2 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{6} y_i \\ \sum_{i=1}^{6} x_i y_i \end{bmatrix},$$

unde:

$$\sum_{i=1}^{6} x_i = 1 + 3 + 5 + 7 + 9 + 11 = 36$$

$$\sum_{i=1}^{6} x_i^2 = 1 + 9 + 25 + 49 + 81 + 121 = 286$$

$$\sum_{i=1}^{6} y_i = 6.12 + 7.36 + 8.01 + 9.18 + 10.15 + 11.05 = 51.87$$

$$\sum_{i=1}^{6} x_i y_i = 1 \cdot 6.12 + 3 \cdot 7.36 + 5 \cdot 8.01 + 7 \cdot 9.18 + 9 \cdot 10.15 + 11 \cdot 11.05 = 345.41$$

Sistemul de ecuații devine

$$\begin{bmatrix} 6 & 36 \\ 36 & 286 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 51.87 \\ 345.41 \end{bmatrix}.$$

Rezolvând sistemul, obținem coeficienții

$$a = 0.4961, \quad b = 5.6531$$

și astfel, dreapta de regresie liniară (estimatorul liniar) este:

$$y = 0.4961x + 5.6531.$$

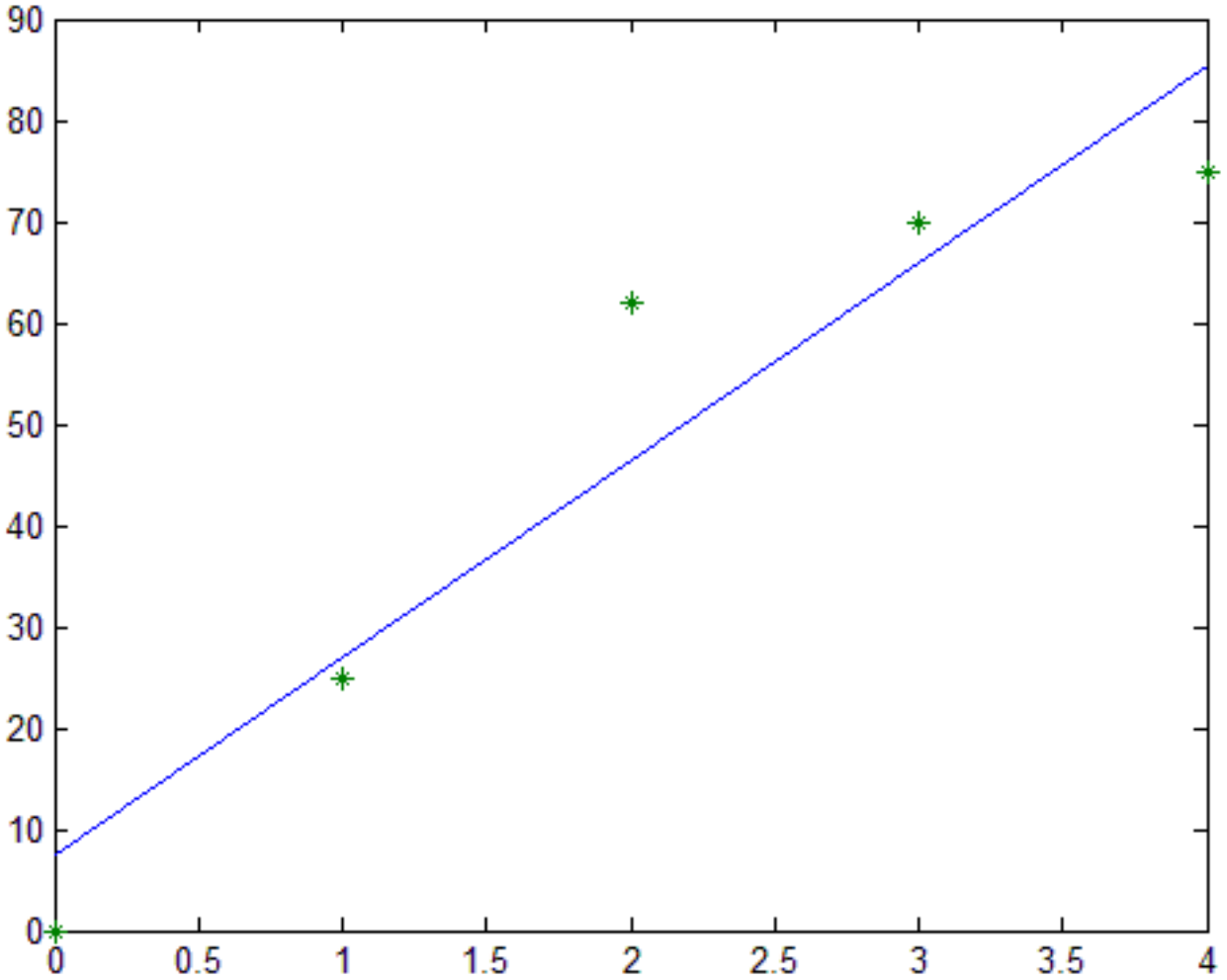
Comenzi Matlab pentru regresia liniară

În Matlab, determinarea parametrilor a și b ai dreptei de aproximare $y = ax + b$ se face utilizând functia `polyfit()`, care se apelează cu sintaxa `p=polyfit(x,y,n)`, iar `polyval(p.x)` evaluează un polinom în valorile precizate ale variabilei.

Exemplu. Avem setul de date $x = [0, 1, 2, 3, 4]$, $y = [0, 25, 62, 70, 75]$. Să se aproximeze în sensul celor mai mici pătrate cu o regresie liniară aceste date.

În Matlab avem secvențele de cod:

```
x=[0,1,2,3,4];
y=[0,25,62,70,75];
c=polyfit(x,y,1);
e=polyval(c,x);
axis([-2,10,-50,150])
plot(x,e,x,y,'*')
```


90
80
70
60
50
40
30
20
10
0
0
0.5
1
1.5
2
2.5
3
3.5
4

Regresia polinomială

$\longrightarrow$reprezintă o aproximare a unui set de date printr-un polinom de forma

$$F\left(x; a_j\right) = a_0 + a_1 x + a_2 x^2 + ... + a_m x^m, \quad j = 0, 1, ..., m.$$

Considerăm cunoscute perechile $(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n)$.

- Ecuația generală a estimatorului pătratic este

$$y = a_0 + a_1 x + a_2 x^2.$$

- Pentru determinarea estimatorului pătratic pentru baza de date experimentale obtinem coeficienții a_0, a_1 si a_2.
- Erorile de aproximare a valorilor experimentale, prin folosirea estimatorului pătratic sunt date de

$$e_i = |y_i - y| = \left| y_i - \left(a_0 + a_1 x_i + a_2 x_i^2 \right) \right|, \quad i = 0, 1, 2, ..., n.$$

Suma pătratelor acestor erori este o funcție depinzând de a_0, a_1 și a_2 sub forma

$$S(a_0, a_1, a_2) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 \right)^2 .$$

Minimizarea erorilor de aproximare implică îndeplinirea condițiilor:

$$\begin{aligned}
\frac{\partial S}{\partial a_0} &= -2 \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 \right) = 0 \\
\frac{\partial S}{\partial a_1} &= -2 \sum_{i=1}^{n} x_i \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 \right) = 0 \\
\frac{\partial S}{\partial a_2} &= -2 \sum_{i=1}^{n} x_i^2 \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 \right) = 0
\end{aligned}$$

Putem determina coeficienții a_0, a_1, a_2 din ecuațiile care rezultă, introducând simbolul sumă în paranteze:

$$\begin{aligned} na_0 + a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3 &= \sum_{i=1}^{n} x_i y_i \\ a_0 \sum_{i=1}^{n} x_i^2 + a_1 \sum_{i=1}^{n} x_i^3 + a_2 \sum_{i=1}^{n} x_i^4 &= \sum_{i=1}^{n} x_i^2 y_i . \end{aligned}$$

Scris sub formă matriceală, avem:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i^2 y_i \end{bmatrix} .$$

Exemplu

Să se obțină estimatorul pătratic, în sensul celor mai mici pătrate, pentru baza de date experimentale:

x_i	-2	-1	0	1	2
y_i	2.17	13.05	35.44	56.33	92.29

.

Solutie.

Necunoscutele a, b, c din ecuația $y = a + bx + cx^2$ se obțin prin rezolvarea sistemului

$$\begin{bmatrix} 5 & \sum_{i=1}^{5} x_i & \sum_{i=1}^{5} x_i^2 \\ \sum_{i=1}^{5} x_i & \sum_{i=1}^{5} x_i^2 & \sum_{i=1}^{5} x_i^3 \\ \sum_{i=1}^{5} x_i^2 & \sum_{i=1}^{5} x_i^3 & \sum_{i=1}^{5} x_i^4 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{5} y_i \\ \sum_{i=1}^{5} x_i y_i \\ \sum_{i=1}^{5} x_i^2 y_i \end{bmatrix},$$

unde avem:

$$\sum_{i=1}^{5} x_i = (-2) + (-1) + 0 + 1 + 2 = 0$$

$$\sum_{i=1}^{5} x_i^2 = 4+1+0+1+4 = 10$$

$$\sum_{i=1}^{5} x_i^3 = (-8)+(-1)+0+1+8 = 0$$

$$\sum_{i=1}^{5} x_i^4 = 16+1+0+1+16 = 34$$

$$\sum_{i=1}^{5} y_i = 2.17+13.05+35.44+56.33+92.29 = 199.28$$

$$\sum_{i=1}^{5} x_i y_i = (-2)\cdot 2.17+(-1)\cdot 13.05+0\cdot 35.44+1\cdot 56.33+2\cdot 92.29 = 223.52$$

$$\sum_{i=1}^{5} x_i^2 y_i = 4\cdot 2.17+1\cdot 13.05+0\cdot 35.44+1\cdot 56.33+4\cdot 92.29 = 447.22$$

Rezultă sistemul de ecuații:

$$\begin{bmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 199.28 \\ 223.52 \\ 447.22 \end{bmatrix}.$$

Rezolvând acest sistem de ecuații, obținem:

$$a = 32.9046, \quad b = 22.3520, \quad c = 3.4757.$$

Astfel, dreapta de regresie polinomială (estimatorul pătratic) este de forma:

$$y = 32.9046 + 22.3520x + 3.4757x^2.$$

Comenzi Matlab pentru regresia polinomială

→determinarea parametrilor a, b și c ai dreptei de aproximare $y = a + bx + cx^2$ se face tot cu ajutorul funcțiilor Matlab `polyfit()`, `polyval()`
→`polyfit()` - determină coeficienții de regresie cuadratică (regresie polinomială, estimator pătratic)

Exemplu. Să se construiască un polinom de grad 2 care să aproximeze următoarele puncte obținute prin măsurători experimentale:

$$\begin{aligned} x &= [0.1, 0.4, 0.5, 0.7, 0.7, 0.9] \\ y &= [0.61, 0.92, 0.99, 1.52, 1.47, 2.03]. \end{aligned}$$

Să se reprezinte grafic.
Soluție.

```
x=[0.1,0.4,0.5,0.7,0.7,0.9];y=[0.61,0.92,0.99,1.52,1.47,2.03];
cc=polyfit(x,y,2)
xx=x(1):0.1:x(length(x));
yy=polyval(cc,xx)
plot(xx,yy,'r-');hold on
plot(x,y,'x')
axis([0,1,0,3])
xlabel('x');ylabel('y')
```

Obținem:

```
cc =
 1.7295 0.0591 0.5871
```