Curs #14 Curbe de regresie

1. Regresie liniară

Fie dat următorul tabel de date:

x_i	y_i
1	1,3
2	3,5
3	4,2
4	5
5	7
6	8,8
7	10,1
8	12,5
9	13,0
10	15,6

Reprezentarea grafică a punctelor (x_i, y_i), $i=\overline{1,10}$ se poate vedea în Fig. 1.

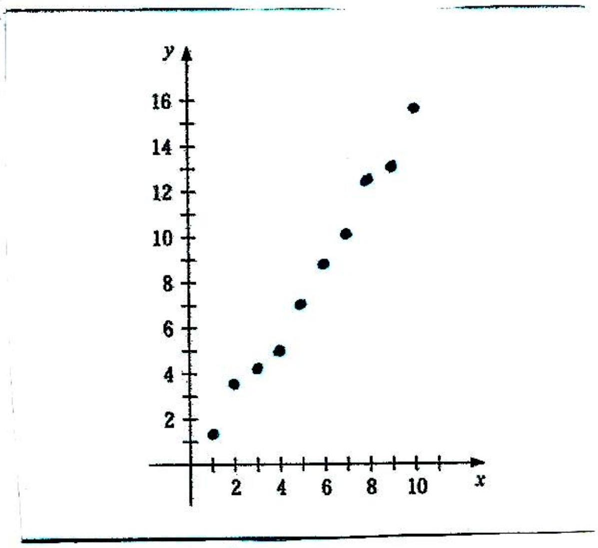

Fig. 1

Conform graficului se observă că relația dintre x și y este una liniară. Motivul pentru care aceste puncte nu se află pe dreaptă este acela că datele experimentale au fost obținute cu o marjă de eroare.

În acest caz nu este indicat să se construiască o funcție care să interpoleze datele, ci mai degrabă să le aproximeze.

Prin interpolare, cum ar fi interpolarea Lagrange se introduc oscilații care nu sunt naturale. De exemplu, reprezentarea grafică a polinomului $P_9(x)$ care interpolează setul date de mai sus este dată în Fig. 2.

Vom construi în continuare o dreaptă de ecuație $y = ax + b$ care trece cel mai aproape de punctele date. Coeficienții a și b sunt necunoscuți și urmează a fi calculați.

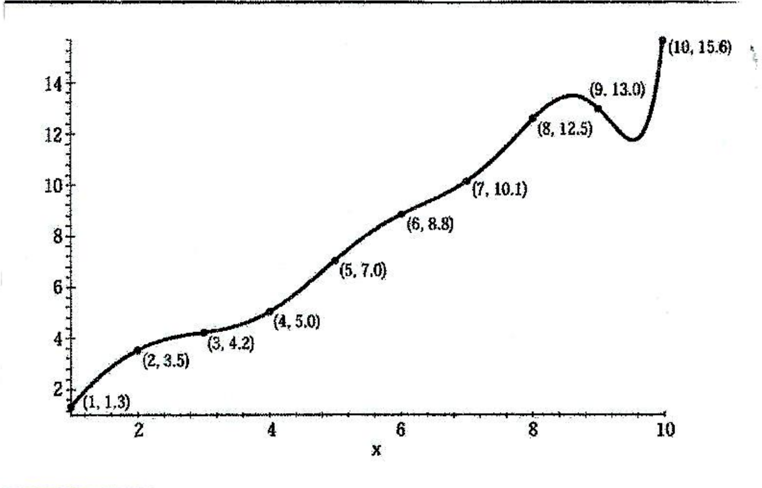

Fig.2

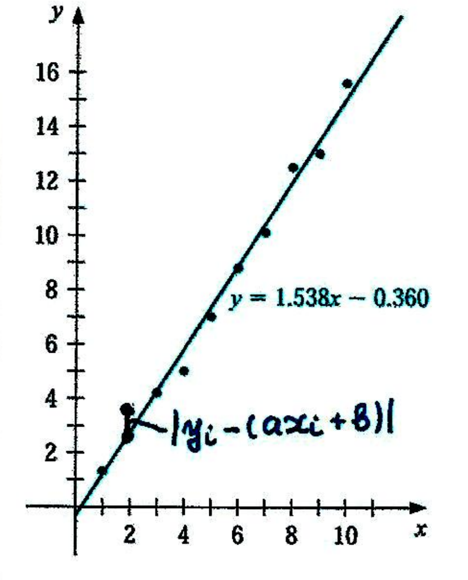

Fig 3

Distanța pe verticală dintre punctul (x_i, y_i) și cel situat pe dreaptă, (x_i, ax_i+b) este $|y_i - (ax_i+b)|$.

Fie vectorii $y = (y_1, y_2, \ldots, y_n)$ și $\bar{y} = (ax_1+b, ax_2+b, \ldots, ax_n+b)$.

Coeficienții a și b îi vom calcula minimizând funcția:

$$E(a,b) = \|y - \bar{y}\|^2 = \sum_{i=1}^{n} \left(y_i - (ax_i+b)\right)^2$$

Precizăm că $\|a\| = \sqrt{a_1^2 + a_2^2 + \ldots + a_n^2}$, unde $a = (a_1, a_2, \ldots, a_n)$

Minimizând funcția E se obține poziția optimă a dreptei $y = ax+b$ față de punctele (x_i, y_i), $i = \overline{1,n}$.

Funcția E admite un punct de minim care este și punct staționar. Rezolvăm sistemul:

$$\begin{cases} \dfrac{\partial E}{\partial a}(a,b) = 0 \\[2mm] \dfrac{\partial E}{\partial b}(a,b) = 0 \end{cases} \iff \begin{cases} 2\sum\limits_{i=1}^{n}\left(y_i - (ax_i+b)\right)\cdot\left(y_i - (ax_i+b)\right)'_a = 0 \\[2mm] 2\sum\limits_{i=1}^{n}\left(y_i - (ax_i+b)\right)\cdot\left(y_i - (ax_i+b)\right)'_b = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n}(y_i-(ax_i+b))\cdot(-x_i)=0 \\ \sum_{i=1}^{n}(y_i-(ax_i+b))\cdot(-1)=0 \end{cases}$$

$$\Leftrightarrow \begin{cases} -\sum_{i=1}^{n} x_i y_i + a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = 0 \\ -\sum_{i=1}^{n} y_i + a\sum_{i=1}^{n} x_i + b\underbrace{\sum_{i=1}^{n} 1}_{\substack{\parallel \\ n}} = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} a\cdot\sum_{i=1}^{n} x_i^2 + b\cdot\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i\cdot y_i \\ a\cdot\sum_{i=1}^{n} x_i + b\cdot n = \sum_{i=1}^{n} y_i \end{cases}$$

Sistemul obținut este un sistem liniar. Matricea asociată sistemului este:

$$A = \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \underbrace{\sum_{i=1}^{n} x_i^0}_{\substack{\parallel \\ n}} \end{pmatrix}$$

iar vectorul termenilor liberi

$$w = \begin{pmatrix} \sum_{i=1}^{n} x_i\cdot y_i \\ \sum_{i=1}^{n} y_i \end{pmatrix}$$

Coeficienții a și b se determină în urma rezolvării sistemului $A\cdot\begin{pmatrix} a \\ b \end{pmatrix} = w$.

Obs: Pentru rezolvarea numerică a sistemului anterior se poate aplica una din metodele de eliminare Gauss.

Pentru exemplul de mai sus funcția liniară $y=ax+b$ este

$$y = 1,538x - 0,360 \quad (\text{vezi Fig. } 3)$$

2. Regresie polinomială

Dacă punctele (x_i, y_i), $i=\overline{1,n}$ se află pe o curbă polinomială atunci funcția care aproximează datele o putem alege de forma unui polinom de grad m

$$P_m(x) = a_m x^m + a_{m-1}x^{m-1} + \dots + a_1 x + a_0$$

Considerăm în continuare cazul $m=2$, caz care corespunde regresiei pătratice.

$$P_2(x) = ax^2 + bx + c$$

Minimizând funcția $E(a,b,c) = \| y - \bar{y} \|^2 = \sum_{i=1}^{n} \left(y_i - (ax_i^2 + bx_i + c)\right)^2$ obținem poziția optimă a parabolei față de punctele (x_i, y_i). Aflăm punctul staționar pentru funcția E

$$\begin{cases} \dfrac{\partial E}{\partial a}(a,b,c) = 0 \\[2mm] \dfrac{\partial E}{\partial b}(a,b,c) = 0 \\[2mm] \dfrac{\partial E}{\partial c}(a,b,c) = 0 \end{cases} \Leftrightarrow \begin{cases} 2\sum_{i=1}^{n} \left(y_i - (ax_i^2 + bx_i + c)\right)\cdot(-x_i^2) = 0 \\ 2\sum_{i=1}^{n} \left(y_i - (ax_i + bx_i + c)\right)\cdot(-x_i) = 0 \\ 2\sum_{i=1}^{n} \left(y_i - (ax_i + bx_i + c)\right)\cdot(-1) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} a\cdot\sum_{i=1}^{n} x_i^4 + b\cdot\sum_{i=1}^{n} x_i^3 + c\cdot\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i\cdot x_i^2 \\ a\cdot\sum_{i=1}^{n} x_i^3 + b\cdot\sum_{i=1}^{n} x_i^2 + c\cdot\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i\cdot x_i \\ a\cdot\sum_{i=1}^{n} x_i^2 + b\cdot\sum_{i=1}^{n} x_i + c\cdot n = \sum_{i=1}^{n} y_i \end{cases}$$

În acest caz matricea asociată sistemului este:

$$A = \begin{pmatrix} \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{4} x_i^3 & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^1 & \sum_{i=1}^{n} x_i^0 \end{pmatrix}$$

iar vectorul termenilor liberi

$$W = \begin{pmatrix} \sum_{i=1}^{n} y_i \cdot x_i^2 \\ \sum_{i=1}^{n} y_i x_i^1 \\ \sum_{i=1}^{n} y_i x_i^0 \end{pmatrix}$$

3. Regresia exponențială

Dacă datele experimentale descriu o funcție exponențială, putem căuta curba $y = b e^{ax}$ urmând aceeași strategie.

$$E(a,b) = \sum_{i=1}^{n} (y_i - b \cdot e^{ax_i})^2$$

$$\begin{cases} \frac{\partial E}{\partial a}(a,b) = 0 \\ \frac{\partial E}{\partial b}(a,b) = 0 \end{cases} \Leftrightarrow \begin{cases} 2 \sum_{i=1}^{n} (y_i - b \cdot e^{ax_i}) \cdot (y_i - b e^{ax_i})'_a = 0 \\ 2 \sum_{i=1}^{n} (y_i - b \cdot e^{ax_i}) \cdot (y_i - b \cdot e^{ax_i})'_b = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n} (y_i - b \cdot e^{ax_i}) \cdot (-b e^{ax_i} \cdot x_i) = 0 \\ \sum (y_i - b e^{ax_i}) \cdot (-e^{ax_i}) = 0 \end{cases}$$

Observăm că sistemul obținut nu poate fi rezolvat analitic.

Pentru a determina totuşi coeficienţii a şi b vom logaritma expresia $y = b \cdot e^{ax}$: $\ln y = \ln(b \cdot e^{ax}) \Rightarrow \ln y = \ln b + \ln(e^{ax}) \Rightarrow$

$$\ln y = \ln b + a \cdot x \Rightarrow \ln y = b' + a \cdot x \text{ , cu } b' = \ln b$$

Rescriem funcţia E pentru ecuaţia obţinută, care este echivalentă cu cea de la care am pornit

$$E(a, b') = \sum_{i=1}^{n} (\ln y_i - (b' + a x_i))^2$$

Observăm că E este similară cazului liniar.

$$\begin{cases} \dfrac{\partial E}{\partial a}(a, b') = 0 \\ \dfrac{\partial E}{\partial b'}(a, b') = 0 \end{cases} \Leftrightarrow \begin{cases} 2 \sum\limits_{i=1}^{n} (\ln y_i - (b' + a x_i))(-x_i) = 0 \\ 2 \sum\limits_{i=1}^{n} (\ln y_i - (b' + a x_i)) \cdot (-1) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} a \cdot \sum\limits_{i=1}^{n} x_i^2 + b' \cdot \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} (\ln y_i) \cdot x_i \\ a \cdot \sum\limits_{i=1}^{n} x_i + b' \cdot n = \sum\limits_{i=1}^{n} (\ln y_i) \end{cases}$$

Rezolvând sistemul obţinem valori pentru a şi b'. Cum b' este calculat, din $b' = \ln b$ rezultă $\boxed{b = e^{b'}}$

Obs.: Metoda de determinare a funcţiilor care aproximează un set de date, prezentată în cadrul acestui curs se numeşte <u>metoda celor mai mici pătrate</u>.