REGRESIA

TEHNICI DE DATA MINING

INVATARE SUPERVIZATA

Supervised learning

- CLASIFICAREA – ATRIBUT TINTA DISCRET
- REGRESIE – ATRIBUT TINTA CONTINUU

INVATARE NESUPERVIZATA

Unsupervised learning

- ASOCIERE
- CLUSTERIZARE

REGRESIA

Estimarea valorilor atributului tinta in functie de valorile celorlalte atribute

Atributul tinta: variabila dependenta

Celelalte atribute: variabilele independente

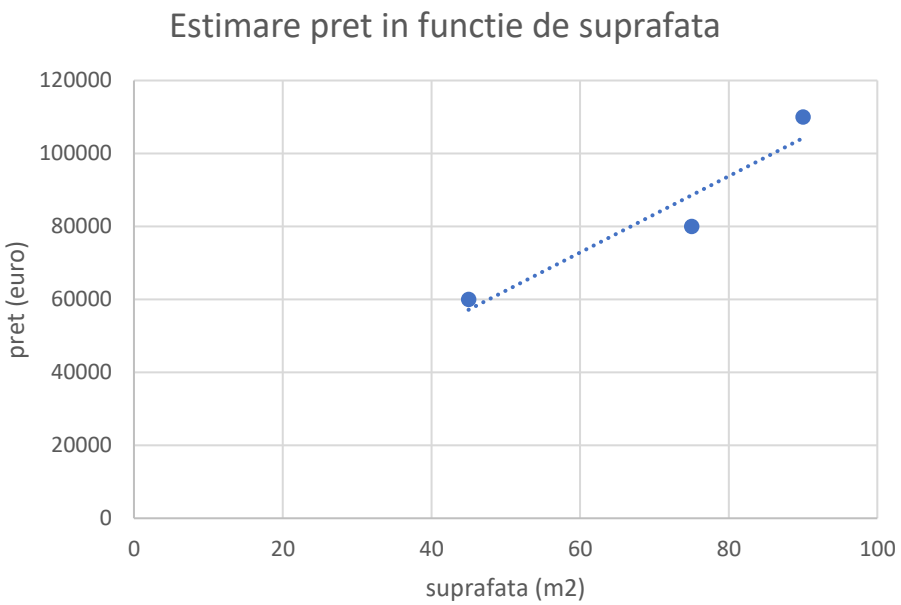
EXEMPLU 1

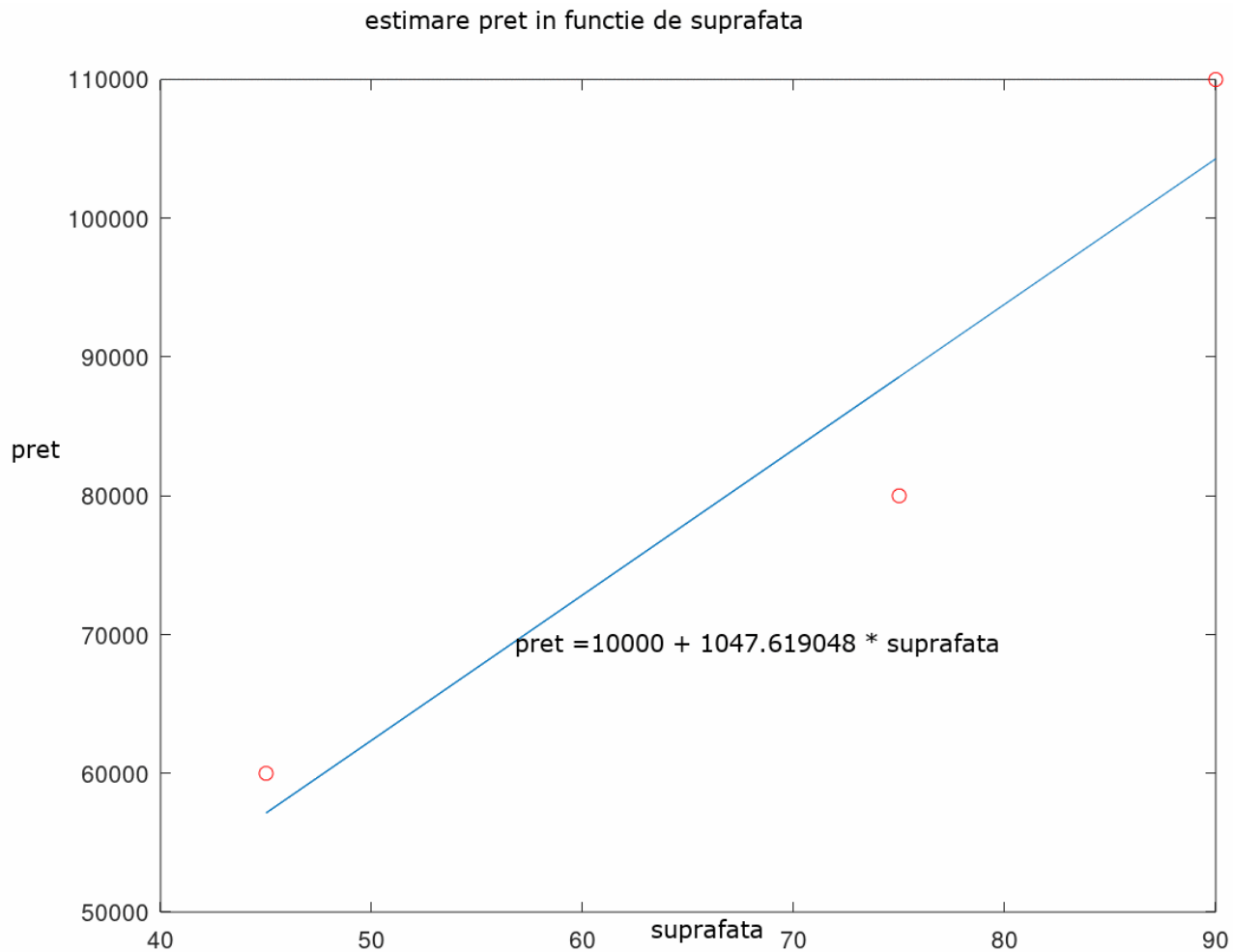
- Analiza relatiei dintre pretul unui apartament si suprafata lui intr-o zona dintr-un oras.
- Estimarea pretului in functie de suprafata

SUPRAFATA (m2)	PRET (Euro)
90	110000
45	60000
75	80000
60	????

- Atribut tinta: pret
- Atribut de intrare: suprafata

SUPRAFATA (m2)	PRET (Euro)
90	110000
45	60000
75	80000




estimare pret in functie de suprafata
pret
suprafata
pret =10000 + 1047.619048 * suprafata
110000
100000
90000
80000
70000
60000
50000
40
50
60
70
80
90

EXEMPLU 2

- Analiza relatiei dintre greutatea unei persoane si inaltimea sa
- Estimarea greutatii in functie de inaltime sau invers

Inaltime (cm)	Greutate (kg)
160	54
165	60
170	85
185	90

- Atribut tinta: greutate
- Atribut de intrare: inaltime

EXEMPLU 3

- Estimarea punctajului la un test al unui student in functie de nr de ore de studiu pentru test

- Atribut tinta: punctaj
- Atribut de intrare: nr ore studiu

Nr ore studiu	Punctaj test
4	54
8	60
5	85
7	90
16	100
6	???

Tipuri de regresie

Regresie simpla

Estimarea atributului tinta in functie de un singur atribut independent se numeste **regresie simpla.**

- Y = f(X)

Regresie multipla

- Daca estimarea se face in functie de mai multe attribute independente atunci avem **regresie multipla.**
- Y=f(X1,X2,…, Xk)

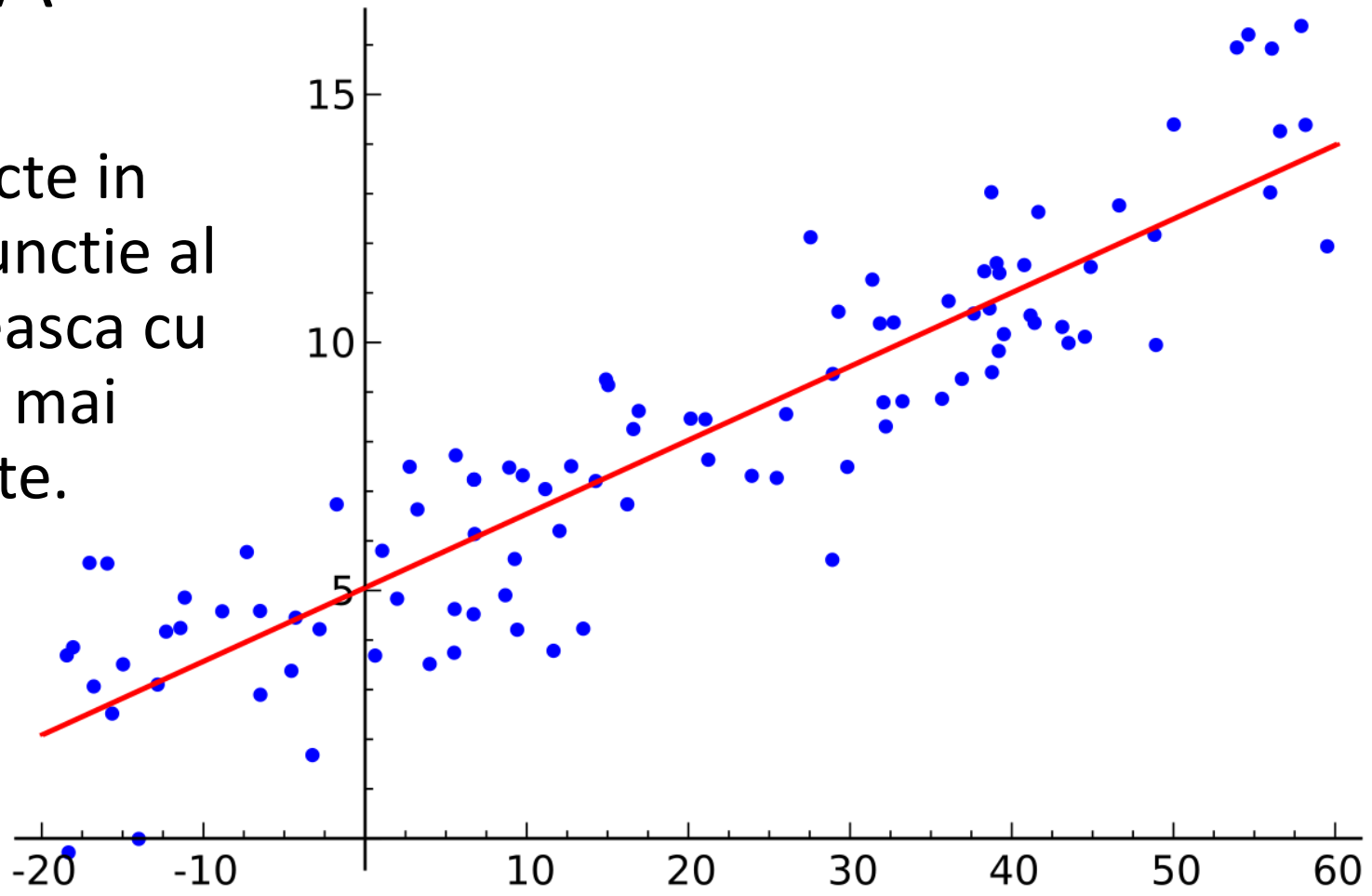
EXEMPLU 4

CYLINDER	DISPLACEMENT	HORSEPOWER	WEIGHT	ACCELERATION	MPG
8	307	130	3504	12	18
8	350	165	3693	11.5	15
8	318	150	3436	11	18
8	304	150	3433	12	16
8	302	140	3449	10.5	17

- Estimarea consumului de combustibil al unei masini in functie de cateva caracteristici ale sale:
- greutate
- marimea motorului
- puterea motorului
- nr. cilindri
- acceleratie
- displacement

REGRESIE SIMPLA

- Se da o multime de puncte in plan. Sa se gaseasca o functie al carui grafic sa se potriveasca cu aceste puncte, sa fie cat mai aproape de aceste puncte.



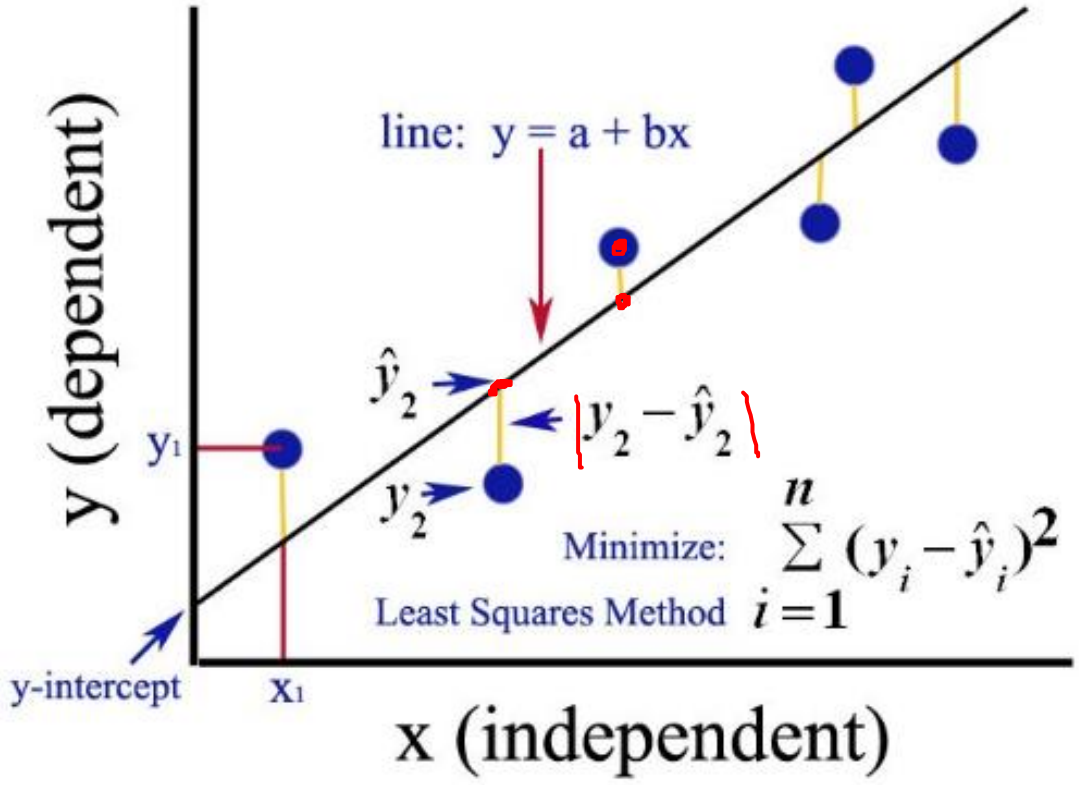
wikipedia

In imagine aceasta functie este un polinom de grad I, ax+b al carui graphic este o dreapta.

Acest tip de regresie se numeste regresie liniara.

Regresia liniara

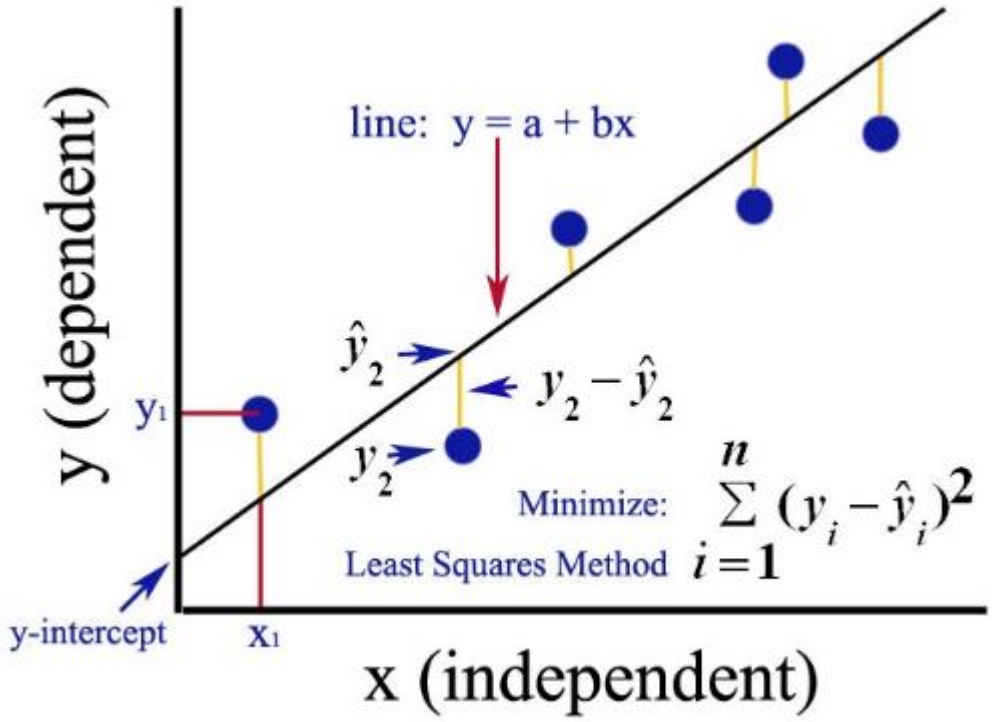
- Multimea de puncte este aproximata printr-o dreapta.
- Functia care va estima relatia dintre variabila de intrare si variabila de iesire (numit atributul tinta, in acest caz) este liniara
- Y=a + b X



- Date punctele
- $(x_1,y_1), (x_2, y_2),\ldots,(x_n,y_n)$
- vrem sa determinam a si b astfel incat dreapta
- Y=a +b x

sa aproximeze bine aceste puncte adica

sa treaca sufficient de aproape de puncte



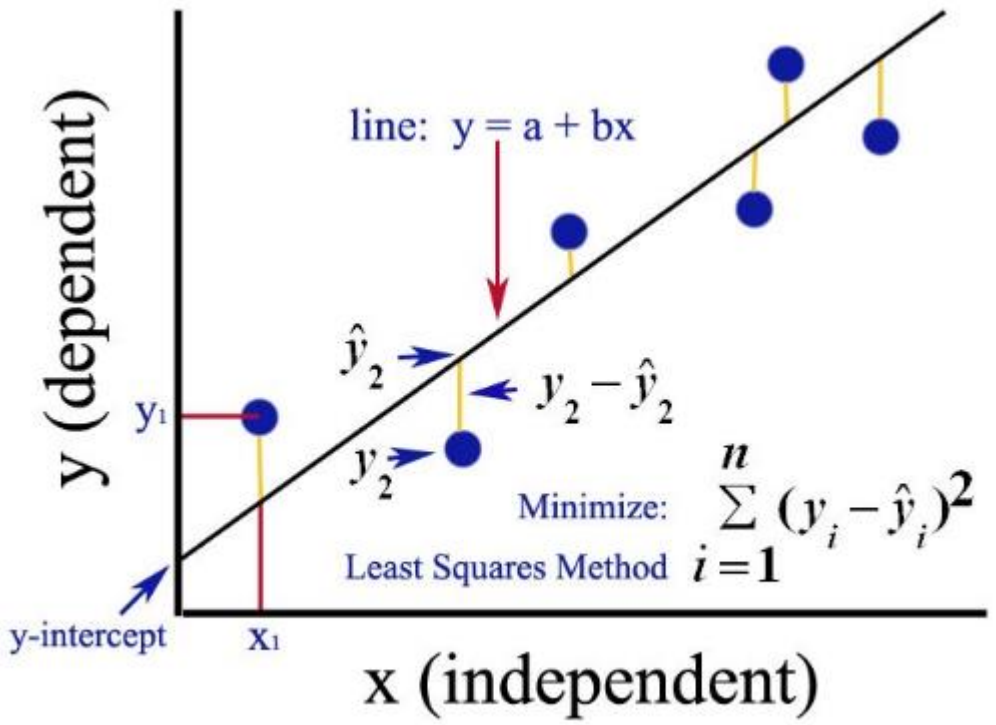
Metoda celor mai mici patrate

- Masuram distantele de la fiecare punct la punctul corespunzator de pe dreapta,
- (x_i, y_i) -> $(x_i, \hat{y}_i)$

distanta = $| y_i - \hat{y}_i |$ pt fiecare i=1,n

S= suma patratelor acestor distante

Cautam a si b astfel incat S sa fie minima.



Metoda celor mai mici patrate

- $S = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 =$
- $= \sum_{i=1}^{n}(y_i - a - b\,x_i)^2$
- S = S(a,b) = functie de doua variabile.
- S este minima cand $\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = 0$

- Se rezolva sistemul

- $a * n + b \sum x_i = \sum y_i$
- $a * \sum x_i + b \sum x_i^2 = \sum x_i y_i$

Exemplu – Regresie liniara

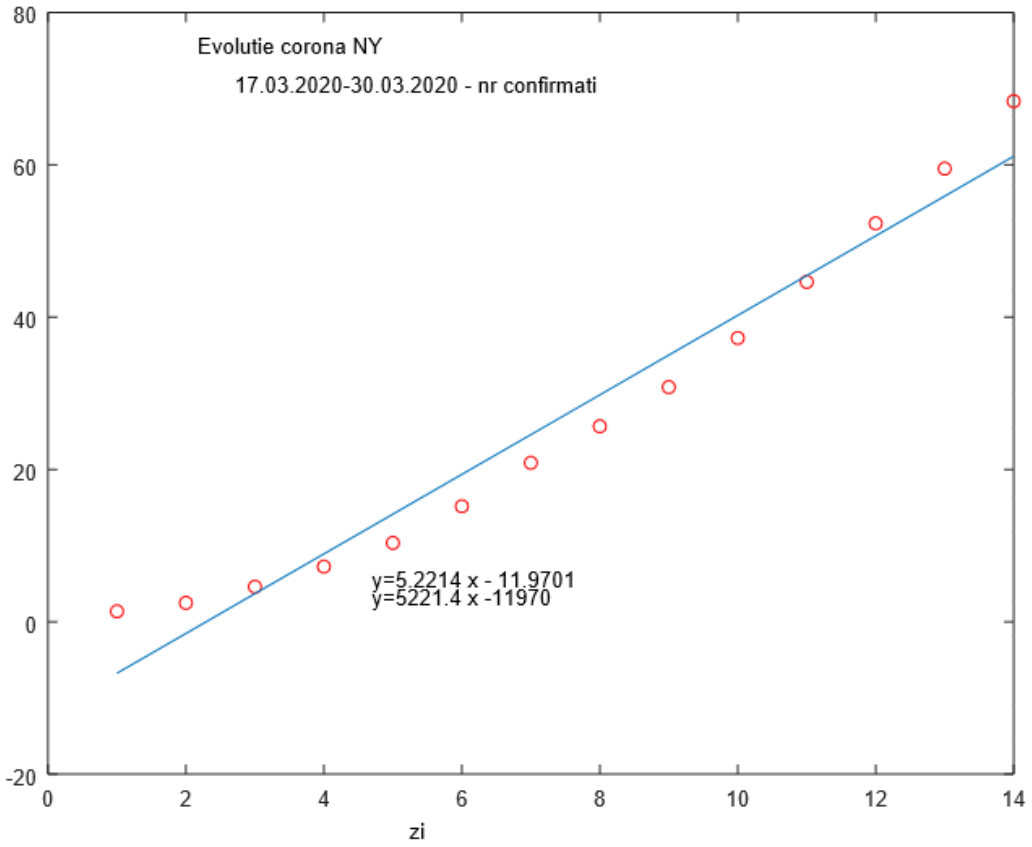
rez in MATLAB

$Ax = b$

$x = A \backslash b$

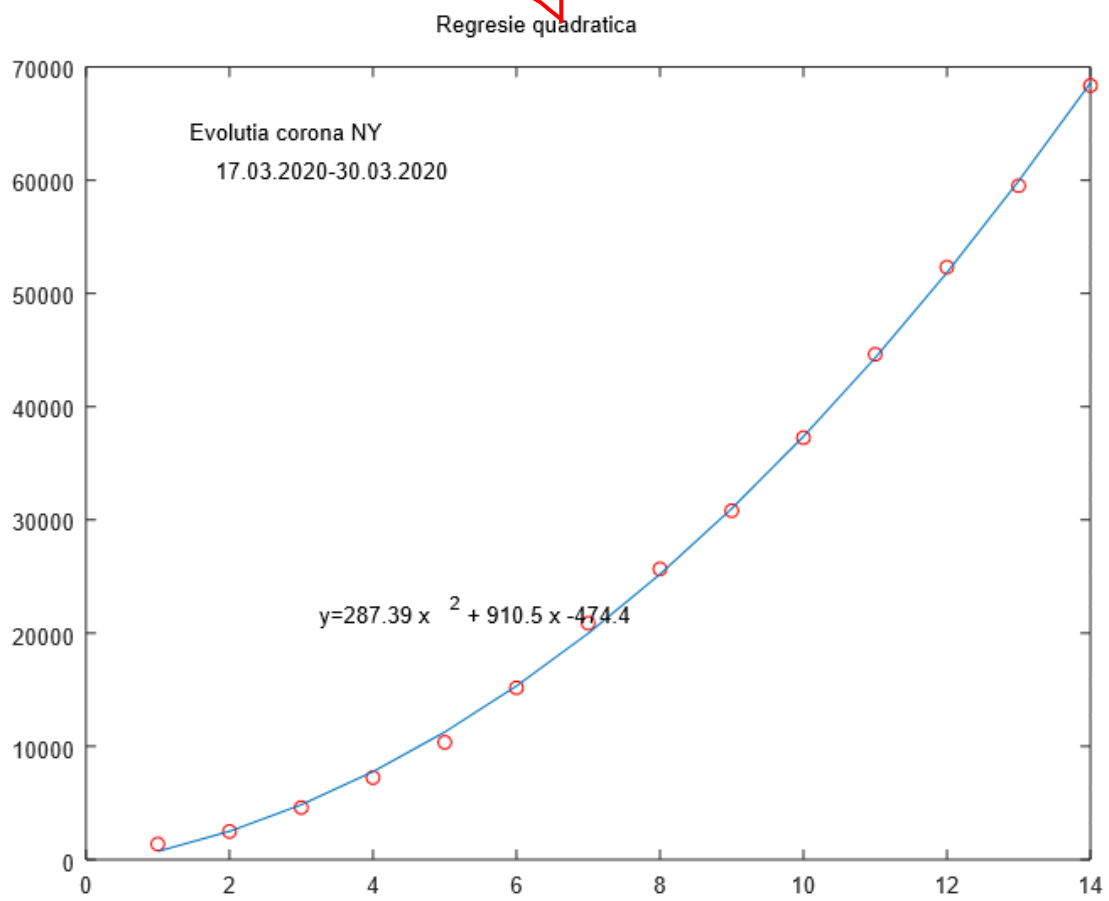
$y = ax + b$

Evolutia nr cazurilor confirmate de COV-ID 19 in statul New York	
Perioada: 17.03.2020 - 30.03.2020	
Data	**Nr cazuri confirmate**
3/17/2020	1374
3/18/2020	2481
3/19/2020	4597
3/20/2020	7245
3/21/2020	10356
3/22/2020	15168
3/23/2020	20875
3/24/2020	25665
3/25/2020	30811
3/26/2020	37258
3/27/2020	44635
3/28/2020	52318
3/29/2020	59513
3/30/2020	68369



Regresie patratica

$y = ax^2 + bx + c$



EVALUAREA MODELELOR DE REGRESIE

Metrics

Mean Absolute Error	10.323824
Root Mean Squared Error	12.903307
Relative Absolute Error	0.87377
Relative Squared Error	0.745059
Coefficient of Determination	0.254941

MAE: Mean Absolute Error

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - \hat{y}_i \right|$$

Root Mean Squared Error

RMSE este radical din MSE

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{MSE}$$

Coefficient of determination = R2-score

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$