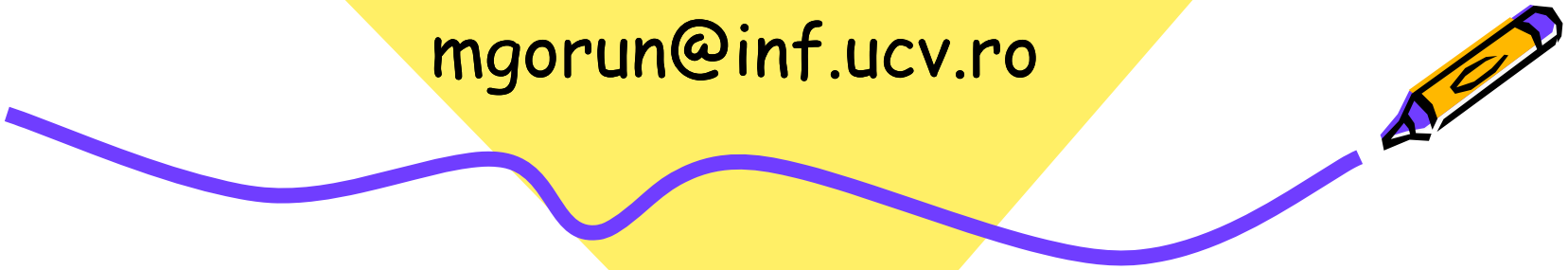
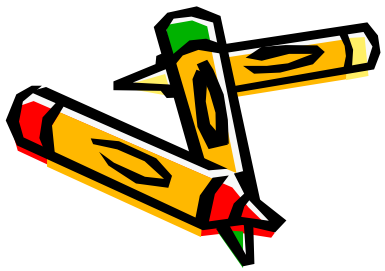
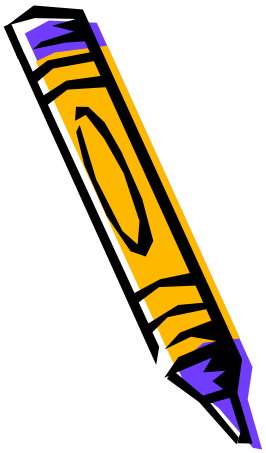
Regresia liniara

Marina Gorunescu
mgorun@inf.ucv.ro

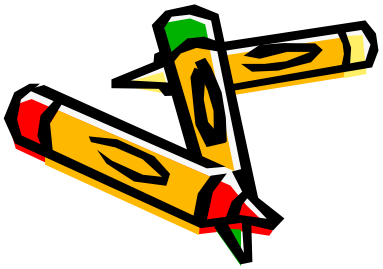


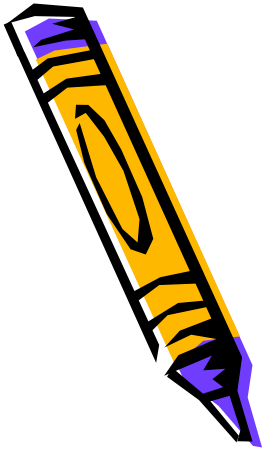
despre prognoza





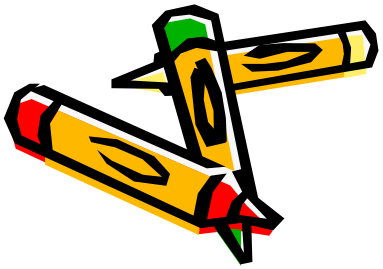
Prognoza reprezintă procesul estimării unei anumite mărimi, pe baza datelor istorice cunoscute, de exemplu, prognoza vremii în următoarele 24 de ore sau mai mult, prognoza prețului petrolului pentru perioada de vară, prognoza cursului leu/euro în următorul semestru etc.

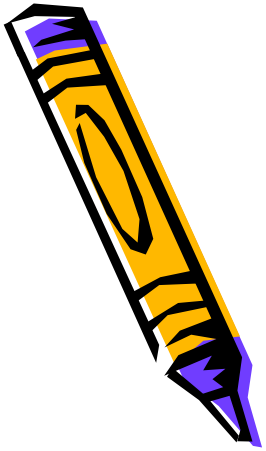




Deciziile care implică incertitudine au absolută nevoie de prognoză. Din vastul domeniu al prognozei ne vom limita la a prezenta două aspecte importante:

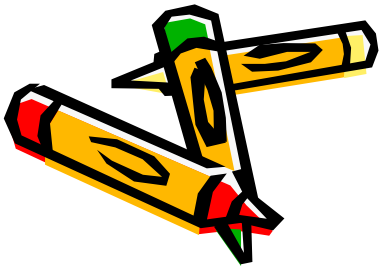
- Metoda de prognoză bazată pe regresia liniară multiplă;
- Prognoza în cazul seriilor temporale.

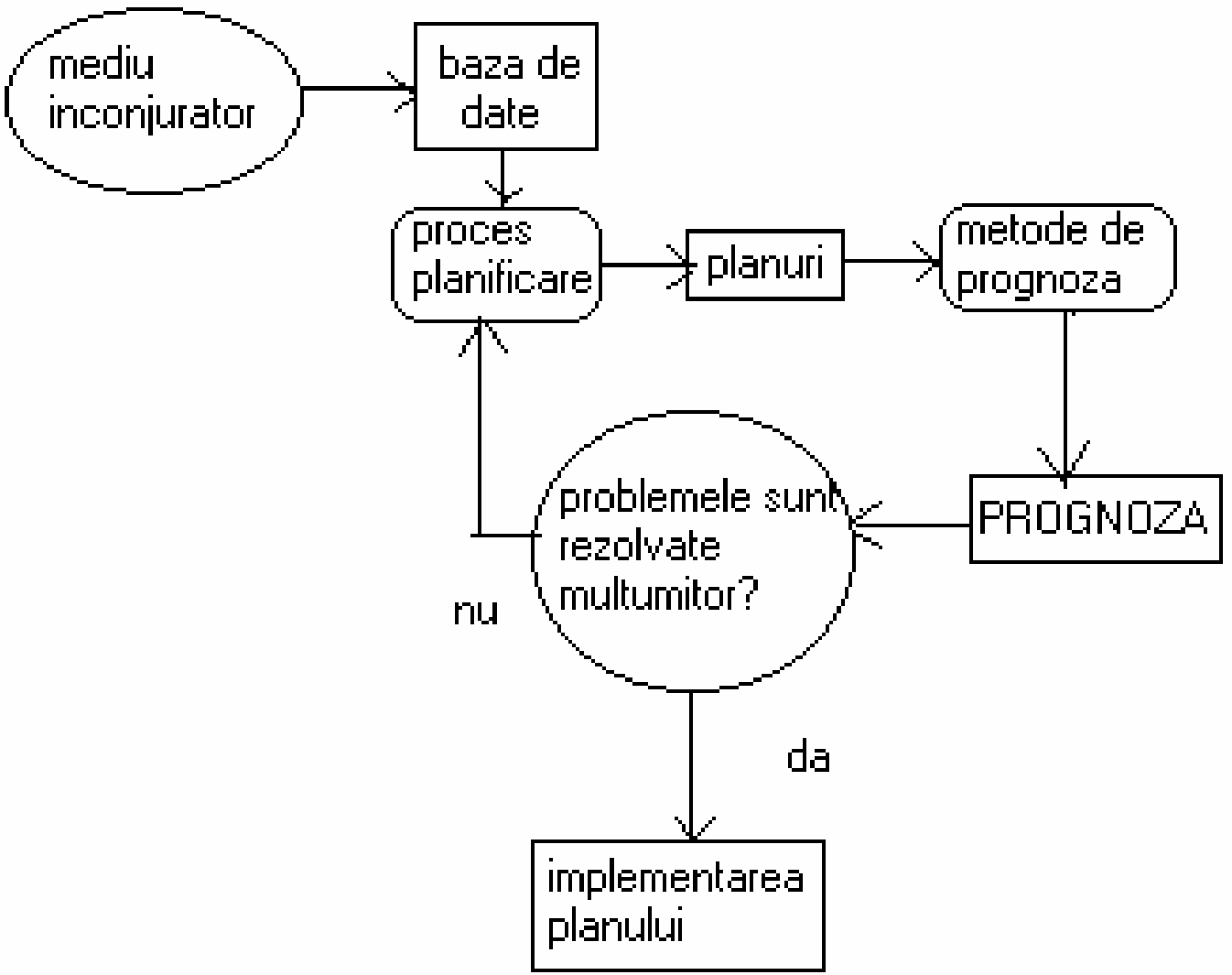


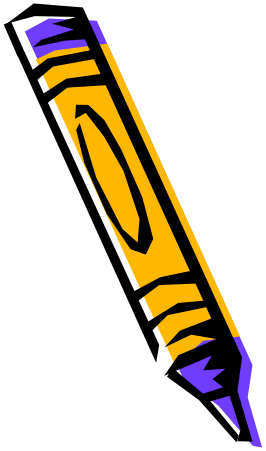


Cei ce se ocupă de planificare folosesc metodele prognozei pentru simularea şi vizualizarea rezultatelor planurilor construite.

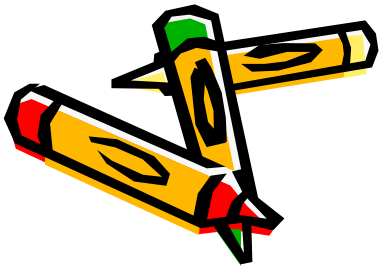
Dacă rezultatul nu este mulțumitor este necesară revizuirea planurilor şi modelelor, din nou se estimează rezultatele şi procesul continuă până la obținerea unui rezultat satisfăcător.



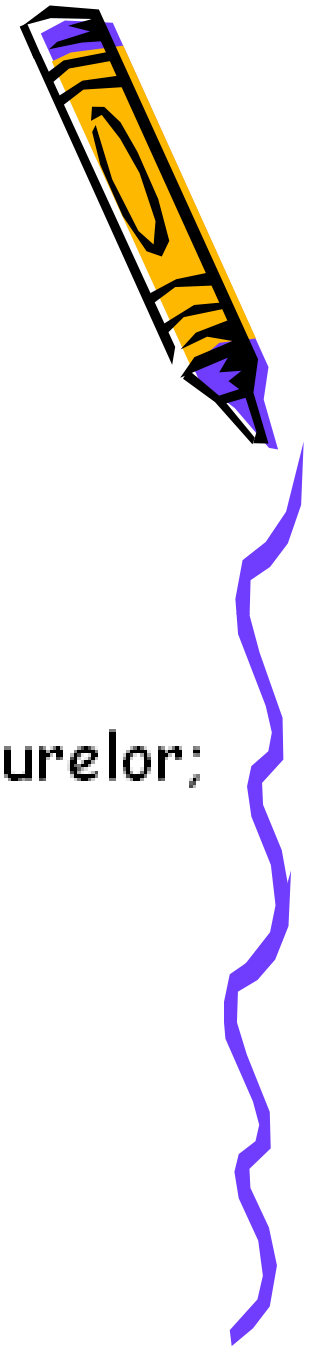

mediu inconjurator
baza de date
proces planificare
planuri
metode de prognoza
PROGNOZA
problemele sunt rezolvate multumitor?
nu
da
implementarea planului



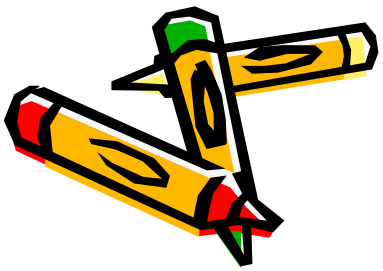
O asemenea abordare a problemei pare naturală, în practică, însă deseori se revizuie prognoza și nu planurile. Pe baza prognozelor corecte se pot face planificări în viitorul mai mult sau mai puțin apropiat și, mai ales, se pot alege deciziile corecte.



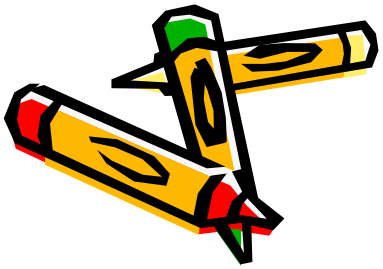
aplicatii



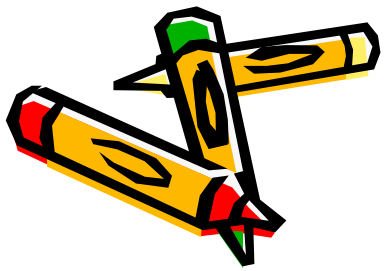
- Prognoza vremii;
- Prognoza uraganelor/cutremurelor;
- Prognoze economice.
- Planificarea transporturilor.



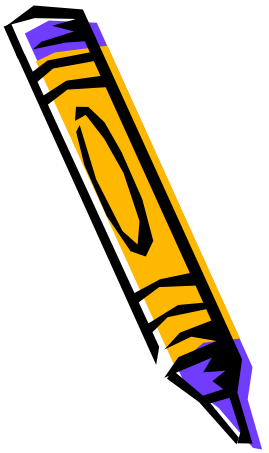
În cadrul modelului regresiv prezentat anterior (modelul logistic) şi utilizat la clasificare, variabila dependentă era categorială şi, variabila predictoare şi cea dependentă erau corelate.



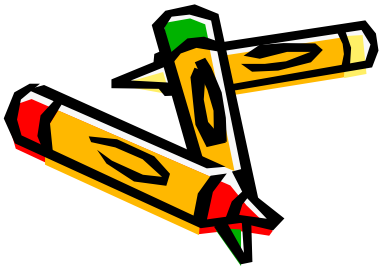
Scopul final al acestei abordări era prognoza comportamentului variabilei dependente pe baza valorilor variabilei explicative, prognoză efectuată utilizând ecuația ce descrie legătura dintre cele două seturi de date, corespunzătoare celor două variabile.



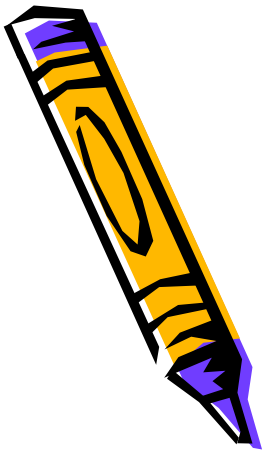
regresia liniara



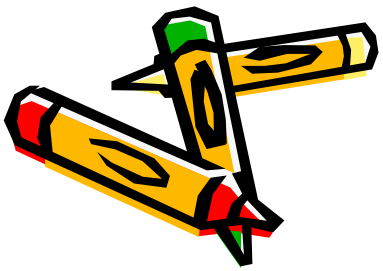
Modul de prezentare a legăturii liniare dintre două variabile, în general numerice, atunci când aceasta există, se numește *metoda regresiei liniare* (*regresia liniară*).



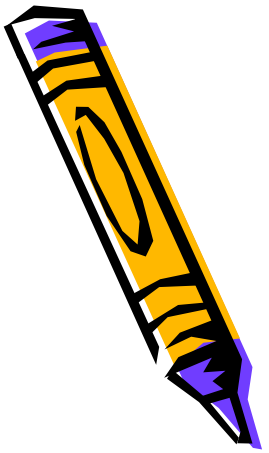
variabila predictor, variabila raspuns



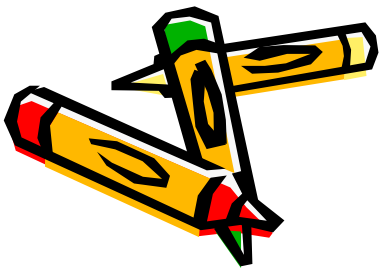
Se consideră una dintre variabile ca *variabilă independentă* sau *variabilă predictor*, iar cealaltă variabilă ca *variabilă dependentă* sau *variabilă răspuns*.

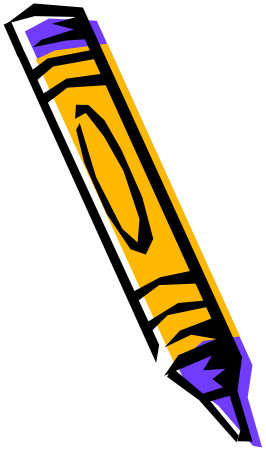


dreapta de regresie

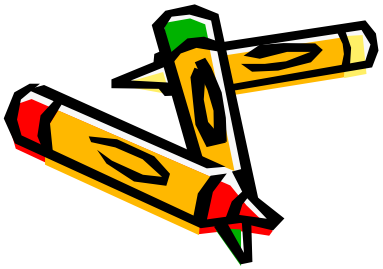


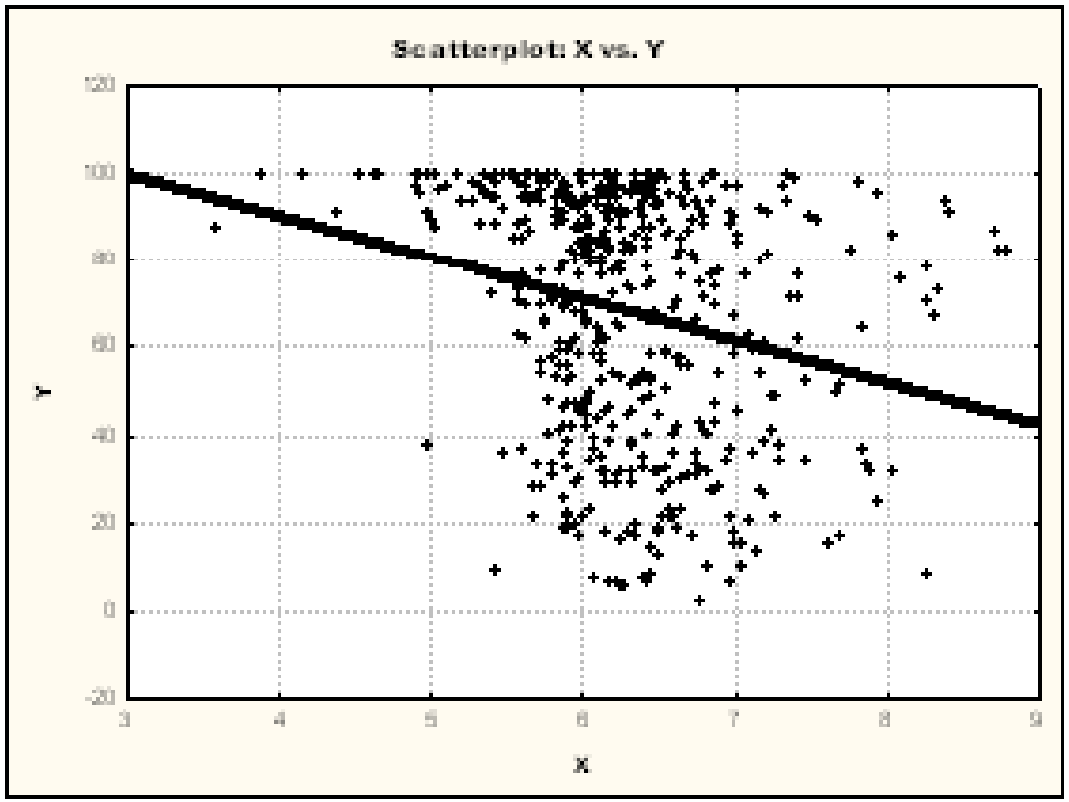
Legătura liniară dintre cele două variabile este descrisă de o ecuație liniară, *ecuația de regresie*, căreia îi corespunde geometric *dreapta de regresie*.



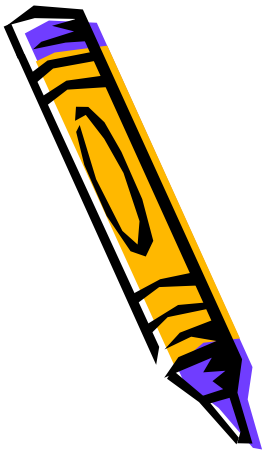


Ca metodologie, în cazul variabilelor numerice, variabila dependentă se distribuie pe axa ordonatelor, în timp ce variabila independentă se distribuie pe axa absciselor.




Scatterplot: X vs. Y
120
100
80
60
40
20
0
-20
Y
3
4
5
6
7
8
9
X

metoda celor mai mici patrate



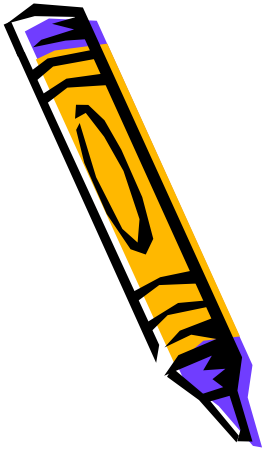
Ecuația dreptei de regresie se stabileşte pe baza metodei "*celor mai mici pătrate*".

Presupunem că obținem următorul tabel de valori, rezultat al măsurătorii unei mărimi fizice $f(x)$.

x_0	x_1	x_2	.....	x_n
y_0	y_1	y_2	.....	y_n

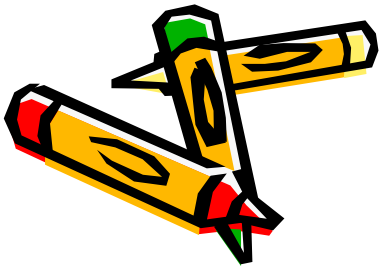
astfel încât $y_i = f(x_i), 0 \le i \le n$,

fiecare dintre valori fiind calculată cu o anumită eroare.



Suntem interesați de a calcula valoarea lui f în puncte distincte de $x_i, 0 \le i \le n$.

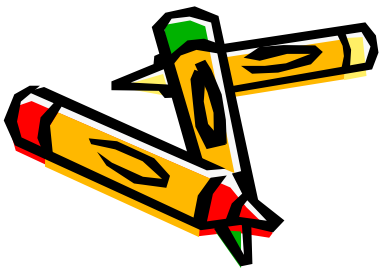
În anumite situații este important să cunoaștem cât de mult se abate graficul lui f de la o dreaptă $g(x) = a + bx$, unde a și b sunt parametri reali.

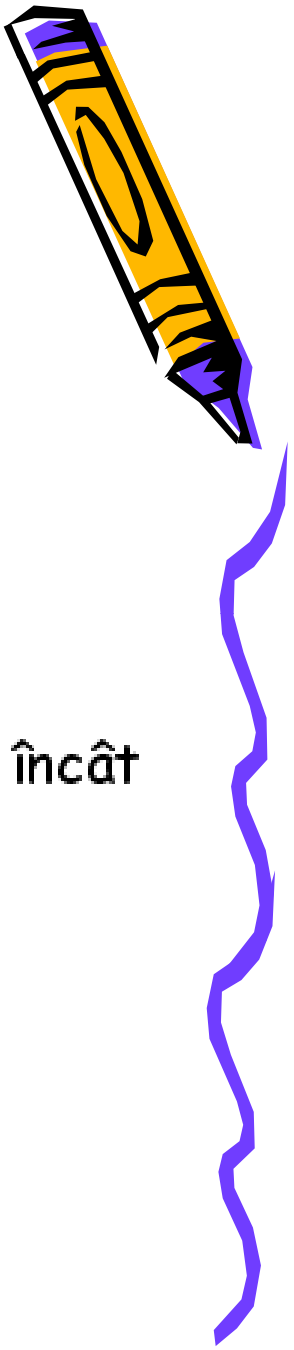


abaterea medie patratica

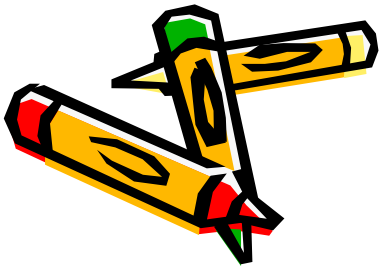
Considerăm *abaterea medie pătratică* în formula de aproximare, care este o funcție de două variabile, definită prin:

$$U(a,b) = \sum_{i=0}^{n} (f(x_i) - g(x_i))^2 = \sum_{i=0}^{n} (y_i - a - bx_i)^2$$



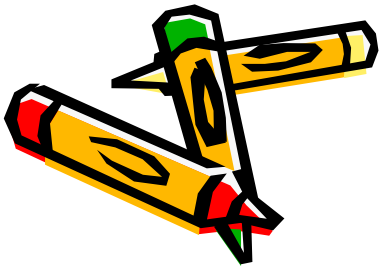


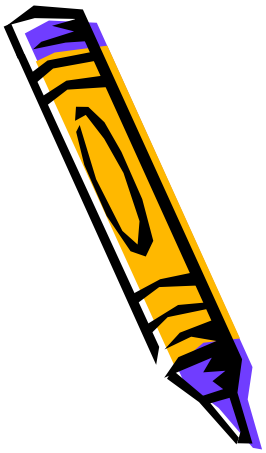
Metoda celor mai mici pătrate constă în determinarea parametrilor a şi b , astfel încât funcția $U(a,b)$ să fie minimă.



Rezolvând sistemul

$$\begin{cases} \dfrac{\partial U}{\partial a} = -2\sum_{i=0}^{n}(y_i - a - bx_i) = 0 \\ \dfrac{\partial U}{\partial b} = -2\sum_{i=0}^{n}(y_i - a - bx_i)\cdot x_i = 0 \end{cases}$$

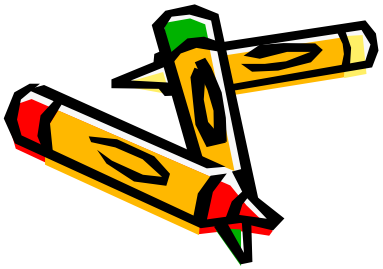


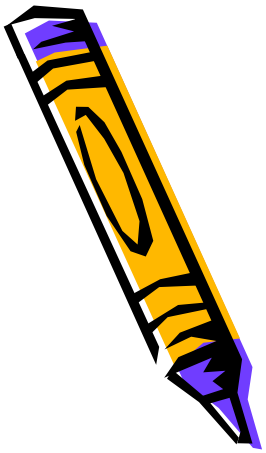


sistem echivalent cu:

$$\begin{cases} b\sum_{i=0}^{n} x_i + (n+1)\cdot a = \sum_{i=0}^{n} y_i \\ b\sum_{i=0}^{n} x_i^2 + a\sum_{i=0}^{n} x_i = \sum_{i=0}^{n} x_i y_i \end{cases},$$

obținem punctele critice (a_0, b_0).



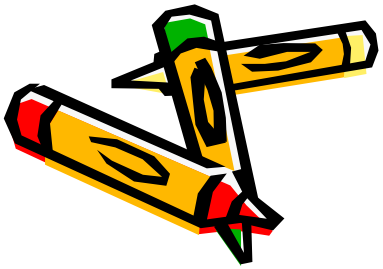


Avem:

$$\frac{\partial^2 U}{\partial a^2} = 2(n+1);$$

$$\frac{\partial^2 U}{\partial a \partial b} = 2\sum_{i=0}^{n} x_i ;$$

$$\frac{\partial^2 U}{\partial b^2} = 2\sum_{i=0}^{n} x_i^2 ,$$



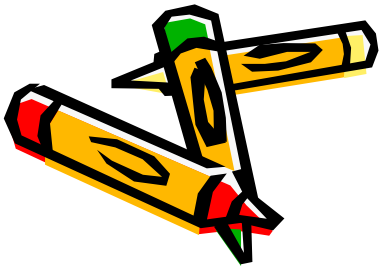
deci

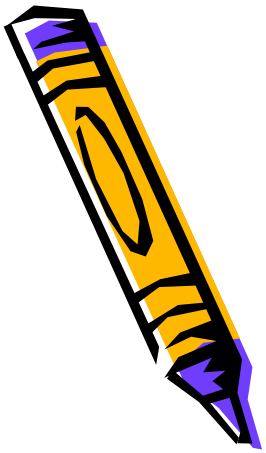
$$\frac{\partial^2 U}{\partial a^2} \geq 0;$$

$$\frac{\partial^2 U}{\partial a^2} \cdot \frac{\partial^2 U}{\partial b^2} - \left(\frac{\partial^2 U}{\partial a \partial b}\right)^2 = 4\left((n+1)\sum_{i=0}^{n} x_i^2 - \left(\sum_{i=0}^{n} x_i\right)^2\right) \geq 0,$$

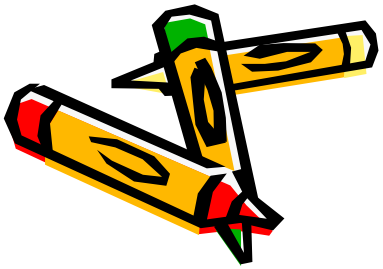
(conform inegalității Cauchy-Schwarz)

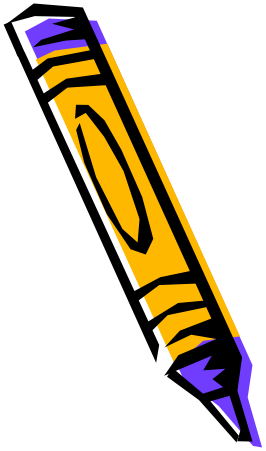
În concluzie, (a_0, b_0) este un punct de minim pentru funcția U.





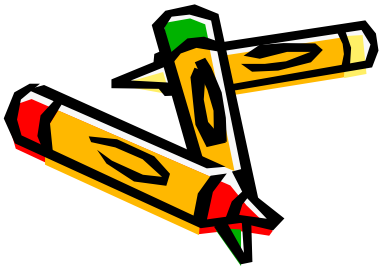
Se spune că dreapta $y = a_0 + b_0 x$ liniarizează optim datele experimentale $(x_i, y_i), 0 \leq i \leq n$ şi se numeşte *dreaptă de regresie* a lui y în raport cu x.

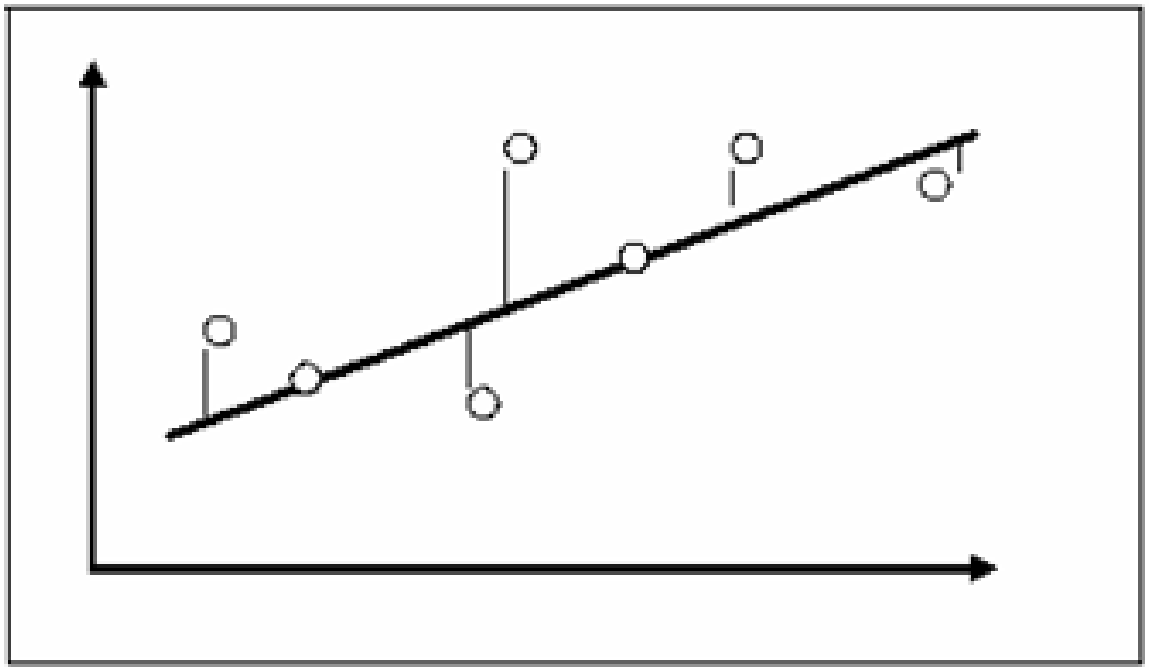


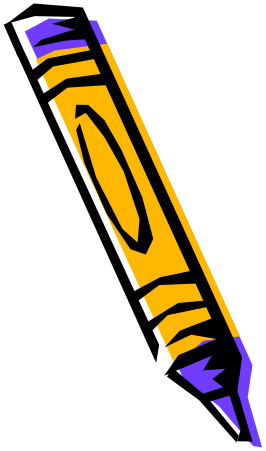


În cazul nostru, metoda constă în calcularea distanțelor (pe verticală) dintre punctele observate (reale) corespunzătoare cuplului de serii statistice şi punctele (imaginare) de pe o anumită dreaptă (de regresie), ce trece prin mijlocul ‚norului' de puncte generate de cuplurile de date.

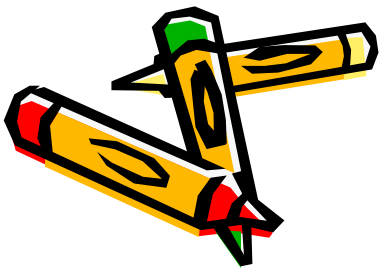
Aceste distanțe sunt cunoscute sub numele de *reziduuri.*

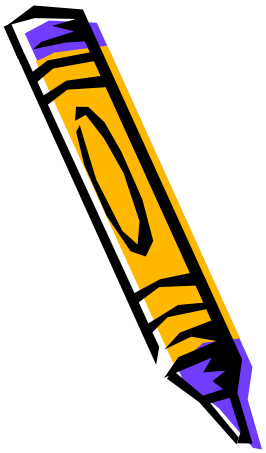




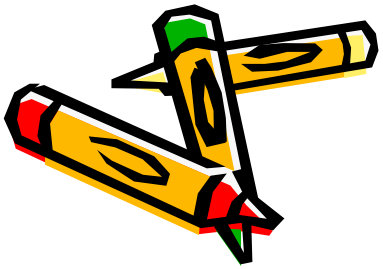


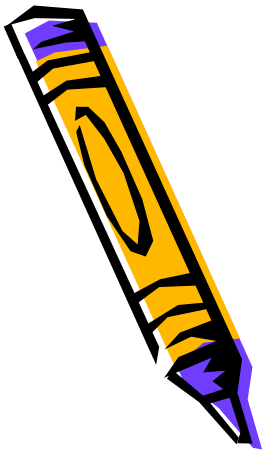
Se alege apoi acea dreaptă care trece prin ‚norul' de puncte, numită *dreapta de regresie*, dreapta pentru care suma pătratelor acestor reziduuri este minimă.





Dreapta de regresie este acea dreaptă (ideală) ce trece prin norul de puncte format de perechile de date ale celor două variabile şi care minimizează, distanța între date şi ea (minimizând suma pătratelor distanțelor).

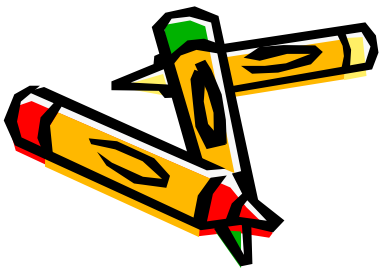


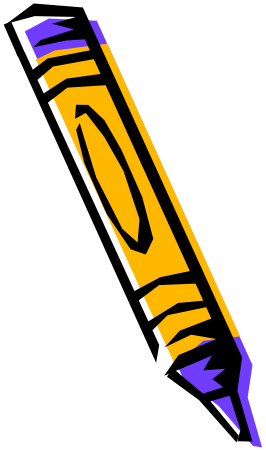


În final, obținem ecuația de regresie sub forma:

$$Y = a + bX$$

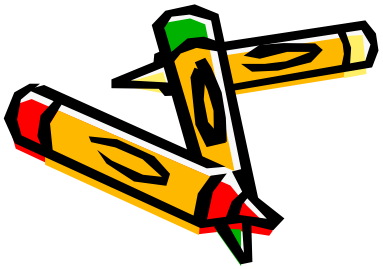
unde *a* se numeşte *interceptor* iar *b* se numeşte *coeficientul de regresie* -panta dreptei de regresie.

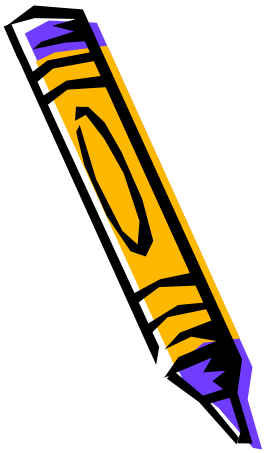




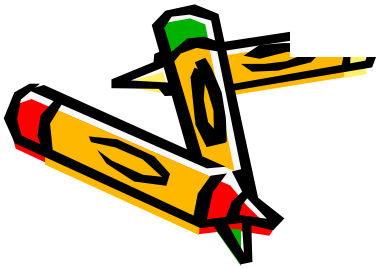
Tehnic vorbind, vom folosi regresia liniară când sunt îndeplinite următoarele trei ipoteze de lucru:

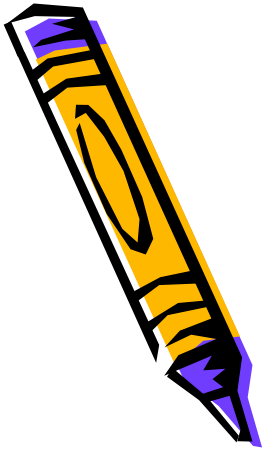
- Valorile variabilei dependente *Y* trebuie să aibă o repartiție normală (gaussiană);





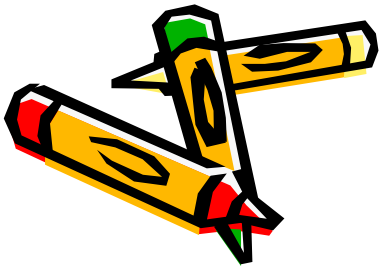
- Variabilitatea variabilei prognozate *Y* trebuie să fie asemănătoare cu cea a predictorului *X* (dispersia sau deviația standard asemănătoare);
- Legătura dintre cele două variabile, predictorul și variabila dependentă trebuie să fie liniară (verificare empirică pe baza 'norului' de puncte, care trebuie să aibă o 'formă' alungită -liniară).





Modul standard de a verifica simultan toate cele trei ipoteze de lucru este analiza statistică a reziduurilor.

Astfel, se poate demonstra că dacă toate cele trei ipoteze sunt verificate simultan, atunci reziduurile sunt normal repartizate de medie zero.



Din punct de vedere matematic, dreapta de regresie este dată de ecuația:

$$y = a + bx\,,$$

unde:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\,, \qquad a = \bar{y} - b \cdot \bar{x}\,.$$

Pentru uşurarea calculelor cu ajutorul unor programe de tipul MS Excel, folosind un tabel ca cel pe care l-am amintit la calculul coeficientului de corelaţie, se pot considera următoarele formule:

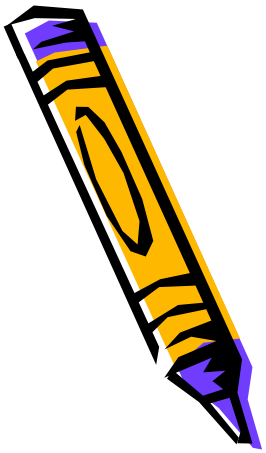
$$\sigma_{XX} = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2;\ \sigma_{YY} = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2;$$

$$\sigma_{XY} = \sum x_i \cdot y_i - \frac{1}{n}\sum x_i \cdot \sum y_i$$

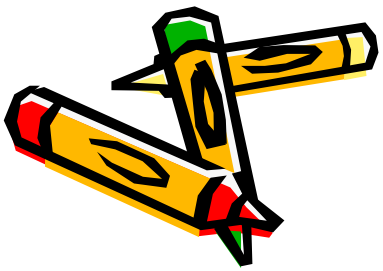
de unde obţinem:

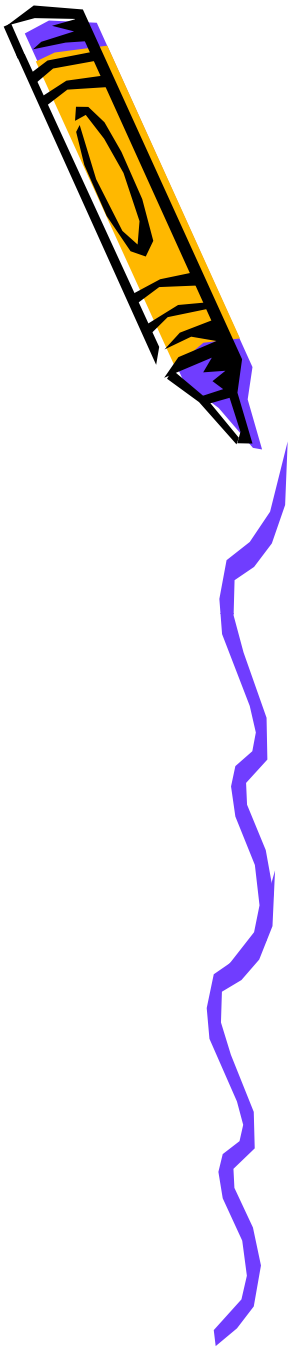
$$b = \frac{\sigma_{XY}}{\sigma_{XX}}.$$

exemple

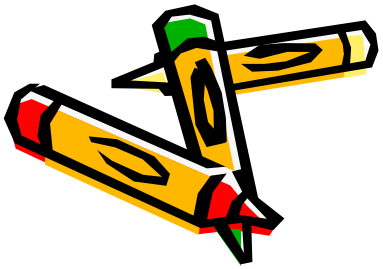


Din baza de date HOMES6, care conține informații despre prețul a 76 de case vândute în Eugene, zona de sud, Oregon, în 2005, alegem 5 cazuri, pentru care cunoaștem prețul (în mii de dolari) Y și suprafața locuibilă (în mii de metri pătrati) X.



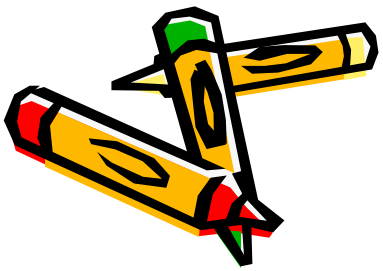


	X = suprafața (mii de m^2)	Y = prețul (în mii $)
1	1.683	259.9
2	1.708	259.9
3	1.922	1.922
4	2.053	2.053
5	2.269	2.269



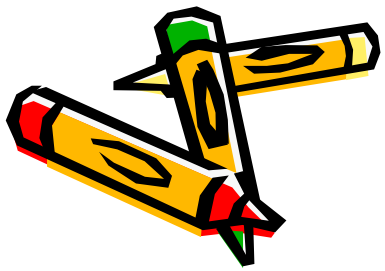
```
»X=[1.683 1.708 1.922 2.053 2.269];
»Y=[259.9 259.9 269.9 270.0 285.0];
»A=[X' Y'];
» corrcoef(A)
ans =
    1.0000    0.9721
    0.9721    1.0000
```

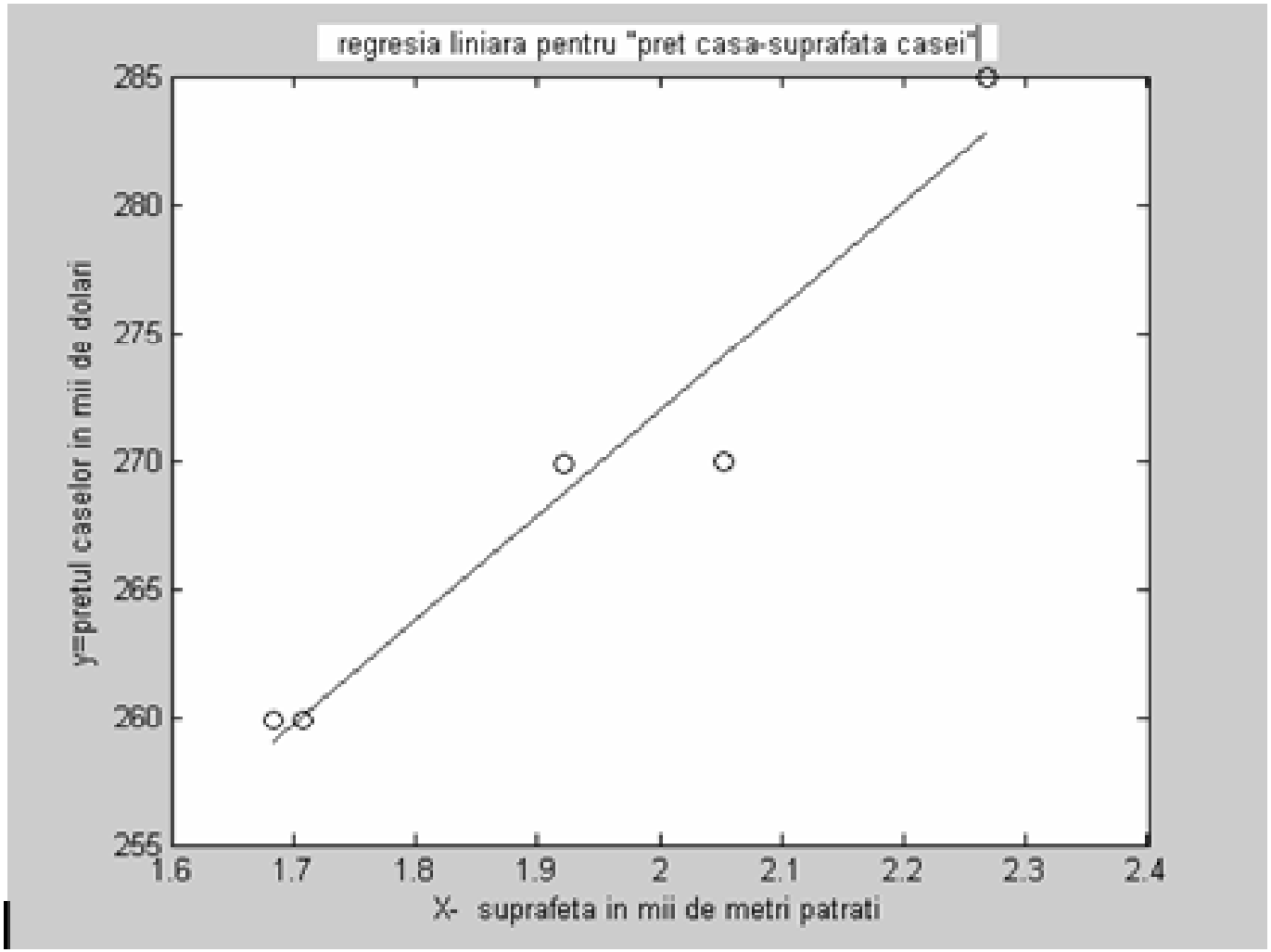
Se observă că variabilele sunt puternic (pozitiv) corelate, coeficientul de corelație fiind 0.9721.

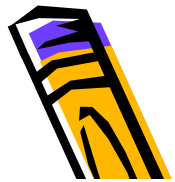


Calculăm dreapta de regresie, folosind formulele prezentate mai sus şi o desenăm în acelaşi cadran cu diagrama împrăştierii:

```
»b=(X'*Y'-5*mean(X)*mean(Y))/((norm(X))^2-5*(mean(X))^2)
   b =
      40.8001
»a=mean(Y)-b*mean(X)
   a =
   190.3182
»plot(X,a+b*X);hold on
» plot(X,Y,'ko');hold off
```




regresia liniara pentru "pret casa-suprafata casei"
y=pretul caselor in mii de dolari
X- suprafeta in mii de metri patrati
285
280
275
270
265
260
255
1.6
1.7
1.8
1.9
2
2.1
2.2
2.3
2.4

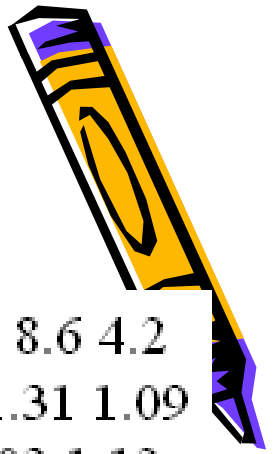


Am prezentat date culese de la un lot de 24 pacienți, având diabet zaharat, privind două variabile:

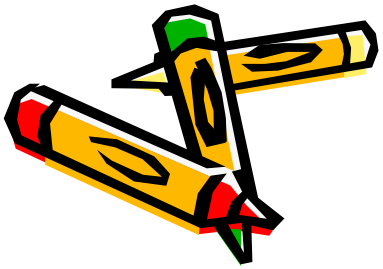
- *G* (mmol/l) - glucoza în sânge
- *Vcf* (% / s) - viteza medie de contracție a ventriculului stâng, obținută prin eco-cardiografie.,

date pentru care am calculat coeficientul de corelație a celor două variabile și anume 0.3755.

Vom calcula dreapta de regresie (*Vcf* ca funcție liniară de glucoza în sânge, valoare mai greu de obținut în practica curentă):

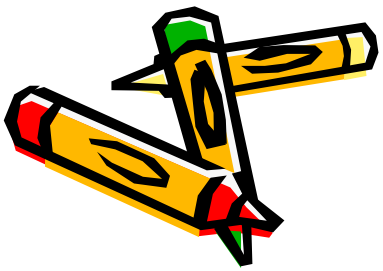


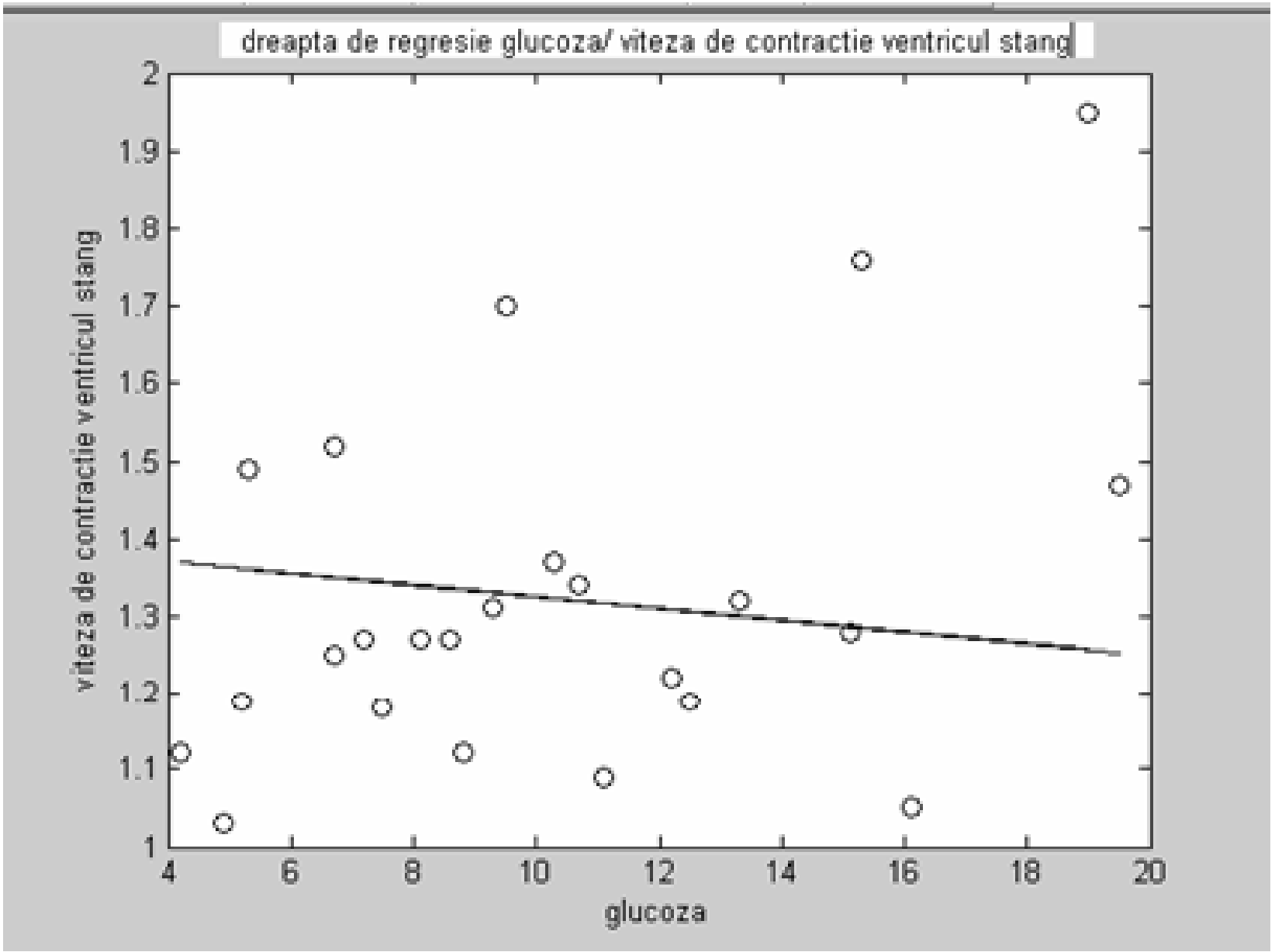
```
»G=[15.3 10.7 8.1 19.5 7.2 5.3 9.3 11.1 7.5 12.2 6.7 5.2 19 15.1 6.7 8.6 4.2
 10.3 12.5 16.1 13.3 4.9 8.8 9.5];V=[1.76 1.34 1.27 1.47 1.27 1.49 1.31 1.09
 1.18 1.22 1.25 1.19 1.95 1.28 1.52 1.65 1.12 1.37 1.19 1.05 1.32 1.03 1.12
 1.70];
» V=[1.76 1.34 1.27 1.47 1.27 1.49 1.31 1.09 1.18 1.22 1.25 1.19 1.95 1.28
 1.52 1.27 1.12 1.37 1.19 1.05 1.32 1.03 1.12 1.70];
» A=[G' V']
» b=(G*V'-24*mean(G)*mean(V))/((norm(G))^2-40*(mean(G))^2)
b =
   -0.0075
» a=mean(V)-b*mean(G)
a =
    1.4010
```



Să desenăm dreapta de regresie şi diagrama împrăştierii:

```
»plot(G,a+b*G);hold on
»plot(G,V,'ko');hold off
```



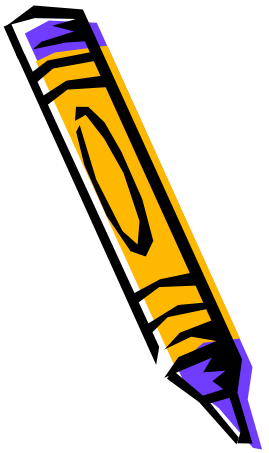

dreapta de regresie glucoza/ viteza de contractie ventricul stang
viteza de contractie ventricul stang
glucoza
2
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1
4
6
8
10
12
14
16
18
20

Cu toate că valoarea coeficientului de corelație nu este edificatoare. nivelul de semnificație (calculat cu STATISTICA) $p = 0.041$ atestă o corelație semnificativă.

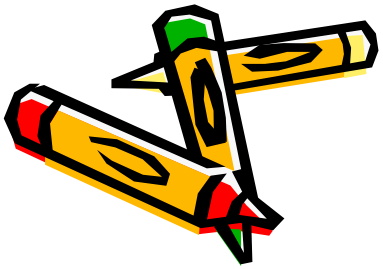
Presupunem că valoarea glucozei în sânge pentru un pacient este 15.1.

Atunci valoarea estimată (prognozată pe baza regresiei liniare) va fi:

```
»V1=a+b*15.1
    V1 =
          1.2871
```

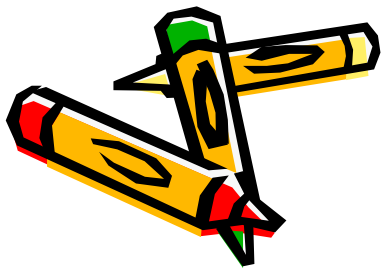


Dacă vom considera perechi de date provenind de la două grupuri diferite de obiecte și având aceleași semnificații, putem folosi dreptele de regresie calculate pentru fiecare grup pentru a compara cele două grupuri.

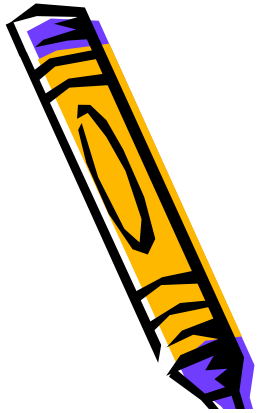


Dacă, de exemplu, cele două drepte de regresie au aproximativ aceiași pantă (sunt paralele), atunci putem considera diferența pe axa verticală (*Y*) ca fiind diferența între mediile variabilei *Y* între cele două grupuri, observație ce este apoi urmată de o testare a semnificației statistice a diferenței.

O asemenea analiză statistică face parte dintr-un studiu statistic mai vast, care se numește *analiza covarianțelor*.



regresia polinomiala



În cazul când legătura dintre cele două variabile statistice nu este liniară și totuși bănuim că există, avem o *regresie neliniară*, de exemplu *regresia polinomială*.
Atunci, în loc de a găsi dreapta de regresie, se găsește curba respectivă de regresie (corespunzătoare, de exemplu, ecuației polinomiale de regresie).
Un exemplu de astfel de regresie polinomială va furniza o ecuație de forma:

$$Y = a + b_1 X + b_2 X^2 + \ldots + b_n X^n ,$$

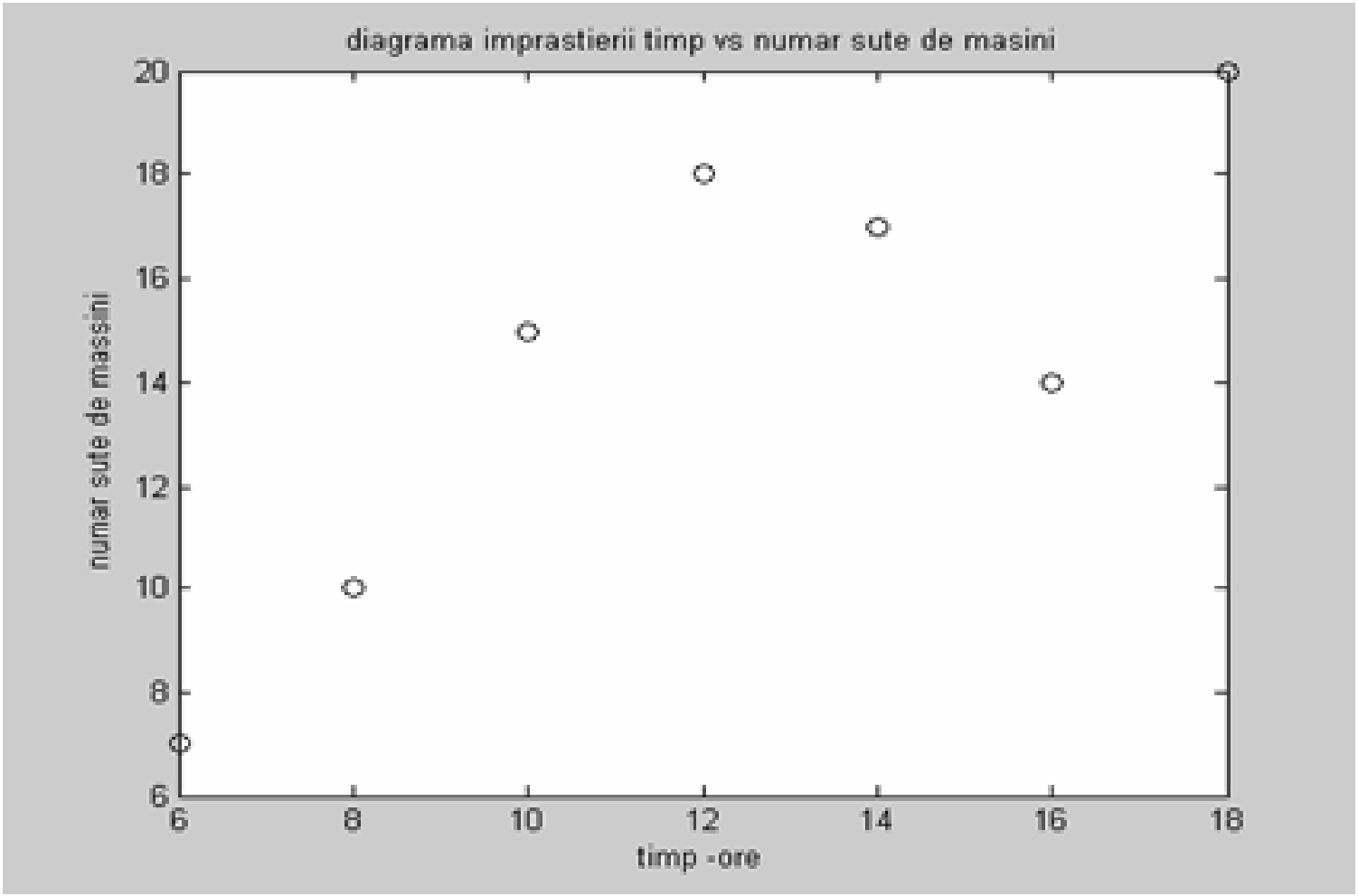
unde, ca de obicei, X este variabila predictoare iar Y este variabila prognozată, adică variabila răspuns.

exemple

Timp de 14 ore (6h-20h) s-a monitorizat traficul dintr-o anumită intersecție, luându-se în considerare totalul mașinilor ce traversează intersecția în intervalul a două ore, cum reiese din următorul tabel

ora	numărul sutelor de mașini
8	7
10	10
12	15
14	18
16	17
18	14
20	20

```
»t=6:2:18;n=[7 10 15 18 17 14 20];
» plot(t,n,'ko')
```



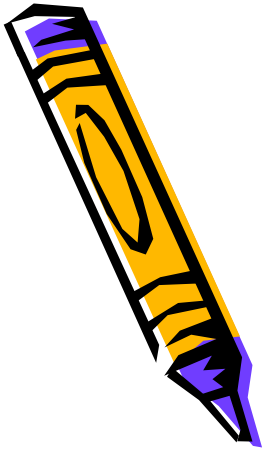
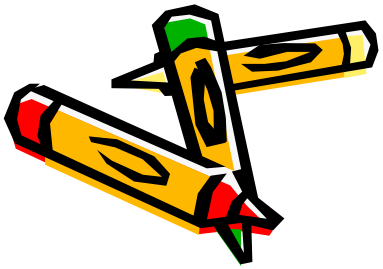
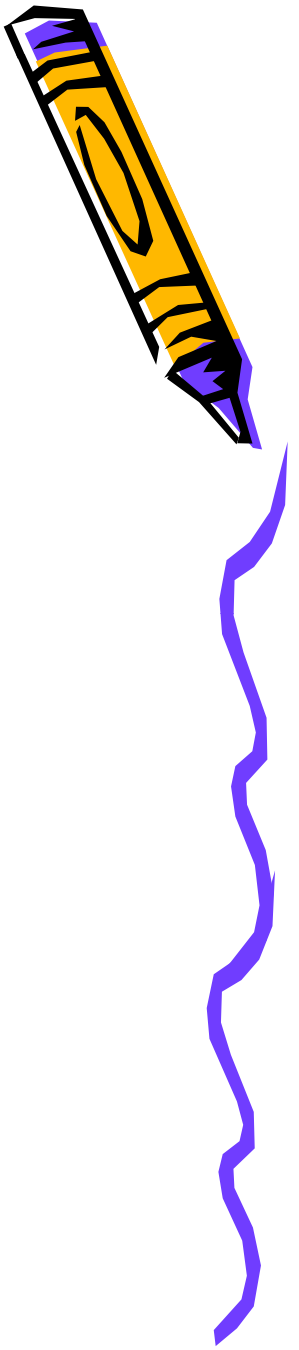


Diagrama împrăştierii sugerează că datele ar putea fi modelate ca o funcție polinomială

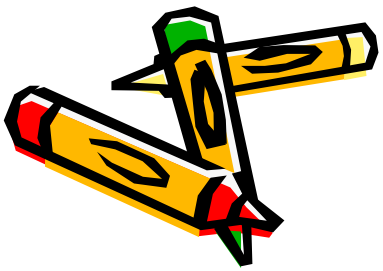
$y = a_0 + a_1 t + a_2 t^2$

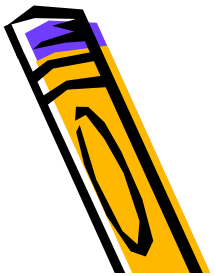
Coeficienții necunoscuți vor fi calculați folosind metoda celor mai mici pătrate. Vom rezolva sistemul de necunoscute a_0, a_1, a_2:





$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix} = \begin{pmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \\ 1 & t_4 & t_4^2 \\ 1 & t_5 & t_5^2 \\ 1 & t_6 & t_6^2 \\ 1 & t_7 & t_7^2 \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}$$

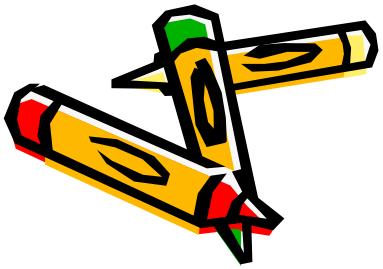


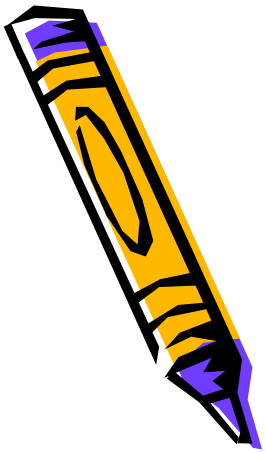


```
» t=6:2:18;n=[7 10 15 18 17 14 20];
» A=[1 t(1) t(1)^2;1 t(2) t(2)^2;1 t(3) t(3)^2;1 t(4) t(4)^2;1 t(5) t(5)^2;
1 t(6) t(6)^2; 1 t(7) t(7)^2];
» a=A\n'
    a =
      -8.6429
       3.2321
      -0.0982
```

Modelul polinomial de ordinul doi al datelor este:

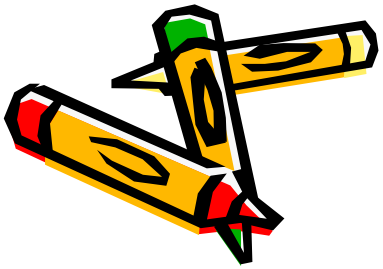
$$y = -8.6429 + 3.2321 \cdot t - 0.0982 \cdot t^2$$





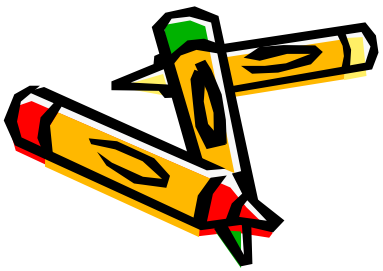
Putem prognoza câte sute de maşini vor trece prin intersecție în intervalul 21h-23h.

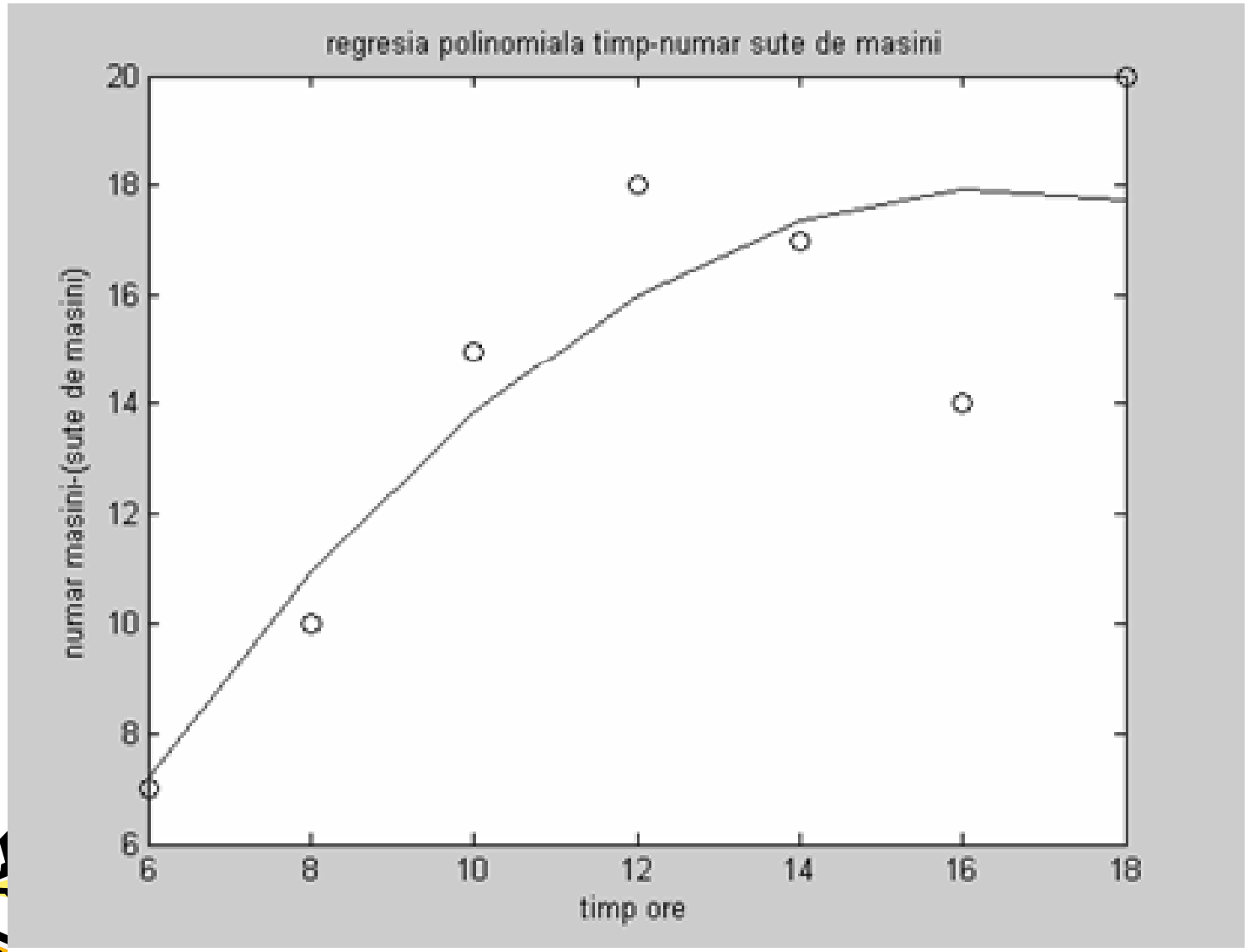
```
» syms t
» f23=subs(f,t,23)
    f23 =
          13.7411
```

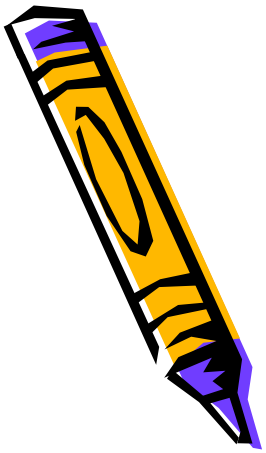


În final desenăm atât diagrama împrăştierii, cât şi curba de regresie polinomială în acelaşi cadran:

```
» plot(t,n,'kO');hold on
» plot(t,a(1)+a(2)*t+a(3)*t.^2);hold off
```




regresia polinomiala timp-numar sute de masini
numar masini-(sute de masini)
timp ore
20
18
16
14
12
10
8
6
6
8
10
12
14
16
18

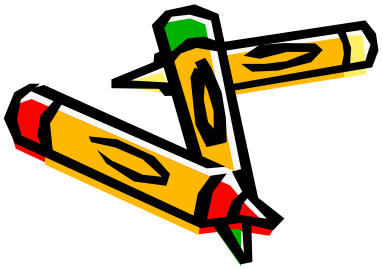


Să înlocuim funcția polinomială cu următoarea funcție:

$$y = a_0 + a_1 \cdot e^{-t} + a_2 \cdot t \cdot e^{-t}$$

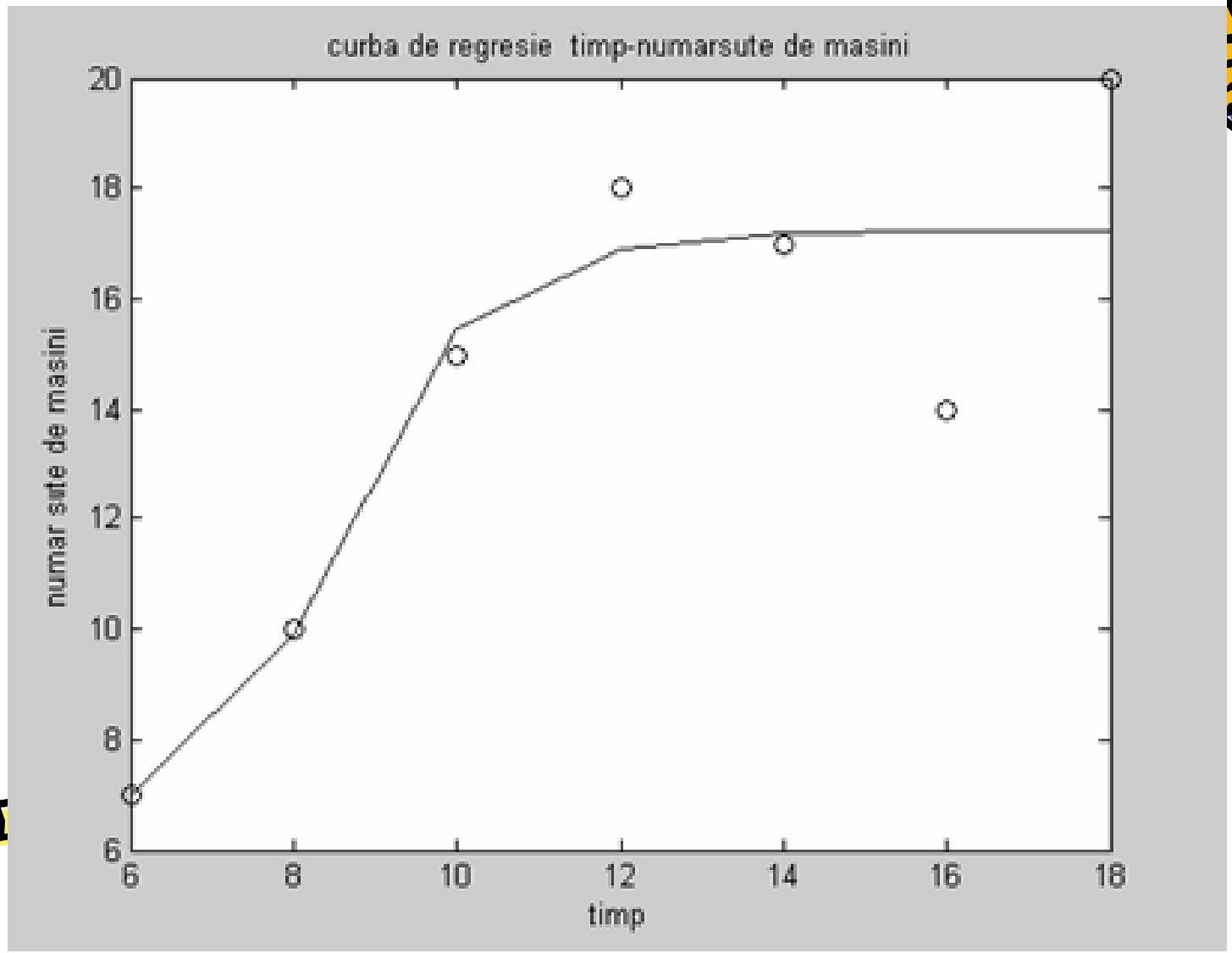
(*linear in the parameters regression*)

Coeficienții a_0, a_1, a_2 vor fi calculați folosind metoda celor mai mici pătrate:



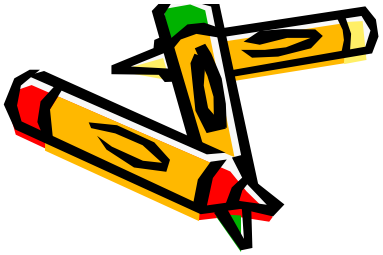
```
» t=6:2:18;n=[7 10 15 18 17 14 20];
» A=[1 exp(-t(1)) t(1)*exp(-t(1));1 exp(-t(2)) t(2)*exp(-t(2));
1 exp(-t(3)) t(3)*exp(-t(3));1 exp(-t(4)) t(4)*exp(-t(4));
1 exp(-t(5)) t(5)*exp( -t(5));1 exp(-t(6)) t(6)*exp( -t(6));
|1 exp(-t(7)) t(7)*exp(-t(7))];
»  a=A\n'
     a =
       1.0e+004 *
        0.0017
        4.9000
       -0.8856

»  plot(t,a(1)+a(2)*exp(-t)+a(3)*t.*exp(-t));hold on
» plot(t,n,'kO');hold off
```


curba de regresie timp-numarsute de masini
numar sute de masini
timp
20
18
16
14
12
10
8
6
6
8
10
12
14
16
18

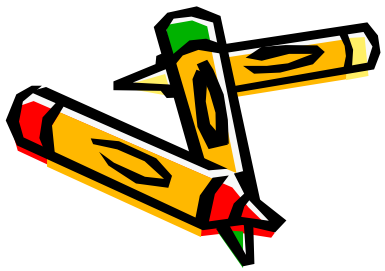
Considerăm situația numărului de pacienți internați într-o secție a unui spital pe parcursul a șase luni, situație reflectată în următorul tabel.

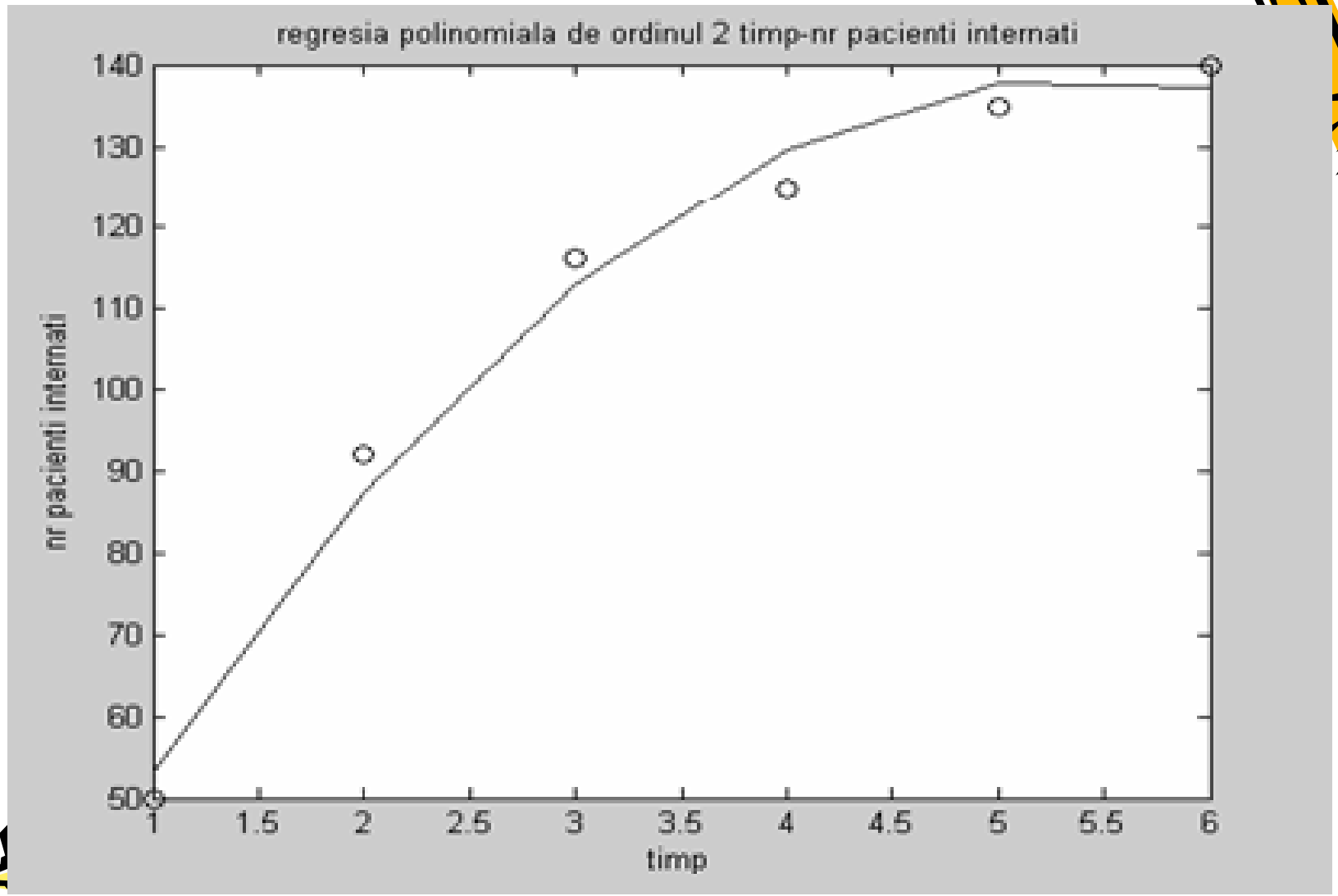
	Număr pacienți internați
1 (august)	50
2 (septembrie)	92
3 (octombrie)	116
4 (noiembrie)	125
5 (decembrie)	135
6 (ianuarie)	140



Pentru început, vom determina diagrama împrăştierii
şi curba de regresie polinomială:

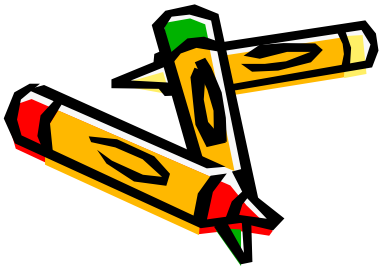
```
» t=1:6;p=[50 92 116 125 135 140];
» A=[1 t(1) t(1)^2;1 t(2) t(2)^2;1 t(3) t(3)^2;1 t(4) t(4)^2;1 t(5) t(5)^2;
1 t(6) t(6)^2];
» a=A\p';
»  plot(t,a(1)+a(2)*t+a(3)*t.^2);hold on
»  plot(t,p,'kO');hold off
```

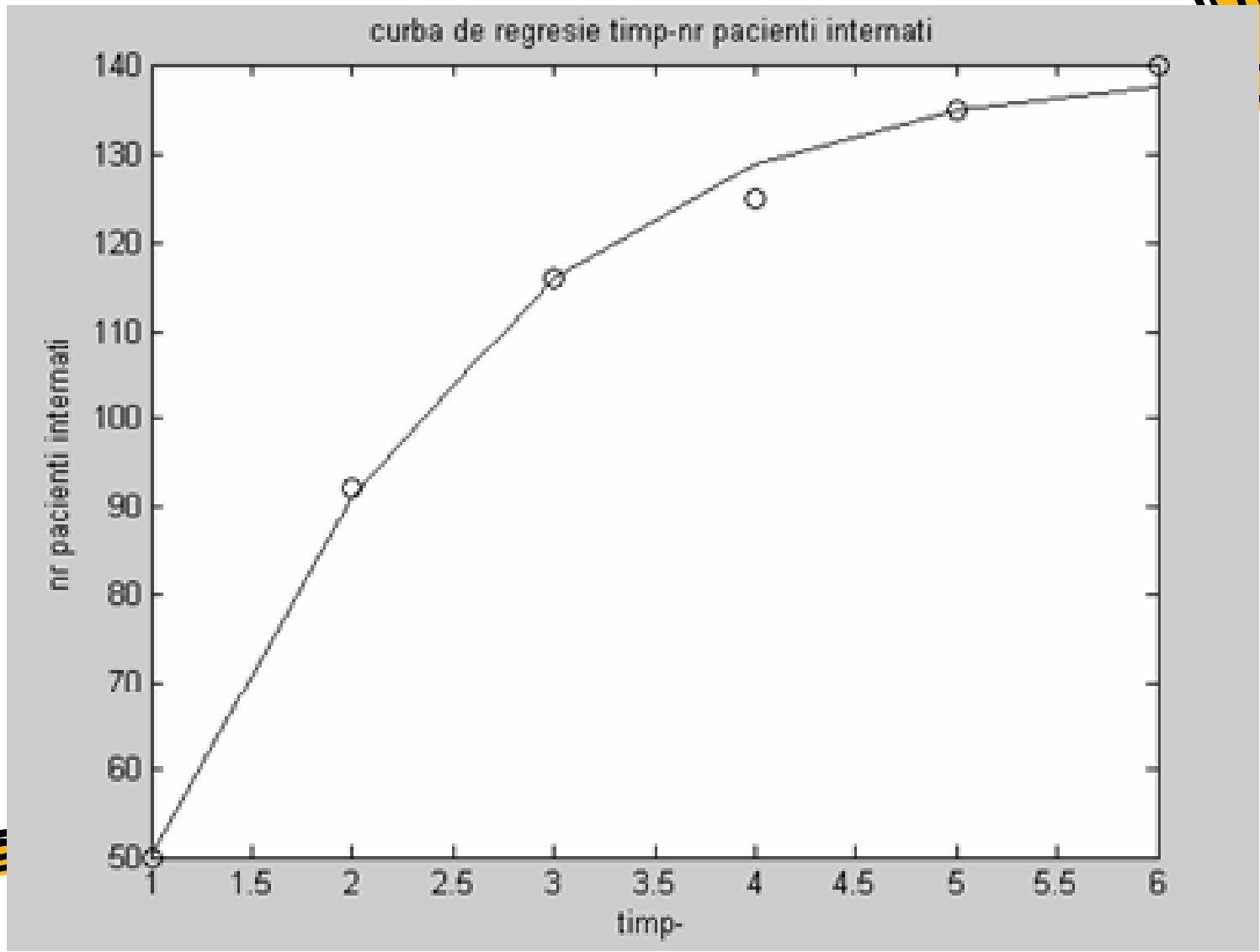



regresia polinomiala de ordinul 2 timp-nr pacienti internati
nr pacienti internati
timp
140
130
120
110
100
90
80
70
60
50
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6

Rezolvăm aceeaşi problemă utilizând „linear in the parameters regression”:

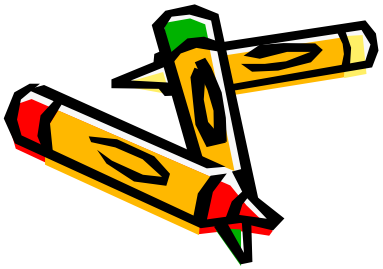
```
» t=1:6;p=[50 92 116 125 135 140];
»  A=[1 exp(-t(1)) t(1)*exp(-t(1));1 exp(-t(2)) t(2)*exp(-t(2));
1 exp(-t(3)) t(3)*exp(-t(3));1 exp(-t(4)) t(4)*exp(-t(4));
1 exp(-t(5)) t(5)*exp( -t(5));1 exp(-t(6)) t(6)*exp( -t(6))];
»  a=A\p';
»  plot(t,a(1)+a(2)*exp(-t)+a(3)*t..*exp(-t));hold on
»  plot(t,n,'ko');hold off
```

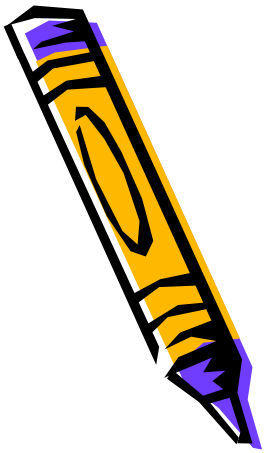



curba de regresie timp-nr pacienti internati
nr pacienti internati
timp-
140
130
120
110
100
90
80
70
60
50
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6

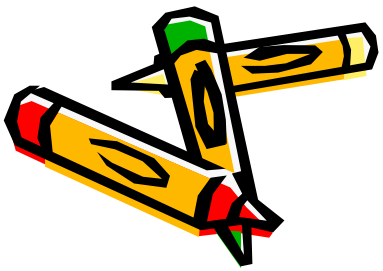
metoda regresiei liniare multiple

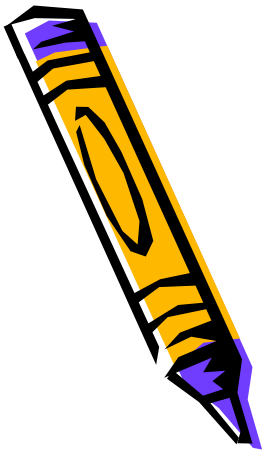
Metoda regresiei liniare se poate extinde de la cupluri de două variabile la mai multe variabile prin metoda *regresiei liniare multiple*, caz în care avem o variabilă dependentă și mai multe variabile predictive.



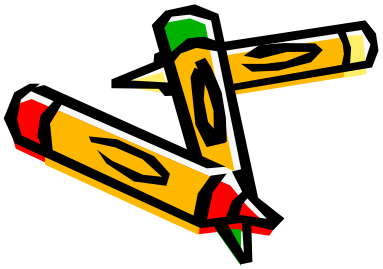


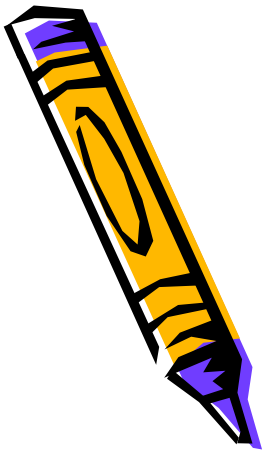
Să considerăm, pe de o parte, un set de *n* variabile aleatoare $X_1, X_2, \ldots, X_n$, considerate independente și care, din punct de vedere probabilistic, reprezintă variabilele ce guvernează *n* factori predictivi independenți (care acționează independent unul față de altul).



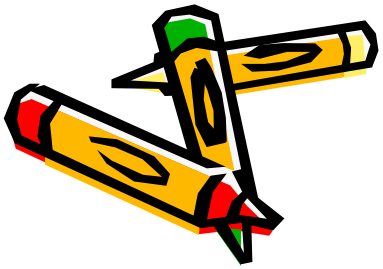


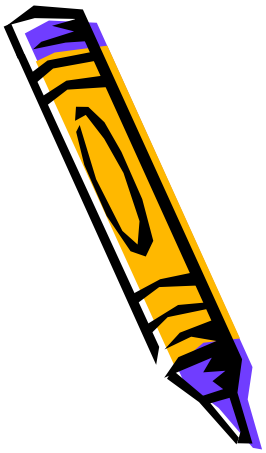
Un exemplu clasic în acest sens este cel din domeniul medical, în cazul unei anumite boli. De exemplu, în cazul oncologiei, putem considera că X_1 reprezintă diametrul mediu al tumorii, X_2 reprezintă vârsta, X_3 reprezintă un factor de risc, ca de exemplu, fumatul ş.a.m.d.



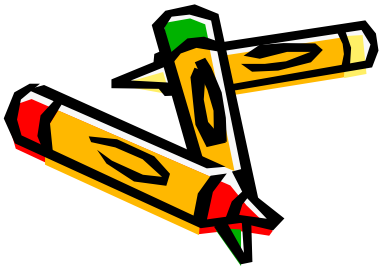


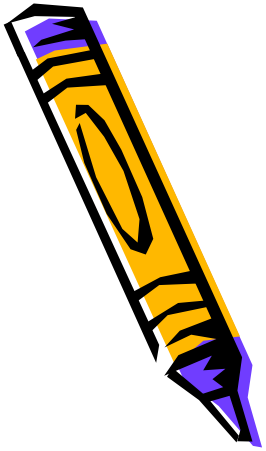
Pe de altă parte, vom considera încă o variabilă aleatoare, notată Y, care va juca un rol special în analiza regresivă. Această variabilă, numită şi variabilă dependentă (prognozată), va fi cea ale cărei valori vor fi estimate pe baza analizei regresive liniare multiple, plecând de la valorile factorilor predictivi.



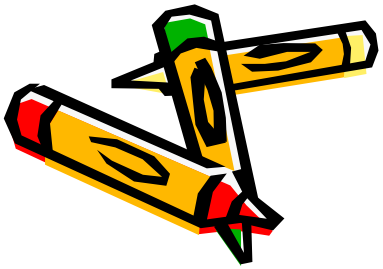


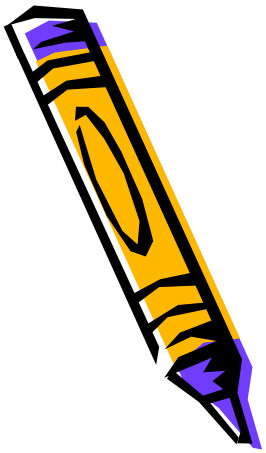
De exemplu, putem considera această variabilă ca fiind variabila care guvernează repartiția timpilor de supraviețuire și ale cărei valori vrem să le evaluăm pe baza valorilor variabilelor predictive.



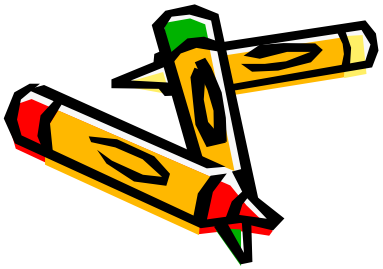


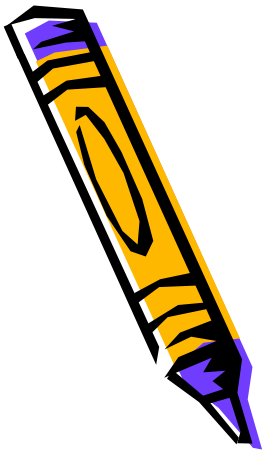
Deducând ecuația de regresie pe baza datelor culese de la un lot de pacienți cu cancer, se poate prognoza timpul pe care îl mai are de trăit un anumit pacient nou, introducând în ecuație valorile particulare ale acestuia pentru fiecare variabilă predictivă în parte.





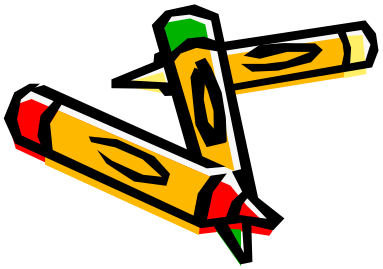
Dispunem, pe de o parte, de *n* variabile predictive sau explicative şi, pe de altă parte, de o variabilă prognozată, explicată, numită şi răspuns sau efect, pe care vrem să o deducem, cunoscând valorile celorlalte variabile.

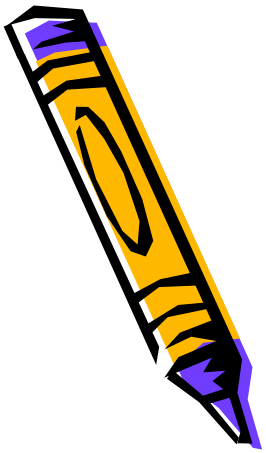




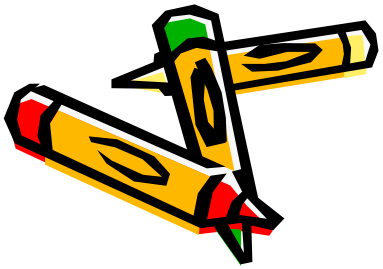
Înainte de începerea analizei statistice, trebuie verificate anumite ipoteze legate de natura variabilelor modelului regresiv.

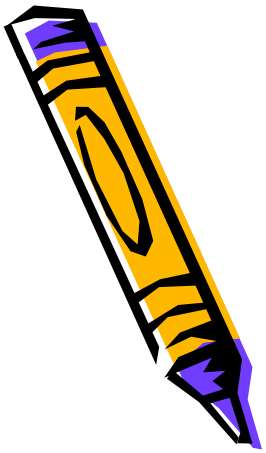
Astfel, variabilele $X_1, X_2, \ldots, X_n$ şi Y trebuie să verifice următoarele două condiții:





- legătura între variabila răspuns Y şi variabilele predictive X_i trebuie să fie una liniară, aşa cum arată şi numele metodei (se verifică cu mijloace statistice clasice),
- efectele fiecărei variabile sunt independente, în principiu, de celelalte.

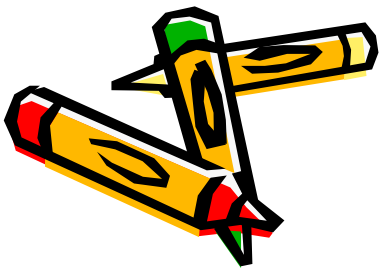


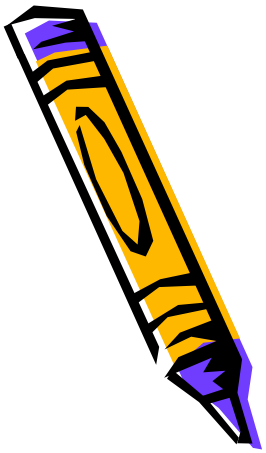


Odată stabilită validitatea condițiilor de aplicare a metodei, se trece la obținerea ecuației de regresie liniară multiplă, care este de forma:

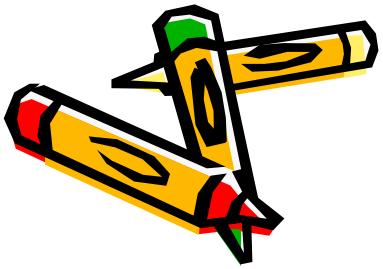
$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + ... + b_n \cdot X_n \ ,$$

unde $b_1, b_2, ..., b_n$ se numesc *coeficienții de regresie*, iar b_0 *interceptor*.



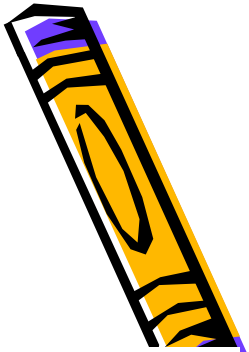


Vom aplica metoda regresiei multiple standard, care reprezintă o generalizare a regresiei simple, considerând hiperplanul care trece prin „norul" multidimensional al tuturor variabilelor. Calculele efective sunt bazate tot pe metoda „celor mai mici pătrate".



exemplu

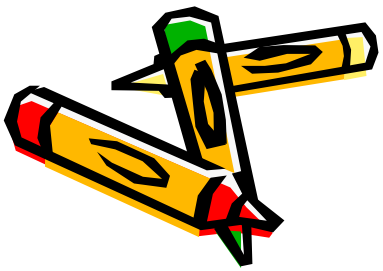
Reluăm exemplul cu prețul caselor din Eugene -Oregon, adăugând un nou predictor, categoria lotului de casă, predictor ce are nevoie de câteva explicații suplimentare. Este natural ca suprafața lotului pe care este situată casa să fie un factor important în stabilirea prețului casei. Pe de alta parte, există o anume diferență de preț pentru loturile de 4000 m^2 și cele de 6000 m^2 și altă diferență de preț între loturile de 24000 m^2 și cele de 26000 m^2.

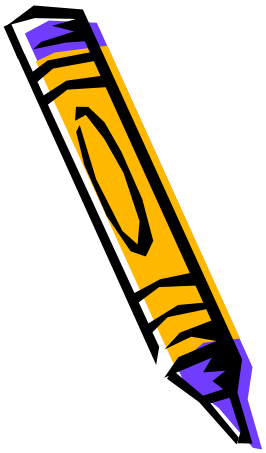


În acest sens, agenții de vânzări au definit niște categorii ale mărimii suprafeței lotului, pe baza experienței lor, categorii ce corespund la creșteri aproximativ egale ale prețului proprietăților.

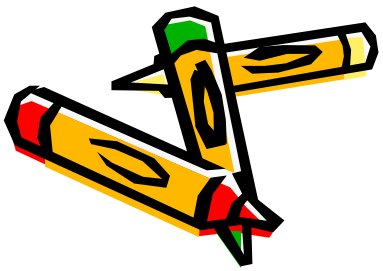
Prezentăm aceste categorii, specificând că este vorba de mii de metri pătrați:

suprafața lotului	0-3	3-5	5-7	7-10	10-15	15-25
categoria	1	2	3	4	5	6

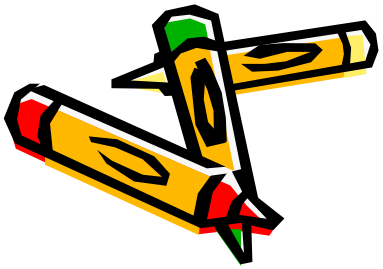


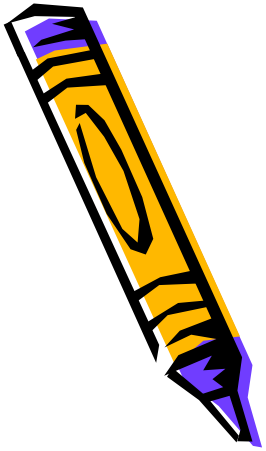


	suprafața casei (mii de m^2)	categoria lotului (mii de m^2)	preț proprietate (mii $)
1	1.888	2	252.5
2	1.683	5	259.9
3	1.708	4	259.9
4	1.922	4	269.9
5	2.053	3	270.0
6	2.269	3	285.0



```
» x1=[1.888 1.683 1.708 1.922 2.053 2.269];
» x2=[2 5 4 4 3 3 ];
» y=[252.5 259.9 259.9 269.9 270 285.0];
» A=[x1' x2' y'];
» corrcoef(A)
ans =
    1.0000   -0.5669    0.8260
   -0.5669    1.0000   -0.0251
    0.8260   -0.0251    1.0000
```

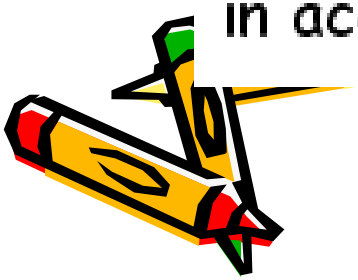


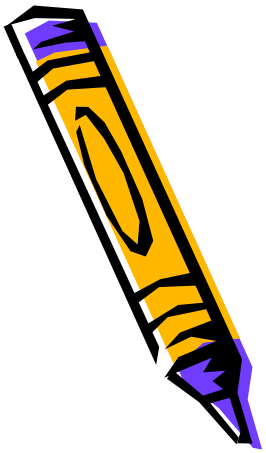


Se observă că prețul proprietății este pozitiv corelat cu suprafața casei;

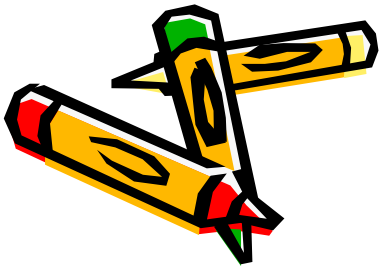
faptul că avem coeficientul de corelație -0.0251 între suprafața casei şi categoria lotului nu implică independența variabilelor *suprafața casei* şi *categoria lotului*.

Calculăm suprafața de regresie (planul de regresie în acest caz) folosind metoda celor mai mici pătrate:



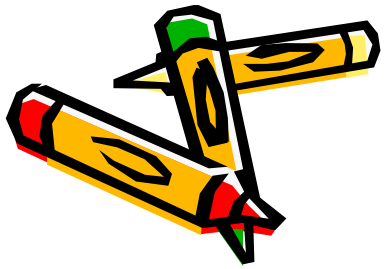


```
» B=[1 x1(1) x2(1);1 x1(2) x2(2);1 x1(3) x2(3);1 x1(4) x2(4);1 x1(5) x2(5);
1 x1(6) x2(6)]
»  a=B\y'
     a =
       122.3570
        61.9756
         7.0911
```

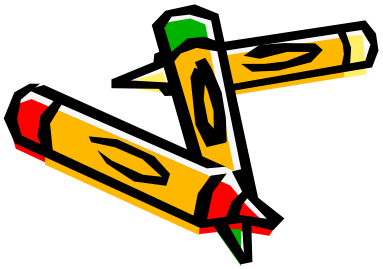


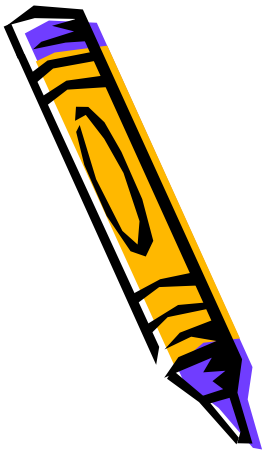
Folosind rezultatul obținut, putem prognoza prețul proprietăților, de exemplu să îl calculăm în următoarele cazuri:

- suprafața casei este de 1805 m^2 și lotul de casă este din categoria 2;
- suprafața casei este de 1200 m^2 și lotul de casă este din categoria 1;

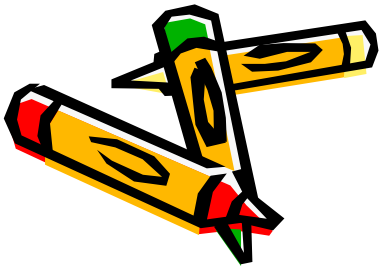


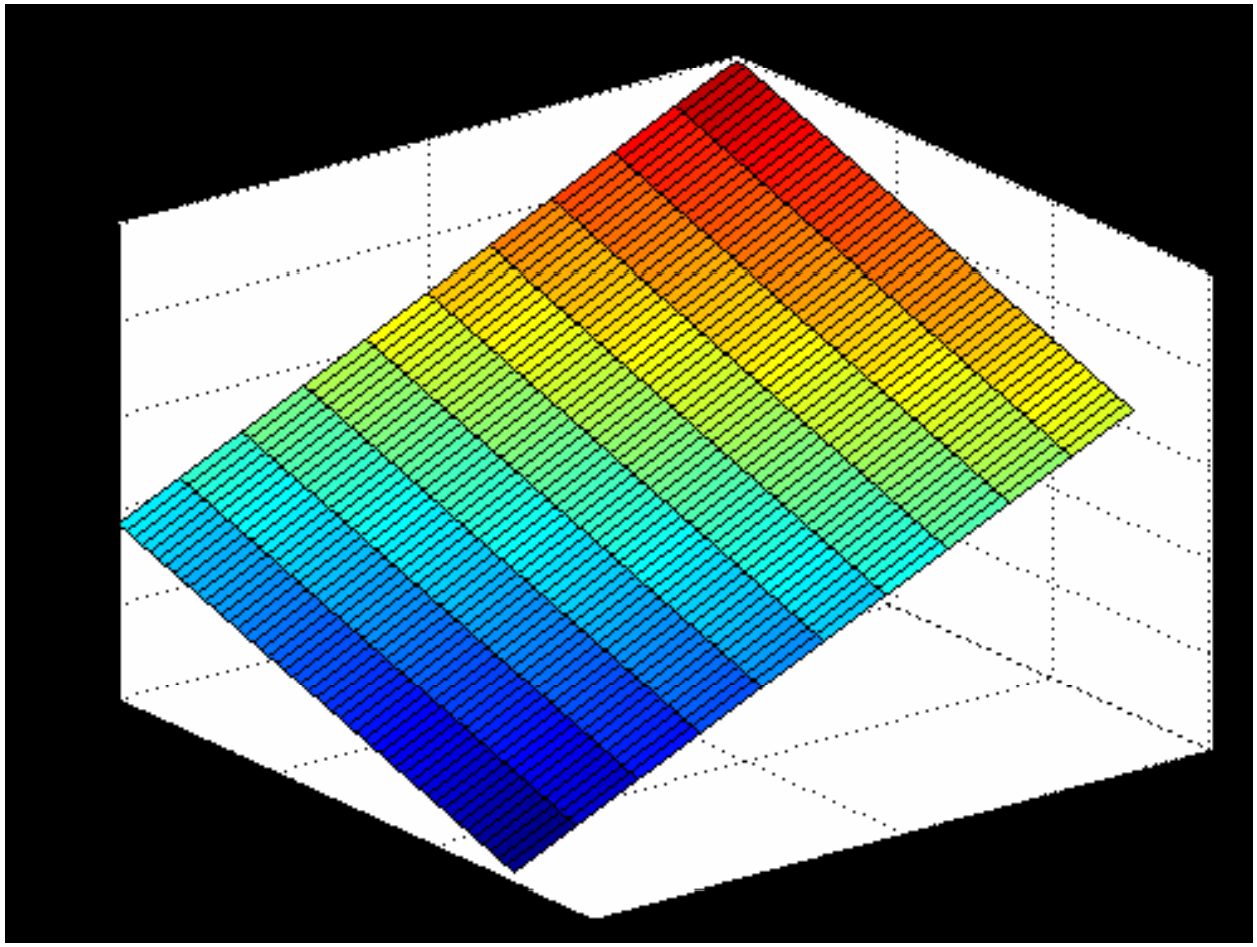
```
» syms x1 x2
»  f=a(1)+a(2)*x1+a(3)*x2
» P1=subs(f,[x1,x2],[1.805,2])
    P1 =
      248.4052
» P2=subs(f,[x1,x2],[1.20,1])
      P2 =
        203.8188
```

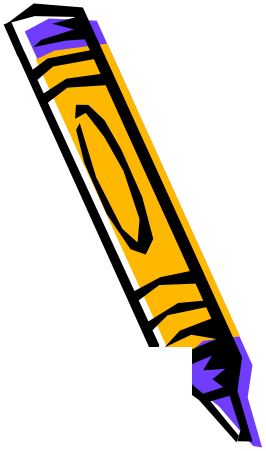




```
» [x1,x2]=meshgrid(1.5:.1:2.5,1:.1:6);
» f=a(1)+a(2)*x1+a(3)*x2;
» surf(x1,x2,f)
```

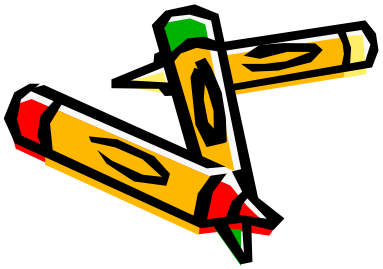






În cazul regresiei liniare multiple apare problema ierarhizării predictorilor în scopul păstrării numai a acelora care, într-adevăr, au o influență semnificativă asupra variabilei *efect*, renunțând astfel la cei nesemnificativi.

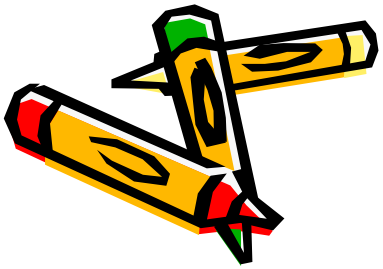
Există două modele importante pentru atingerea acestui scop:



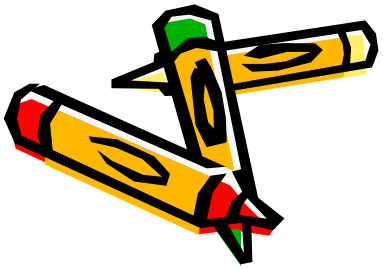
regresia "pas cu pas" directa

1. *regresia „pas cu pas" directă* (*forward stepwise regression*) se bazează pe următorul algoritm simplu:

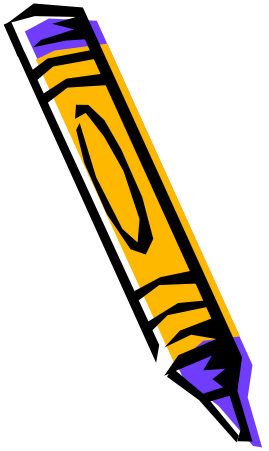
- Se alege predictorul cel mai puternic legat de variabila răspuns. Pentru aceasta se consideră regresia liniară simplă între variabila dependentă şi fiecare predictor în parte, alegându-se acea variabila pentru care nivelul de semnificație *p* este cel mai mic;



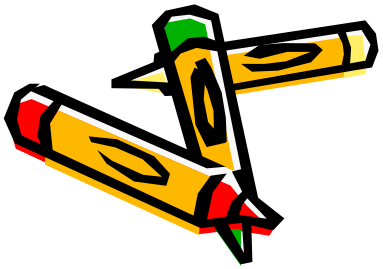
- Se alege, dintre variabilele independente rămase, aceea care are cea mai mare corelație cu *reziduurile* modelului de la pasul anterior;
- Se repetă pasul anterior până când adăugarea unei noi variabile devine nesemnificativă, de exemplu nivelul de semnificație $p = 0.05$ corespunzător corelației cu reziduurile.



regresia "pas cu pas" inversa



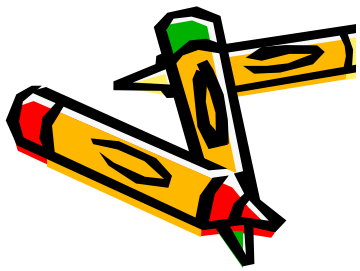
2. *regresia „pas cu pas" inversă* (*backward stepwise regression*). se bazează pe algoritmul invers, adică se pleacă cu modelul complet din care apoi, pe baza nivelului de semnificație *p*, se îndepărtează variabilele neimportante. În principiu, pentru alegerea modelului optim, se consideră ambele modele, după care decizia se ia analizând ambele rezultate.



exemplu

Analiza regresivă multiplă utilizată în predicția *indexului* de rezistență a mușchiului respirator PEmax (exprimat în cm H_2O) în funcție de variabilele predictoare:

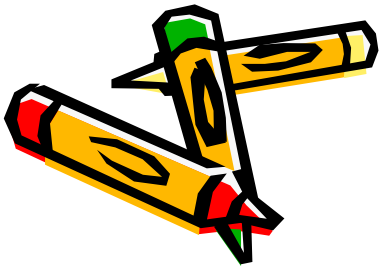
- de înălțime (H -cm),
- greutate (G-kg),
- vârstă (ani),
- sex,
- procentul masei corporale (%, BMP),



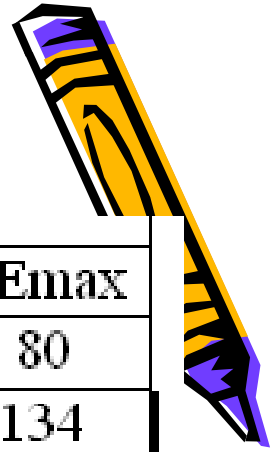
- volumulrespirator forțat per secundă (FEV_1),
- volumul rezidual (RV),
- capacitatea funcțională reziduală (FRC)
- capacitatea totală a plămânului (TLC),

pentru un lot de 25 bolnavii cu fibroză cistică.

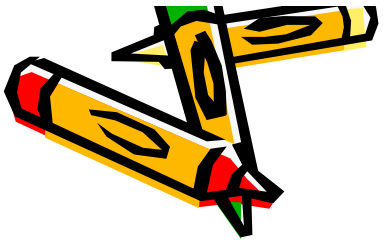
Variabila dependentă a acestui model este reprezentată de indexul de rezistență a mușchiului respirator (PEmax).



Vârstă	Sex	H	G	BMP	FEV_1	RV	FRC	TLC	PEmax
7	0	109	13.1	68	32	258	183	137	95
7	1	112	12.9	65	19	449	245	134	85
8	0	124	14.1	64	22	441	268	147	100
8	1	125	16.2	67	41	234	146	124	85
8	0	127	21.5	93	52	202	131	104	95
9	0	130	17.5	68	44	308	155	118	80
11	1	139	30.7	89	28	305	179	119	65
12	1	150	28.4	69	18	369	198	103	110
12	0	146	25.1	67	24	312	194	128	70
13	1	155	31.5	68	23	413	225	136	95
13	0	156	39.9	89	39	206	142	95	110
14	1	153	42.1	90	26	253	191	121	90
14	0	160	45.6	93	45	174	139	108	100



Vârstă	Sex	H	G	BMP	FEV_1	RV	FRC	TLC	PEmax
15	1	158	51.2	93	45	158	124	90	80
16	1	160	35.9	66	31	302	133	101	134
17	1	153	34.8	70	39	204	118	120	134
17	0	174	44.7	70	49	187	104	103	165
17	1	176	60.1	92	29	188	129	130	120
17	0	171	42.6	69	38	172	130	103	130
19	1	156	37.2	72	21	216	119	81	85
19	0	174	54.6	86	37	184	118	101	85
20	0	178	64.0	86	34	225	148	135	160
23	0	180	73.8	97	57	171	108	98	165
23	0	175	51.5	71	33	224	131	113	95
23	0	179	71.5	95	52	225	127	101	195

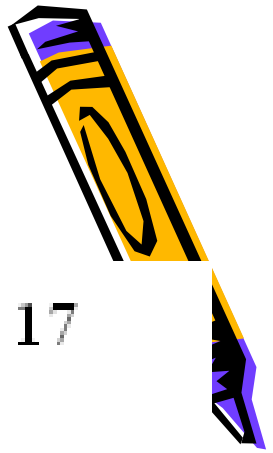


regresia"pas cu pas" directa

Vom aplica regresia „pas cu pas" directă:
vom calcula matricea corelațiilor tuturor variabilelor modelului regresiv, adică matricea coeficienților de corelație.

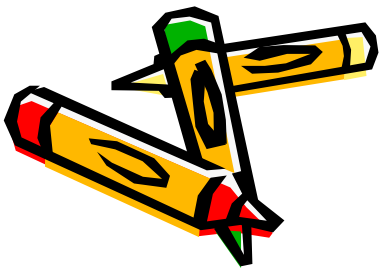
Analiza corelației multiple este necesară pentru a stabili existența legăturii între variabilele analizate.

Menționăm că în cazul de față variabila sex fiind o variabilă aleatoare calitativă (categorială) va fi codată binar: 1 = bărbați și 0 = femei



```
» V=[7 7 8 8 8 9 11 12 12 13 13 14 14 15 16 17 17 17 17
 19 19 20 23 23 23 ];
» S=[0 1 0 1 0 0 1 1 0 1 0 1 0 1 1 1 0 1 0 1 0 0 0 0 0 ];
» H=[109 112 124 125 127 130 139 150 146 155 156 153
 160 158 160 153 174 176 171 156 174 178 180 175 179];
» G=[13.1 12.9 14.1 16.2 21.5 17.5 30.7 28.4 25.1 31.5 39.9
 42.1 45.6 51.7 35.9 34.8 44.7 60.1 42.6 37.2 54.6 64.0 73.8
 51.5 71.5]
» BMP=[68 65 64 67 93 68 89 69 67 68 89 90 93 93 66 70
 70 92 69 72 86 86 97 71 95];
» FEV=[32 19 22 41 52 44 28 18 24 23 39 26 45 45 31 39 49
 29 38 21 37 34 57 33 32];
```

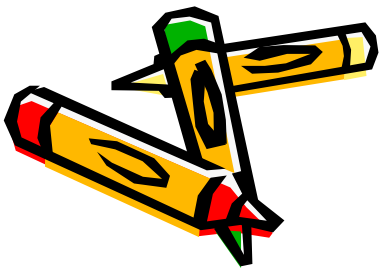
```
» RV=[258 449 441 234 202 308 305 369 312 413 206 253
 174 158 302 204 187 188 172 216 184 225 171 224 225];
» FRC=[183 245 268 146 131 155 179 198 194 225 142 191
 139 124 133 118 104 129 130 119 118 148 108 131 127];
» TLC=[137 134 147 124 104 118 119 103 128 136 95 121
 108 90 101 120 103 130 103 81 101 135 98 113 101];
» PE=[95 85 100 85 95 80 65 110 70 95 110 90 100 80 134
 134 165 120 130 85 85 160 165 95 195];
» A=[V' s' H' G' BMP' FEV' RV' FRC' TLC' PE'];
» corrcoef(A)
```

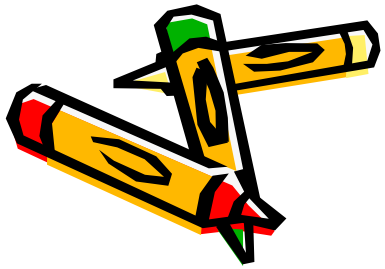


ans =

Columns 1 through 5

1.0000	-0.1671	0.9261	0.9062	0.3778
-0.1671	1.0000	-0.1675	-0.1897	-0.1376
0.9261	-0.1675	1.0000	0.9205	0.4408
0.9062	-0.1897	0.9205	1.0000	0.6724
0.3778	-0.1376	0.4408	0.6724	1.0000
0.1982	-0.4461	0.2389	0.3212	0.4394
-0.5519	0.2714	-0.5695	-0.6222	-0.5824
-0.6394	0.1836	-0.6243	-0.6177	-0.4344
-0.4694	0.0242	-0.4571	-0.4195	-0.3649
0.6135	-0.2886	0.5992	0.6329	0.2295





Columns 6 through 10

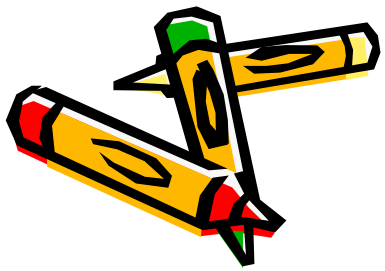
0.1982	-0.5519	-0.6394	-0.4694	0.6135
-0.4461	0.2714	0.1836	0.0242	-0.2886
0.2389	-0.5695	-0.6243	-0.4571	0.5992
0.3212	-0.6222	-0.6177	-0.4195	**0.6329**
0.4394	-0.5824	-0.4344	-0.3649	0.2295
1.0000	-0.6988	-0.6848	-0.3936	0.3061
-0.6988	1.0000	0.9106	0.5891	-0.3156
-0.6848	0.9106	1.0000	0.7044	-0.4172
-0.3936	0.5891	0.7044	1.0000	-0.1816
0.3061	-0.3156	-0.4172	-0.1816	1.0000

Din matricea corelațiilor se observă că:

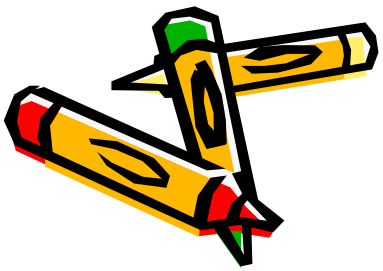
cel mai mare coeficient de corelație este între predictorul *G* și variabila răspuns Pemax. (p = 0.0001).

Calculăm:

- regresia simplă, adică PEmax funcție liniară de *G*
- reziduurile modelului (diferența între valoarea prognozată și cea reală):

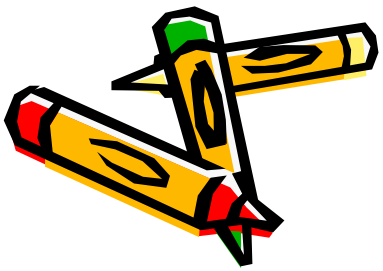


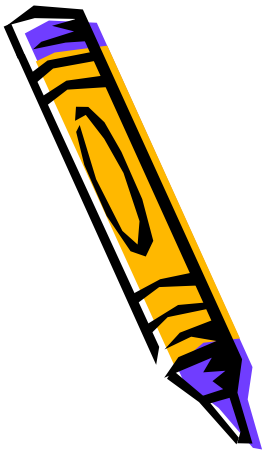
```
» B1=[1 G(1) ;1 G(2) ;1 G(3) ;1 G(4) ;1 G(5) ;1 G(6) ;
   1 G(7) ;1 G(8) ;1 G(9) ;1 G(10) ;1 G(11) ;1 G(12) ;
   1 G(13);1 G(14) ; 1 G(15) ;1 G(16) ;1 G(17) ;1 G(18) ;
   1 G(19) ;1 G(20) ;1 G(21) ;1 G(22) ;1 G(23) ;1 G(24) ;
   1 G(25)];
» b1=B1\PE';
» PE1=b1(1)+b1(2)*G;
» R1=PE-PE1;
```



Alegem, dintre variabilele independente rămase,
aceea care are cea mai mare corelație (ignorând semnul)
cu reziduurile modelului determinat (R1)

```
» A1=[V' s' H'  BMP' FEV' RV' FRC' TLC' R1'];
» corrcoef(A1)
```

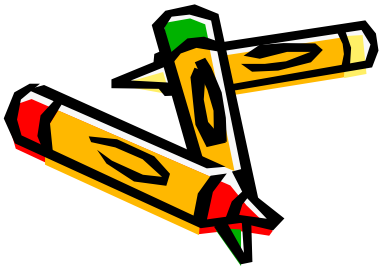


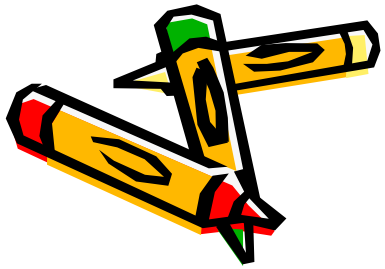


ans =

Columns 1 through 5

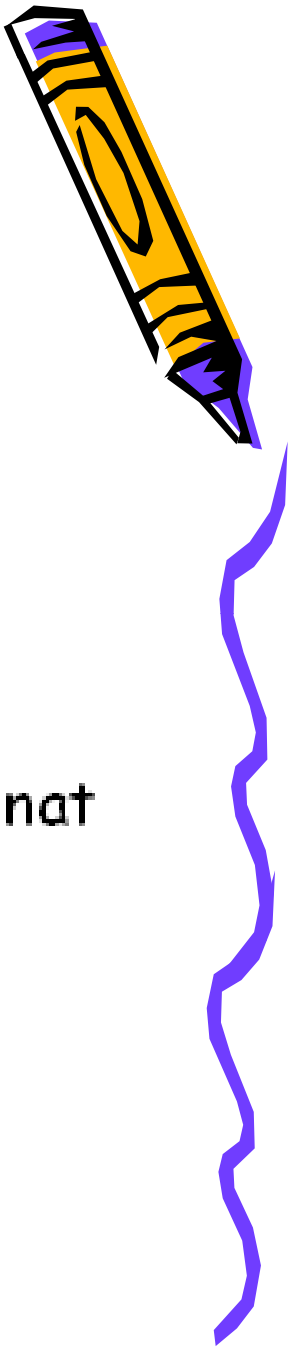
1.0000	-0.1671	0.9261	0.3778	0.1982
-0.1671	1.0000	-0.1675	-0.1376	-0.4461
0.9261	-0.1675	1.0000	0.4408	0.2389
0.3778	-0.1376	0.4408	1.0000	0.4394
0.1982	-0.4461	0.2389	0.4394	1.0000
-0.5519	0.2714	-0.5695	-0.5824	-0.6988
-0.6394	0.1836	-0.6243	-0.4344	-0.6848
-0.4694	0.0242	-0.4571	-0.3649	-0.3936
0.0517	-0.2177	0.0215	-0.2532	0.1328



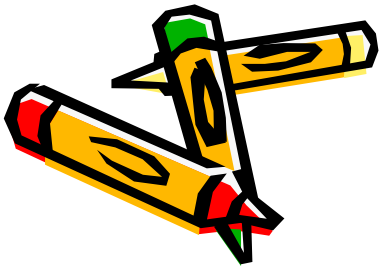


Columns 6 through 9

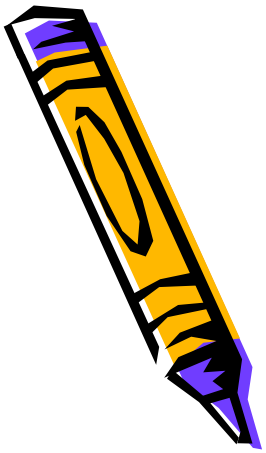
-0.5519	-0.6394	-0.4694	0.0517
0.2714	0.1836	0.0242	-0.2177
-0.5695	-0.6243	-0.4571	0.0215
-0.5824	-0.4344	-0.3649	**-0.2532**
-0.6988	-0.6848	-0.3936	0.1328
1.0000	0.9106	0.5891	0.1010
0.9106	1.0000	0.7044	-0.0340
0.5891	0.7044	1.0000	0.1083
0.1010	-0.0340	0.1083	1.0000



Predictorul BMP are cea mai mare corelație (ignorând semnul) cu reziduurile modelului determinat (R1) (p = 0.0051). Repetăm procedeul:

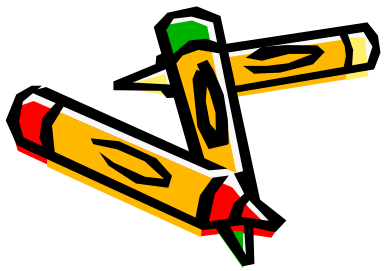


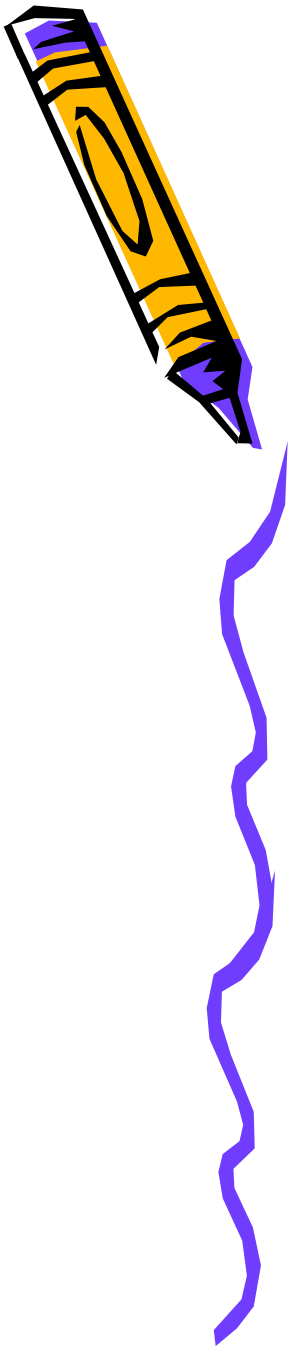
```
»B2=[1 G(1)  BMP(1);1 G(2) BMP(2);1 G(3)  BMP(3);
  1 G(4)  BMP(4);1 G(5)  BMP(5);1 G(6) BMP(6);
  1 G(7)  BMP(7);1 G(8)  BMP(8);1 G(9) BMP(9);
  1 G(10) BMP(10);1 G(11) BMP(11);1 G(12) BMP(12);
  1 G(13) BMP(13);1 G(14) BMP(14) ;1 G(15) BMP(15);
  1 G(16) BMP(16);1 G(17)  BMP(17);1 G(18) BMP(18);
  1 G(19) BMP(19);1 G(20)  BMP(20);1 G(21)  BMP(21);
  1 G(22)  BMP(22);1 G(23) BMP(23);1 G(24)  BMP(24);
  1 G(25) BMP(25)];
» b2=B2\PE';
» PE2=b2(1)+b2(2)*G+b2(3)*BMP;
» R2=PE-PE2;
» A2=[V' s' H'  FEV'  RV' FRC' TLC' R2'];
» corrcoef(A2)
```



Columns 1 through 5

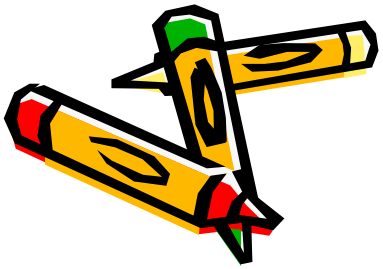
1.0000	-0.1671	0.9261	0.1982	-0.5519
-0.1671	1.0000	-0.1675	-0.4461	0.2714
0.9261	-0.1675	1.0000	0.2389	-0.5695
0.1982	-0.4461	0.2389	1.0000	-0.6988
-0.5519	0.2714	-0.5695	-0.6988	1.0000
-0.6394	0.1836	-0.6243	-0.6848	0.9106
-0.4694	0.0242	-0.4571	-0.3936	0.5891
-0.0589	-0.2366	-0.0647	0.2511	0.0269

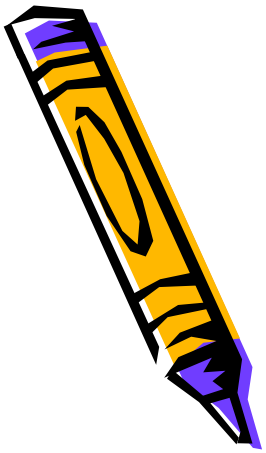




Columns 6 through 8

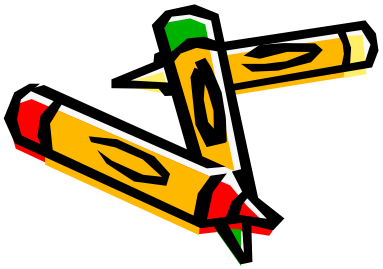
-0.6394	-0.4694	-0.0589
0.1836	0.0242	-0.2366
-0.6243	-0.4571	-0.0647
-0.6848	-0.3936	**0.2511**
0.9106	0.5891	0.0269
1.0000	0.7044	-0.0455
0.7044	1.0000	0.0746
-0.0455	0.0746	1.0000

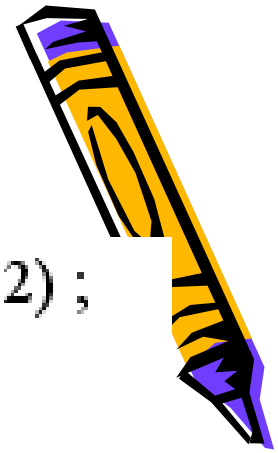




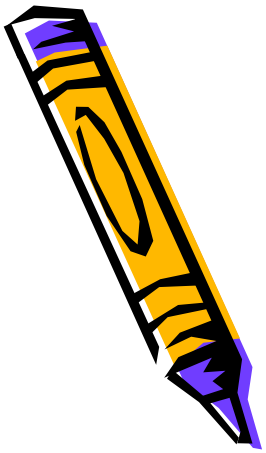
Predictorul FEV are cea mai mare corelație (ignorând semnul) cu reziduurile modelului determinat (R2) ($p = 0.0210$).

Repetăm procedeul:

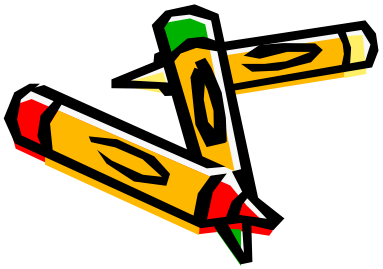




```
»B3=[1 G(1) BMP(1) FEV(1) ;1 G(2) BMP(2) FEV(2) ;
  1 G(3)  BMP(3) FEV(3) ;1 G(4) BMP(4) FEV(4) ;
  1 G(5) BMP(5) FEV(5) ;1 G(6) BMP(6) FEV(6) ;
  1 G(7) BMP(7) FEV(7) ;1 G(8) BMP(8) FEV(8) ;
  1 G(9)  BMP(9) FEV(9);1 G(10) BMP(10) FEV(10) ;
  1 G(11) BMP(11) FEV(11) ;1 G(12) BMP(12) FEV(12) ;
  1 G(13) BMP(13) FEV(13) ;1 G(14)  BMP(14) FEV(14) ;
  1 G(15)  BMP(15) FEV(15);1 G(16)  BMP(16) FEV(16);
  1 G(17)  BMP(17) FEV(17);1 G(18)  BMP(18) FEV(18);
  1 G(19)  BMP(19) FEV(19);1 G(20) BMP(20) FEV(20);
  1 G(21) BMP(21) FEV(21);1 G(22) BMP(22) FEV(22);
  1 G(23) BMP(23) FEV(23);1 G(24) BMP(24) FEV(24);
  1 G(25) BMP(25) FEV(25) ];
```

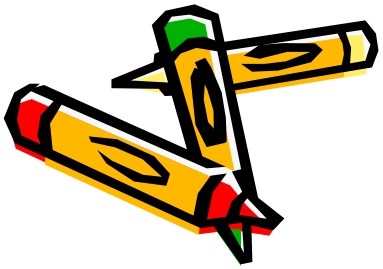


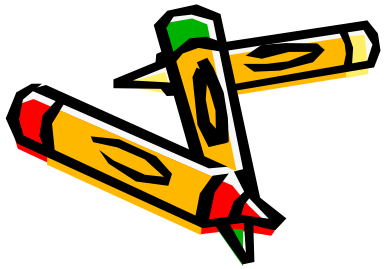
```
» b3=B3\PE';
» PE3=b3(1)+b3(2)*G+b3(3)*BMP+b3(4)*FEV;
» R3=PE-PE3;
»A3=[V' s' H'  RV' FRC' TLC' R3'];
» corrcoef(A3)
```

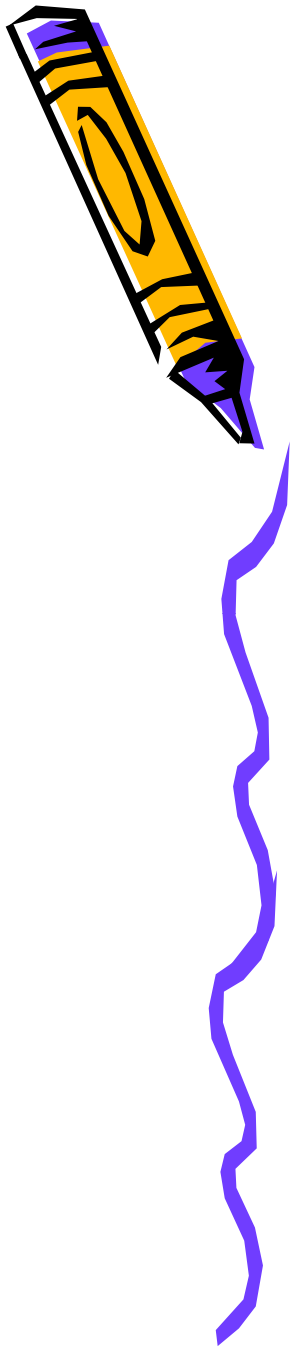


Columns 1 through 5

1.0000	-0.1671	0.9261	-0.5519	-0.6394
-0.1671	1.0000	-0.1675	0.2714	0.1836
0.9261	-0.1675	1.0000	-0.5695	-0.6243
-0.5519	0.2714	-0.5695	1.0000	0.9106
-0.6394	0.1836	-0.6243	0.9106	1.0000
-0.4694	0.0242	-0.4571	0.5891	0.7044
-0.0618	-0.1227	-0.0726	0.1683	0.1080







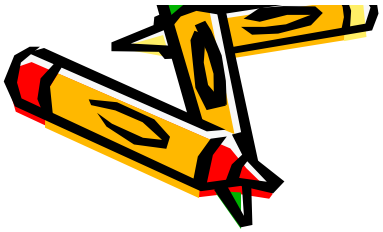
Columns 6 through 7

-0.4694	-0.0618
0.0242	-0.1227
-0.4571	-0.0726
0.5891	**0.1683**
0.7044	0.1080
1.0000	0.1508
0.1508	1.0000

Predictorul RV are cea mai mare corelație (ignorând semnul) cu reziduurile modelului determinat (R3), dar cum nivelul de semnificație este $p = 0.0736 > 0.05$, nu poate face parte din model

Astfel:

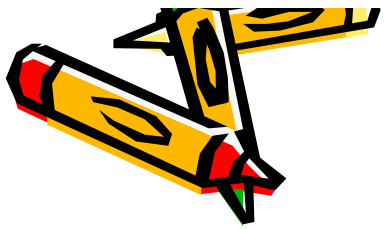
PEmax= 120.8813 + 1.6093 G – 1.2540 BMP+0.7150 FEV

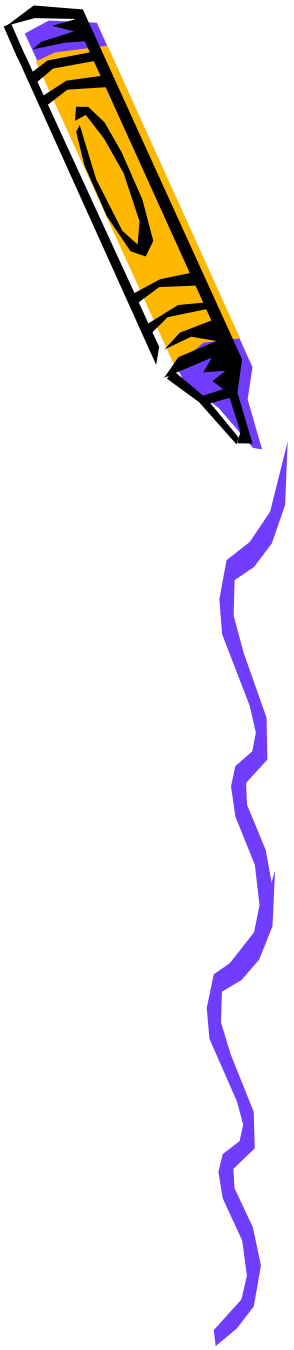


"regresia pas cu pas" posterioara

Datele utilizate la construcția modelului *„regresia pas cu pas posterioară"*, sunt prezentate în tabelul de mai jos.

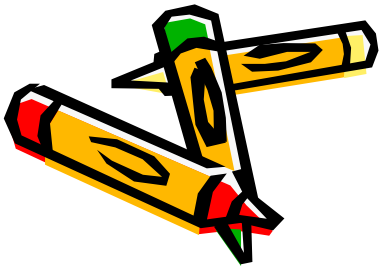
	estimare	eroare standard (b)	t-test	p
Interceptor (b_0)	63.54564	12.70163	5.002952	0.000046
G	1.18671	0.30086	3.944453	0.000646

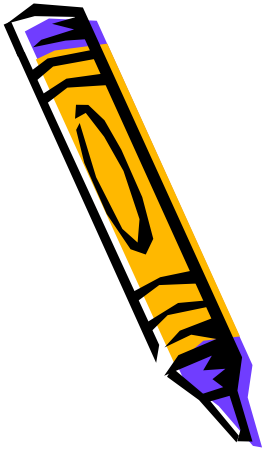




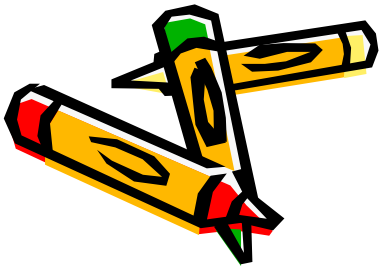
Ecuația de regresie este:

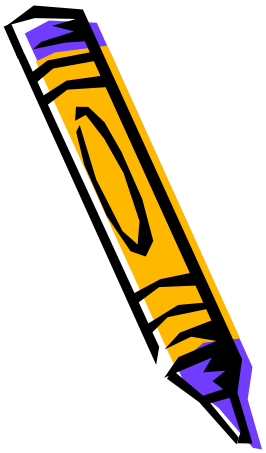
$$\mathrm{PEmax} = 63.55 + 1.19 \cdot G.$$





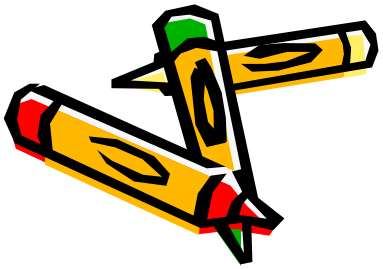
În principiu, nici una dintre cele două variante de regresie multiplă de mai sus nu este cea ideală. Dacă vrem cel mai ‚larg' model, îl alegem pe cel ‚anterior', iar dacă-l dorim pe cel mai ‚strict', îl alegem pe cel ‚posterior'.





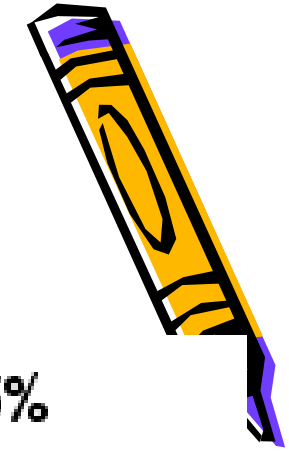
În final, vom aplica metoda regresiei multiple standard.

Calculele efective sunt complexe, pentru aceasta utilizându-se programele specializate.

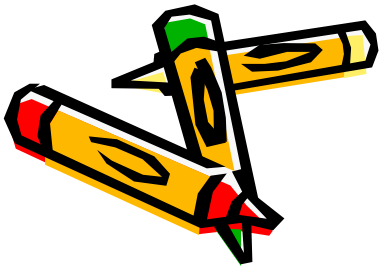


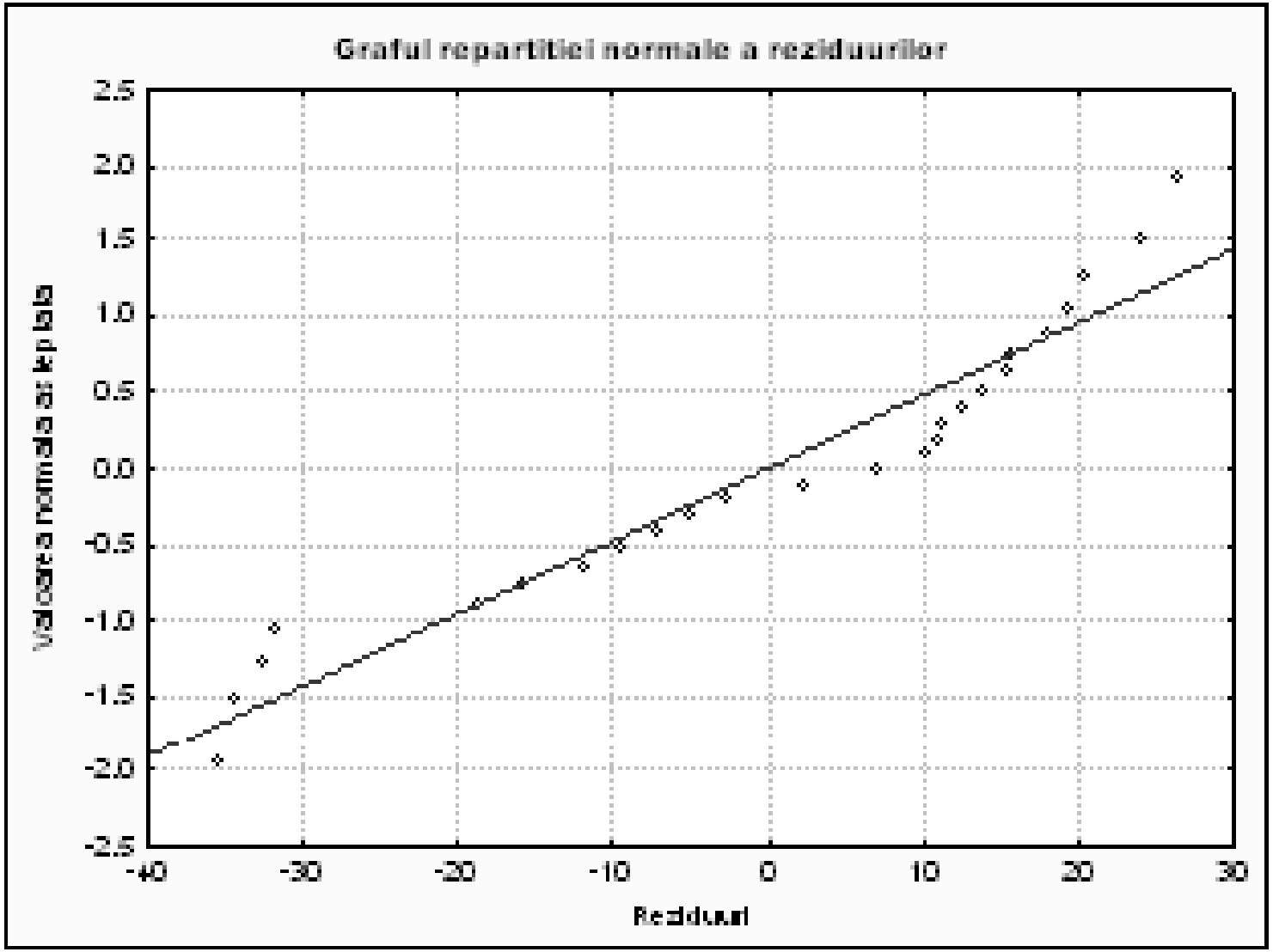
vom lua în considerație toate cele nouă variabile explicative. Rezultatul aplicării acestei metode este tabelul următor.

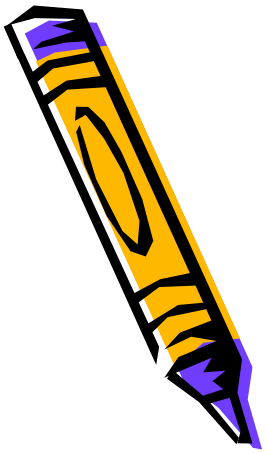
	estimare	eroare standard (b)	t-test (15)	p
interceptor (b_0)	137.229	207.946	0.659	0.519
Vârsta (V)	-2.475	4.367	-0.566	0.579
Sex (S)	-1.388	13.597	-0.102	0.920
H	-0.308	0.864	-0.356	0.726
G	2.878	1.846	1.559	0.139
BMP	-1.797	1.115	-1.610	0.128
Fev1	1.494	0.970	1.539	0.144
RV	0.178	0.186	0.953	0.355
FRC	-0.163	0.475	-0.344	0.735
TLC	0.114	0.477	0.239	0.814



Nicio variabilă predictivă nu trece pragul de 5%
al nivelului de semnificație *p*.
Acesta nu este un criteriu absolut de acceptare a modelului, așa că vom lua în considerație și prognoza astfel obținută, mai ales că reprezentarea grafică a reziduurilor indică un grad de acceptabil de acuratețe a modelului.

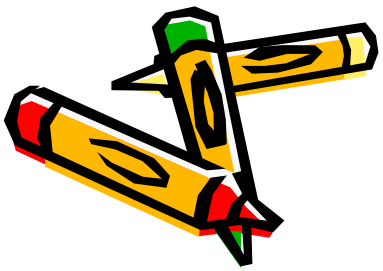


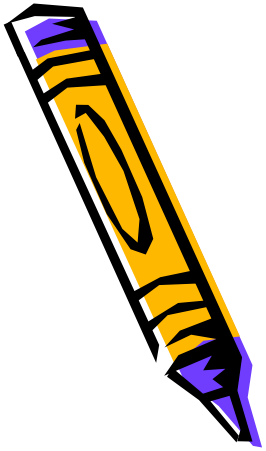

Graful repartitiei normale a reziduurilor
Valoarea normala asteptata
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-40
-30
-20
-10
0
10
20
30
Residual



Ecuația de regresie corespunzătoare este dată de formula:

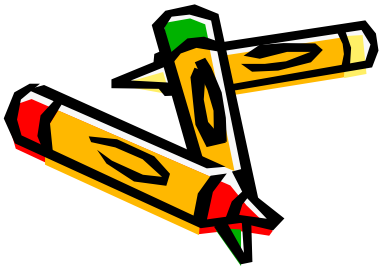
$$PE = 137.23 - 1.79\ BMP + 2.87\ G + 1.49\ FEV_1 + 0.17\ RV - 2.47\ V - 0.3\ H - 0.16\ FRC + 0.11\ TLC - 1.38\ S$$

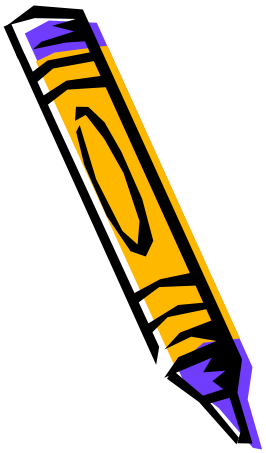




Utilizăm ecuația de regresie multiplă pentru a obține valorile variabilei dependente pentru orice valori individuale ale variabilelor explicative.

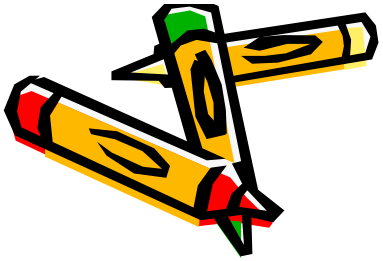
În acest mod, pentru un anumit obiect cu atributele predictive cunoscute, se deduce valoarea atributului necunoscut, considerat ca atribut răspuns (*outcome*).





În cazul de mai sus, pentru un anumit pacient căruia i se cunosc valorile celor nouă parametri medicali predictivi, i se poate prognoza, cu o acuratețe suficientă, valoarea PEmax, prin introducerea în ecuația de regresie a valorilor sale individuale.

Spunem că, astfel, se obține o *valoare prognostic* (*index prognostic*), pe baza datelor cunoscute.



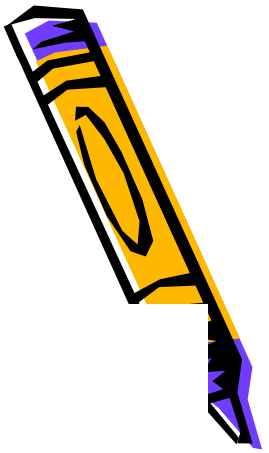
exemple

Prezentăm un exemplu fictiv de aplicare a metodei regresiei liniare multiple în oncologie.
Să presupunem că dispunem de date referitoare la un lot de 101 pacienți suferind de un anumit tip de cancer, date ce reprezintă diametrului mediu al tumorii (în mm), vârsta (în ani), stadiul bolii (I, II, III, IV) și timpul de supraviețuire (în luni) din momentul diagnosticării.

Pentru a putea aplica metoda regresiei, vom coda fiecare pacient ca un vector cu patru elemente, $\mathbf{x} = (x_1, x_2, x_3, y)$.

Din punct de vedere probabilistic, x_i reprezintă o valoare a variabilei aleatoare X_i, unde

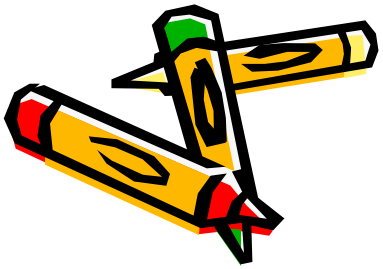
- X_1 reprezintă variabila aleatoare corespunzătoare diametrului mediu al tumorii,
- X_2 reprezintă variabila aleatoare corespunzătoare vârstei,
- X_3 reprezintă variabila aleatoare corespunzătoare stadiului bolii,
- Y reprezintă variabila aleatoare corespunzătoare timpului de supraviețuire.



Am simulat pe computer repartițiile acestor variabile, folosind generatorul de numere (pseudo-) aleatoare, urmând următoarea schemă:

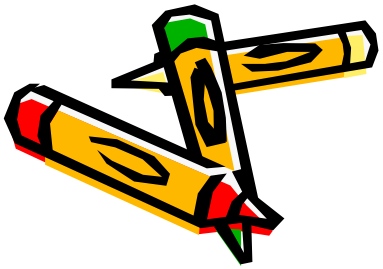
- variabila X_1 -gaussiană de medie 12 și deviație standard 3,
- variabila X_2 -gaussiană de medie 45 și deviație standard 12,
- variabila X_3 -discretă

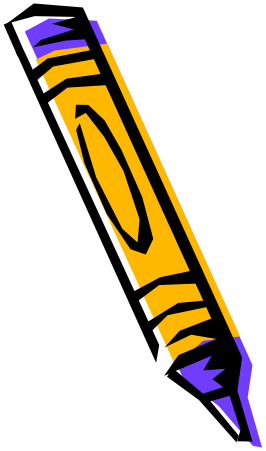
  (valori/probabilități: $\mathrm{I} = 0{,}35$, $\mathrm{II} = 0{,}3$, $\mathrm{III} = 0{,}25$, $\mathrm{IV} = 0{,}1$),
- variabila Y -gaussiană de medie 36 și deviație standard 12.



Scopul analizei regresive efectuate:
de a stabili legătura între variabilele predictive X_1, X_2 şi X_3
şi variabila dependentă Y.

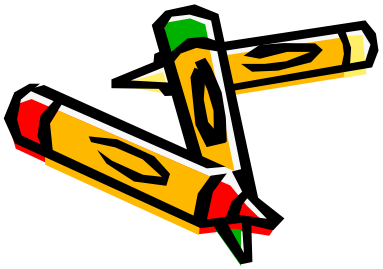
Concret, analizăm legătura între factorii predictivi: diametrul mediu al tumorii, vârstă şi stadiu şi variabila dependentă reprezentând timpul de supraviețuire.



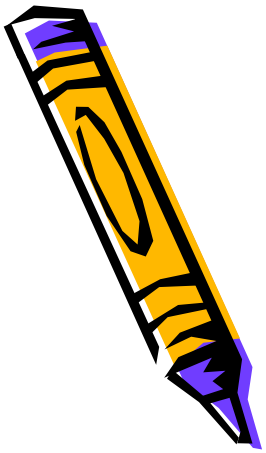


Scopul final: de a putea prognoza timpul de supraviețuire al unui pacient oarecare, care este similar din punct de vedere medical celor din lotul analizat.

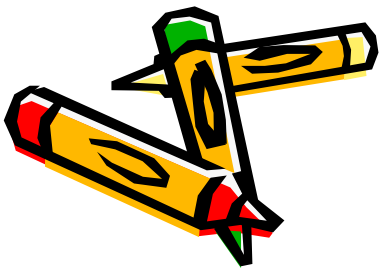
Se dorește extrapolarea rezultatelor obținute pe acest lot la o întreagă populație (inferență statistică), în ipoteza că lotul este reprezentativ pentru aceasta.



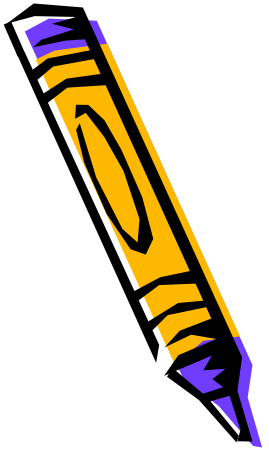
rezultatele analizei regresive



	Beta standardizat	Coeficient b	Nivel semnificaţie p
Interceptor		24.425	0.0001 8
Diametru mediu	0.045	0.046	0.0647
Vârstă	0.226	0.151	0.0571
Stadiu	0.070	0.828	0.0385

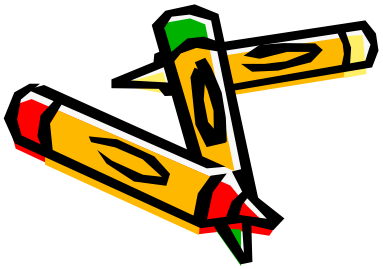


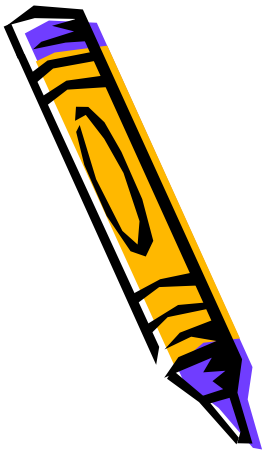
ierarhia predictorilor



ierarhia predictorilor din punctul de vedere al importanței lor în influențarea timpului de supraviețuire este următoarea:

- I stadiul bolii,
- II vârstă
- III diametrul mediu al tumorii.

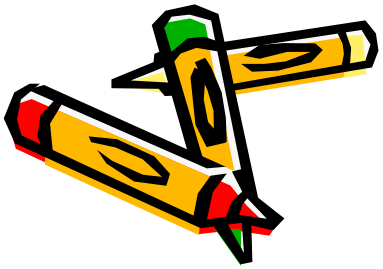




De reținut:

- semnificativ este aici doar stadiul, având un nivel de semnificație $p < 0.05$,
- vârsta este aproape semnificativă,
- iar diametrul se poate considera puțin semnificativ

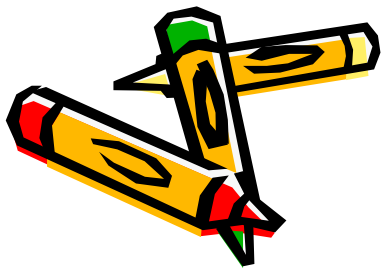
în influențarea timpului de supraviețuire.



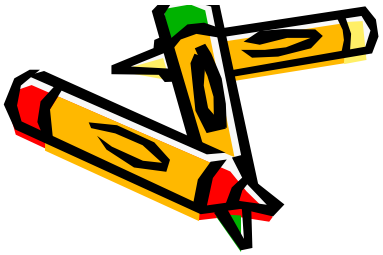
Ecuația de regresie liniară multiplă este dată de:

$$\textit{Timp supraviețuire} = 24.43 + 0.046\ \textit{Diametru} + 0.15\ \textit{Vârstă} + \\ + 0.83\ \textit{Stadiu}.$$

Plecând de la această ecuație, se poate prognoza timpul de supraviețuire al unui nou pacient, care nu face parte din lot dar este similar caracteristicilor lotului.

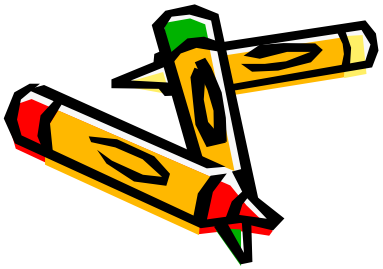


De exemplu, dacă se prezintă un pacient care are diametrul mediu al tumorii de 15 mm, vârsta de 42 ani şi este în stadiul III şi este diagnosticat cu acest tip de cancer, atunci introducând aceste date personale în ecuație obținem timpul de supraviețuire estimat la 34 luni, cu un interval de încredere de 95%, dat de $(31, 37)$, adică timpul de supraviețuire poate lua orice valoare din acest interval cu un grad de încredere de 95%.



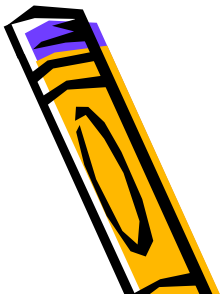
Prezentăm, în final, un exemplu de aplicare a regresiei multiple în domeniul imobiliar, la prognozarea prețului unei case.

Datele se referă la un număr de 76 case din South Eugene, Oregon, USA, din anul 2005 (Pardoe)



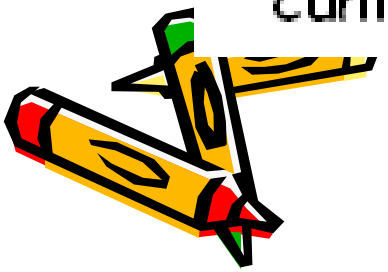
. Variabilele predictive în acest caz sunt următoarele:

- X_1 - suprafața casei (mii de m^2);
- X_2 - categoria mărimii lotului (de la 1 la 11);
- X_3 - numărul de băi;
- X_4 - numărul de dormitoare (între 2 și 6);
- X_5 - vechimea;
- X_6 - mărimea garajului (0, 1, 2 sau 3 mașini);
- D_7 - indicator pentru „listare activă" (așteptare/vândut);



- D_8 - indicator pentru Edison elementar;
- D_9 - indicator pentru Harris elementar;
- D_{10} - indicator pentru Adams elementar;
- D_{11} - indicator pentru Crest elementar;
- D_{12} - indicator pentru Parker elementar

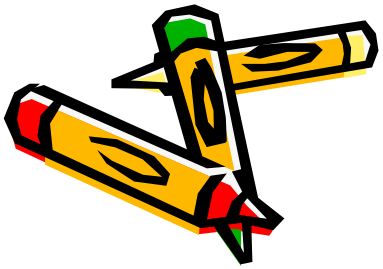
Ultimii indicatori, D_8 - D_{12} reprezintă indicatorii unor şcoli elementare din vecinătatea locuinţelor (localitatea Edgewood), care se iau în considerare la cumpărarea unei case.

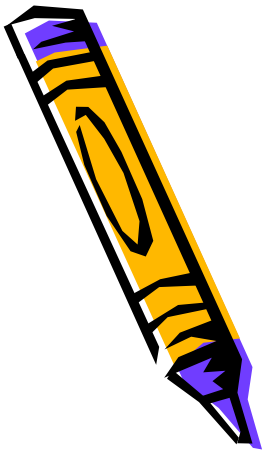


Primul model regresiv propus este de forma:

$$Y = b_0 + b_1 \cdot X_1 + \ldots + b_6 \cdot X_6 + b_7 \cdot D_7 + \ldots + b_{12} \cdot D_{12}$$

Deoarece analiza reziduurilor nu a produs o repartiție suficient de normală a acestora, modelul nu este acceptabil pentru prognoză.

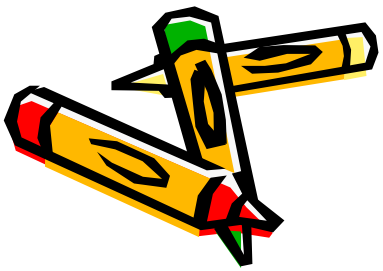




O altă abordare, propune modelul neliniar de ecuație:

$$Y = b_0 + b_1 X_1 + \ldots + b_4 X_4 + b_5 X_3 X_4 + b_6 X_5 + b_7 X_5^2 + b_9 D_7 +$$
$$+ b_{10} D_8 + b_{11} D_9 + b_{12} D_{10} + b_{13} D_{11} + b_{14} D_{12}$$

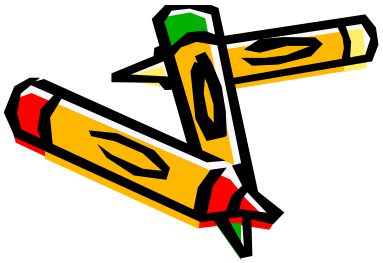
care este validat din punctul de vedere al analizei reziduurilor.



Ultimul model propus, neliniar şi acesta, de ecuație

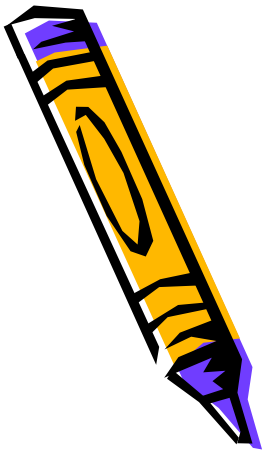
$$Y = b_0 + b_1 X_1 + \ldots + b_4 X_4 + b_5 X_3 X_4 + b_6 X_5 + b_7 X_5^2 + b_9 D_7 + \\ + b_{10} D_8 + b_{11} D_9$$

se dovedeşte cel mai performant.



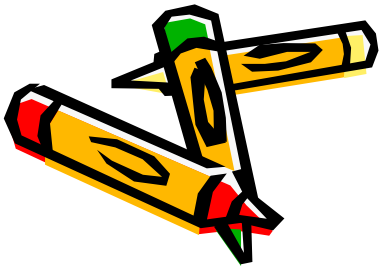
principalele caracteristici ale modelului 3

	Estimare	Eroarea standard se(b)	Nivelul de semnificație p
Constantă	332.478	106.599	0.003
X_1	56.719	27.974	0.047
X_2	9.917	3.438	0.005
X_3	-98.156	42.666	0.025
X_4	-78.910	27.752	0.006
X_3 X_4	30.390	11.878	0.013
X_5	3.301	3.169	0.302
X_5^2	1.641	0.733	0.029
X_6	13.119	8.285	0.118
D_7	27.424	10.988	0.015
D_8	67.062	16.822	0.000
D_9	47.273	14.844	0.002

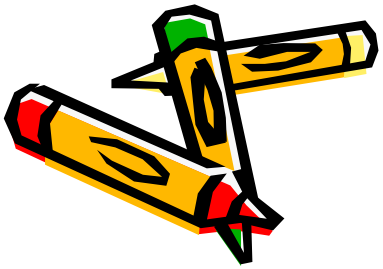


Rezultă că ecuația de regresie (neliniară) corespunzătoare este dată de:

$$Y = 332.48 + 56.72X_1 + 9.92X_2 - 98.16X_3 - 78.91X_4 +$$
$$+ 30.39X_3X_4 + 3.30X_5 + 27.42D_7 + 67.06D_8 + 47.27D_9$$



În concluzie, un cumpărător oarecare poate să completeze variabilele predictive ale modelului cu datele sale concrete, obținând astfel o prognoză a prețului pe care ar trebui să-l plătească ca să dobândească o locuință în acea zonă.



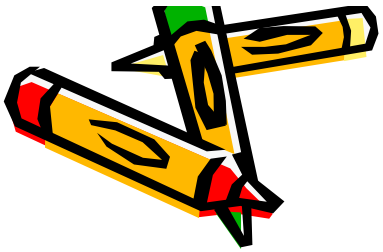
IMPORTANT

- Se va evita un model provenind dintr-un eşantion mic dar având multe variabile predictive.

  Ca regulă generală, numărul de predictori nu trebuie să depăşească valoarea *n*/10, unde *n* este volumul eşantionului.

- Alegerea automată a modelului pe baza unui program statistic adecvat este normală, dar nu trebuie ignorat bunul simţ al practicianului în evaluarea şi validarea finală a modelului.

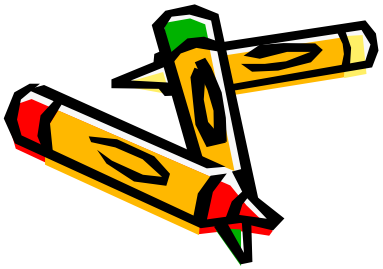
- Variabilele explicative, puternic corelate între ele, vor fi de așa natură selectate, încât să fie inclus în model doar un ‚reprezentant' al lor și nu toate, pentru evitarea redundanței.
- Pentru mai multă siguranță, se verifică capacitatea modelului pe alt eșantion, dacă acest lucru este posibil.

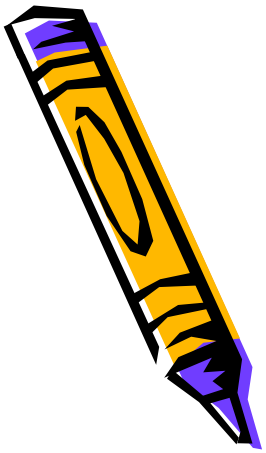


studiu de caz cu MATLAB

Aproximarea unui set de date $\{(x_i, y_i), 1 \le i \le n\}$ printr-o funcție polinomială, respectiv mixt-exponențială, (funcțiile f_1 și f_2) folosind metoda celor mai mici pătrate.

Este posibil ca pe curba aproximantă să nu se găsească nici un punct al setului de date, lucru ce deosebește net aproximarea de procesul de interpolare, caz în care toate punctele sunt situate pe curbă.

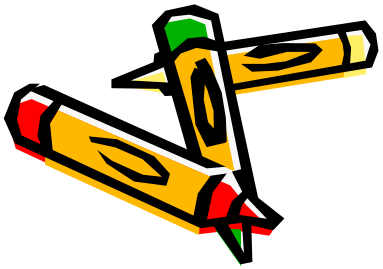




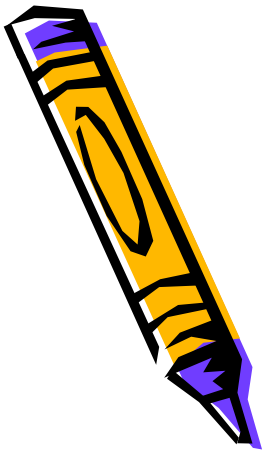
Funcțiile folosite sunt:

polyfit - funcție ce returnează coeficienții polinomului;

polyval - funcție ce returnează valorile polinomului în punctele $\{x_i, 1 \le i \le n\}$.



exemplu



Considerăm situația numărului de pacienți internați într-o secție a unui spital pe parcursul unui an, situație reflectată în următorul tabel, unde am codificat lunile anului în sensul:

Ianuarie = 1;

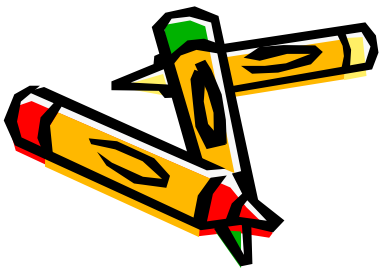
Februarie = 2;

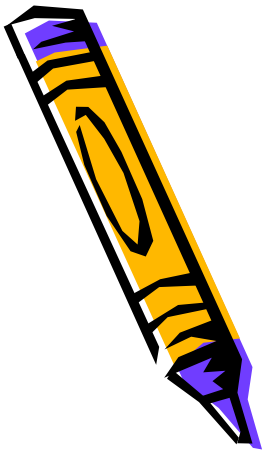
.....

Decembrie = 12

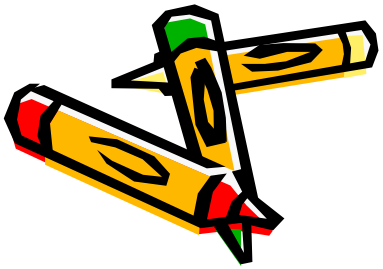
1	2	3	4	5	6	7	8	9	10	11	12
140	110	120	114	105	90	70	50	92	116	125	105

```
» X=1:12;Y=[140 110 120 114 105 90 70 50 92 116 125 105];
» polyfit(X,Y,4)
   ans =
      -0.0869    2.3396  -19.4963   48.6599   96.3081
```



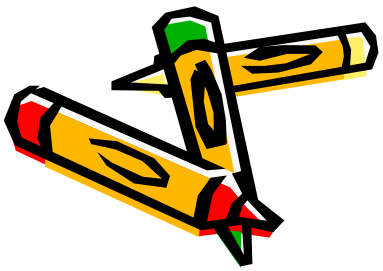


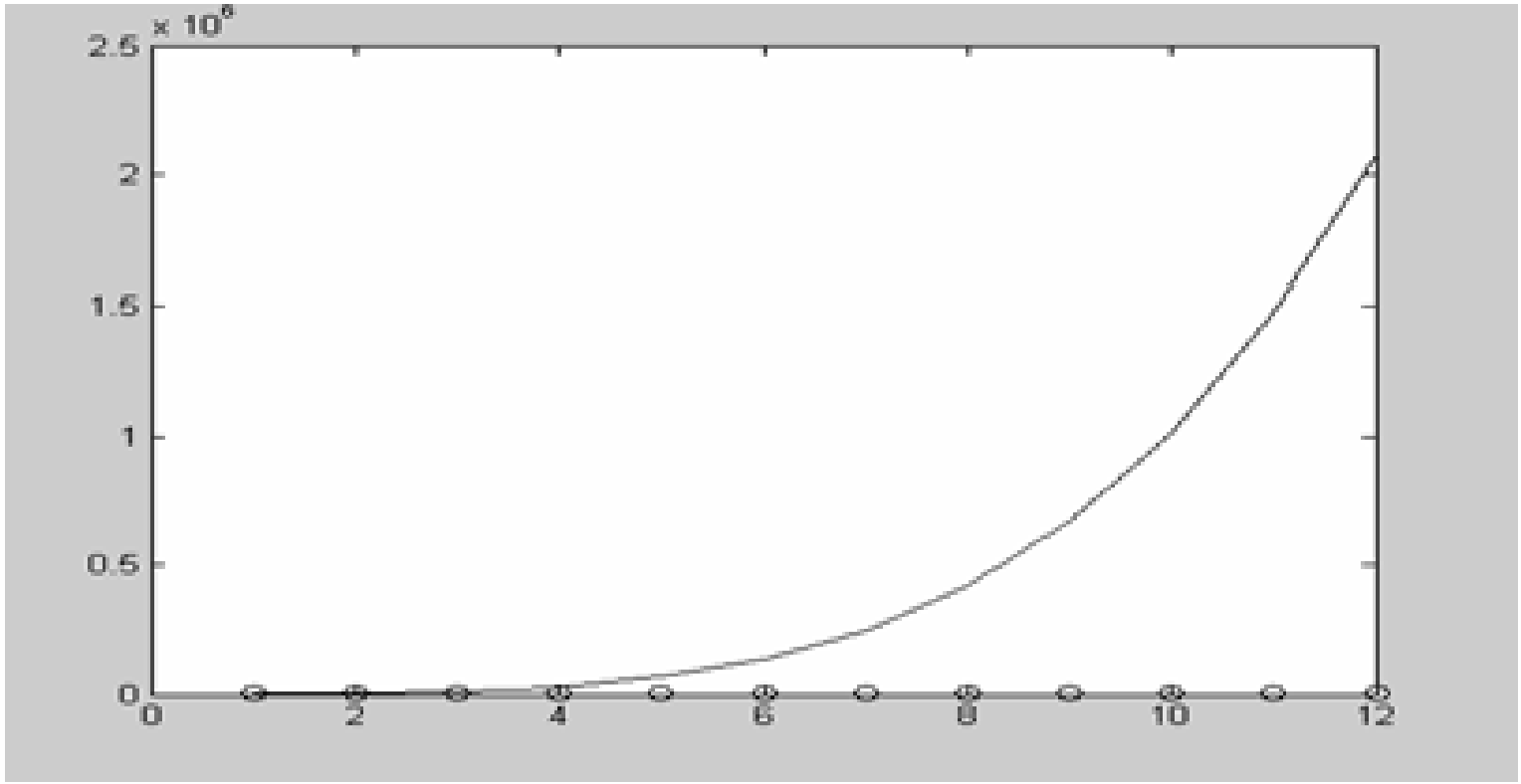
```
» pol = polyval(p,X,Y)
      pol =
    Columns 1 through 5
    127.7244  132.9691  122.9516  106.4955   90.3388
    Columns 6 through 10
     79.1336   75.4468   79.7591   90.4660  103.8770
    Columns 11 through 12
    114.2162  113.6218
```



Să desenăm în acelaşi sistem de axe setul de date şi polinomul de gradul 4 ce-l aproximează:

```
» y1=p(1)+p(2)*X+p(3)*X.^2+p(4)*X.^3+p(5)*X.^4;
» plot(X,y1,X,Y,'O')
```

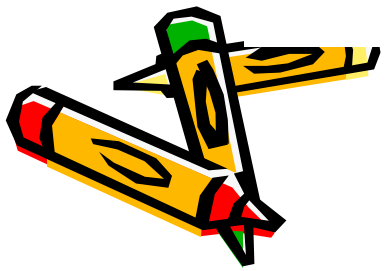


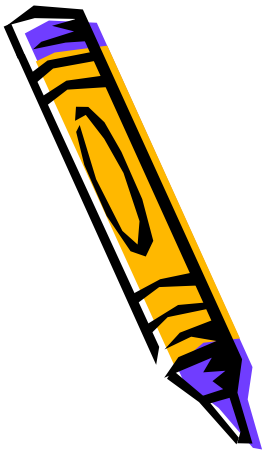

x 10^6
2.5
2
1.5
1
0.5
0
0
2
4
6
8
10
12

Rezultatul nu este mulțumitor, datorită valorilor mari ale ordonatelor.

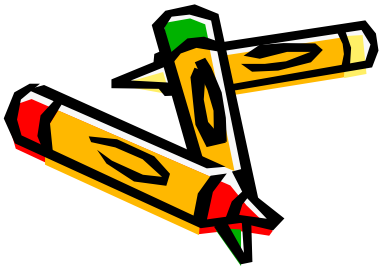
În acest caz este necesară normalizarea datelor.

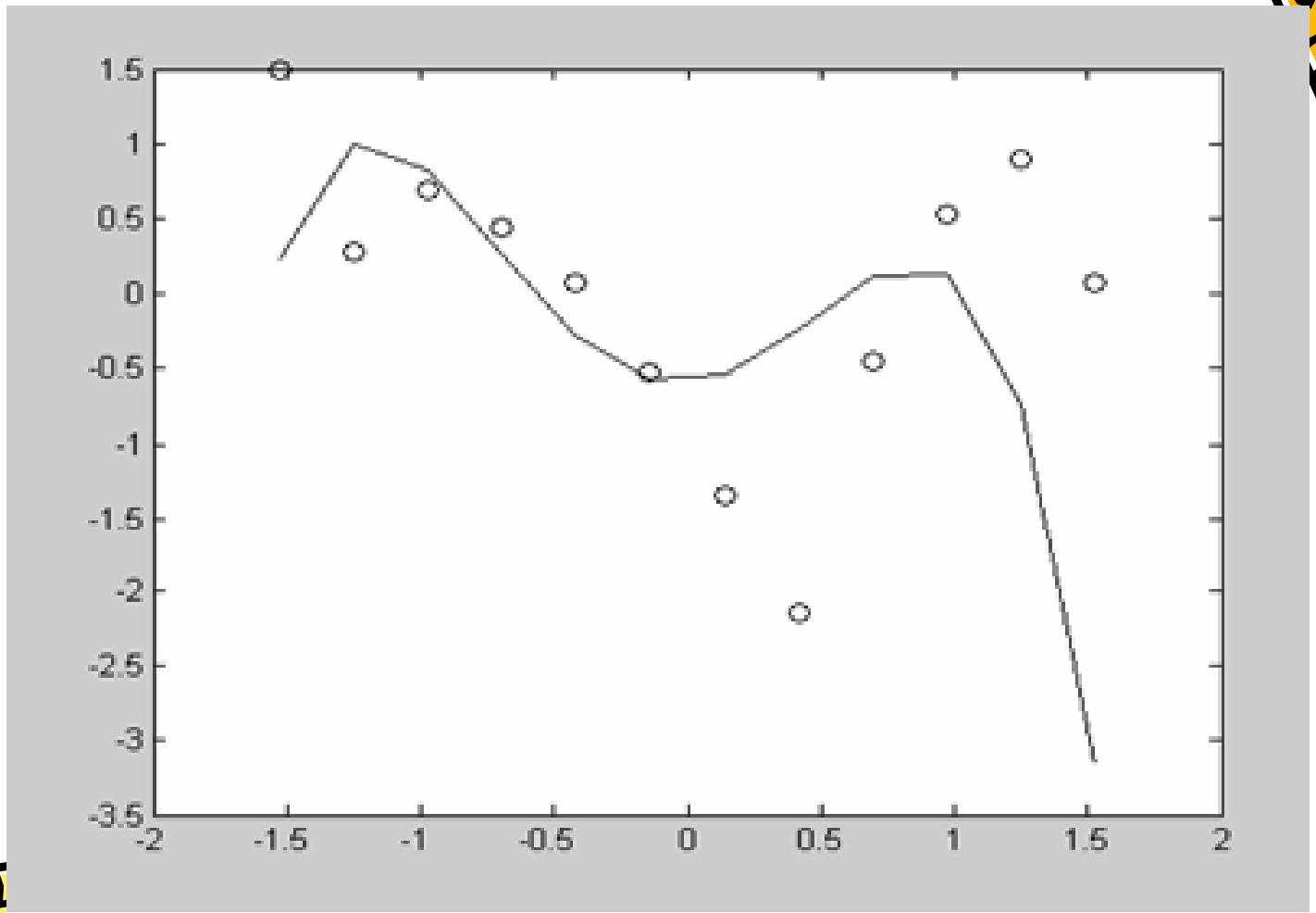
```
» Xs=(X-mean(X))./std(X)
    Xs =
      Columns 1 through 5
   -1.5254   -1.2481   -0.9707   -0.6934   -0.4160
      Columns 6 through 10
    -0.1387    0.1387    0.4160    0.6934    0.9707
     Columns 11 through 12
     1.2481    1.5254
```

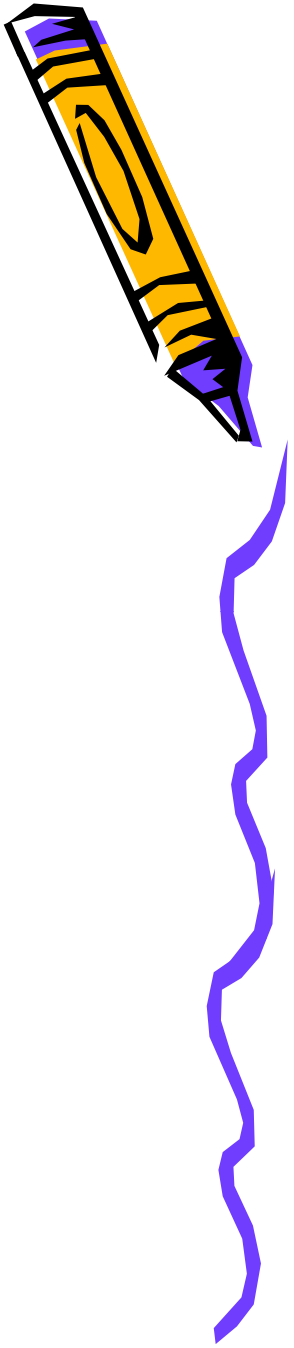




```
» Ys=(Y-mean(Y))./std(Y):
» ps=polyfit(Xs,Ys,4)
    ps =
       -0.5961    0.1525    2.1616   -0.5425   -1.0883
»y2=ps(1)+ps(2)*Xs+ps(3)*Xs.^2+ps(4)*Xs.^3+ps(5)*Xs.^4;
» plot(Xs,y2,'k',Xs,Ys,'ko')
```

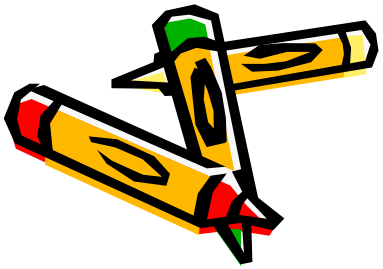



1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

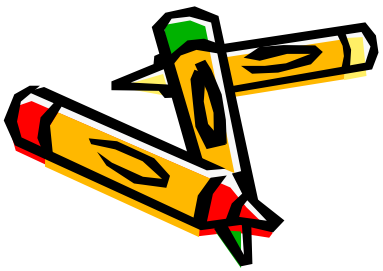


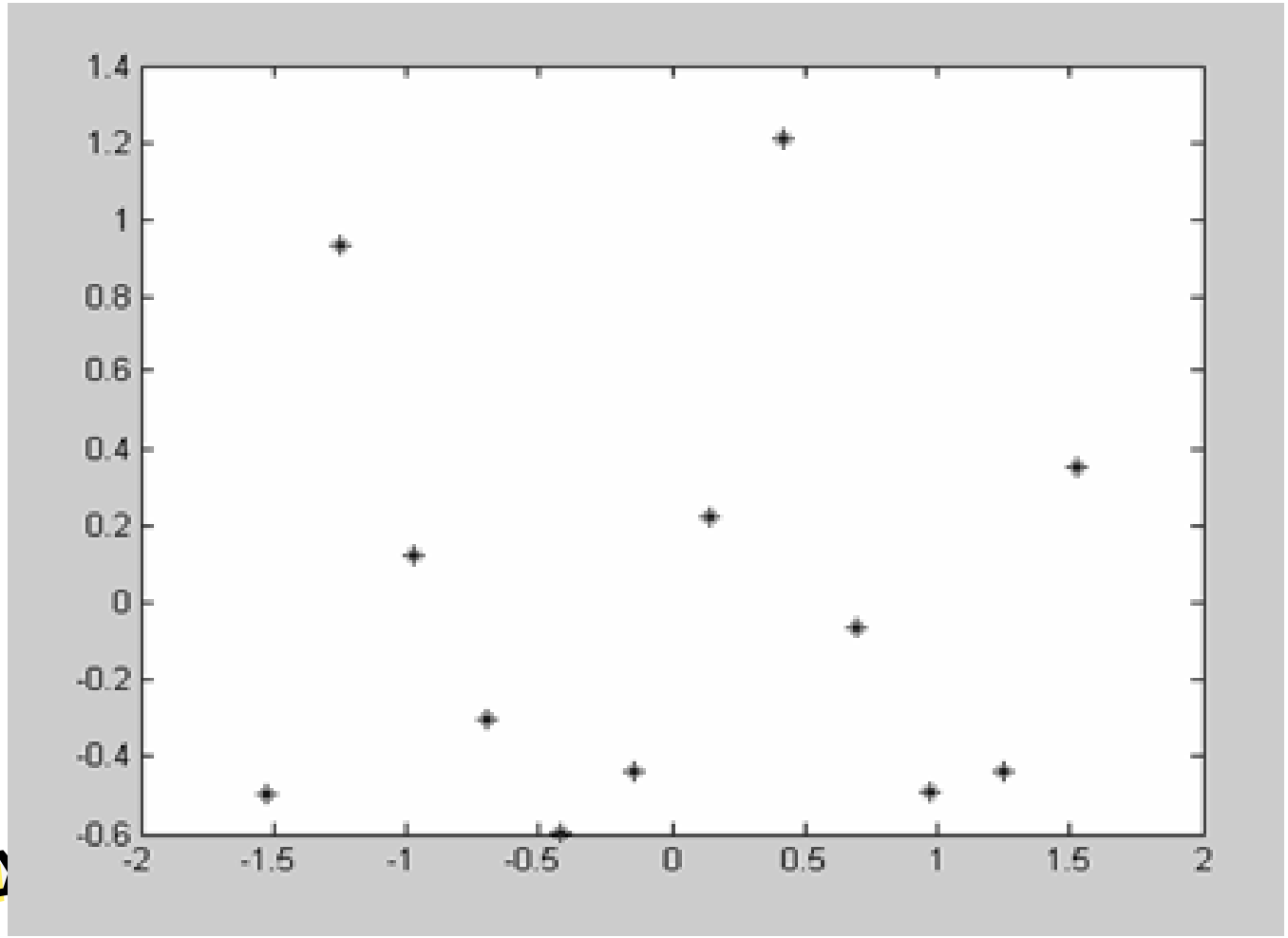
Reziduurile, fiind diferențele între valorile prognozate şi valorile reale, măsoară acuratețea aproximării cu ajutorul polinomului construit.

Vom reprezenta grafic aceste reziduuri:

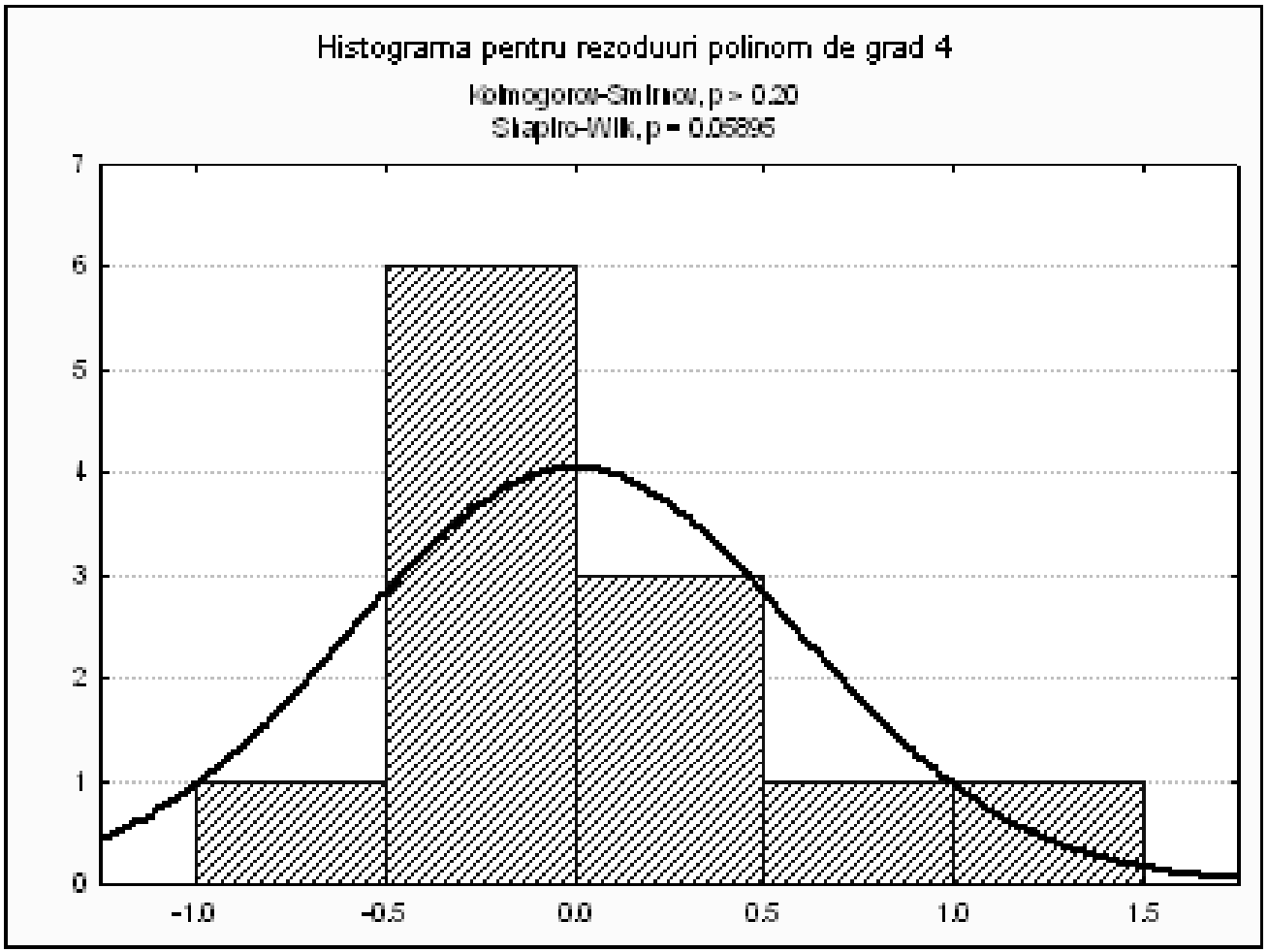


```
» rez=pol-Ys
  rez =
    Columns 1 through 5
     -0.4983    0.9323    0.1198   -0.3046   -0.5951
    Columns 6 through 10
     -0.4411    0.2211    1.2079   -0.0623   -0.4921
    Columns 11 through 12
     -0.4377    0.3500
» plot(Xs,rez,'*')
```



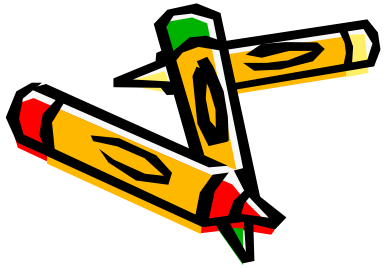

1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

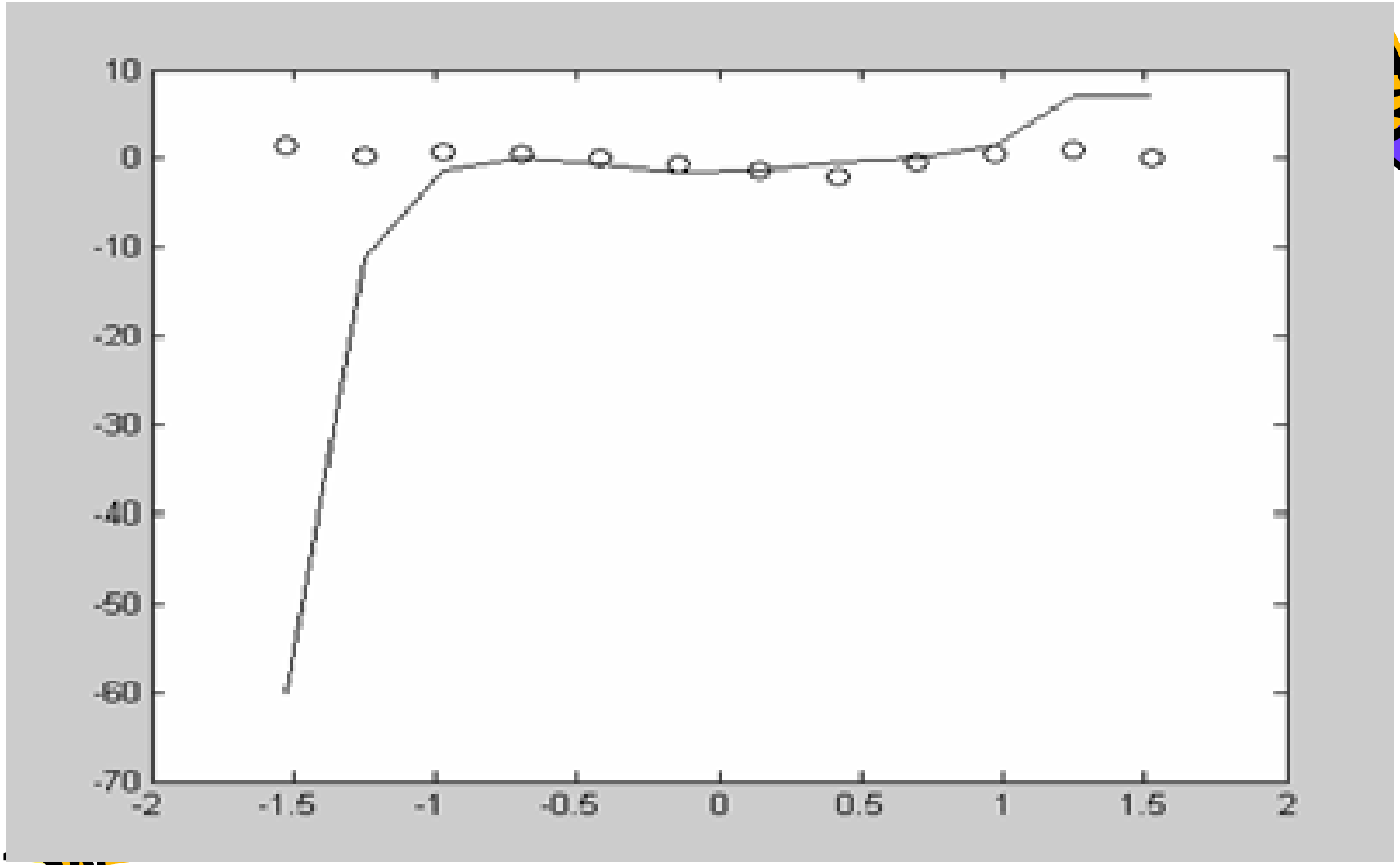
histograma reziduurilor



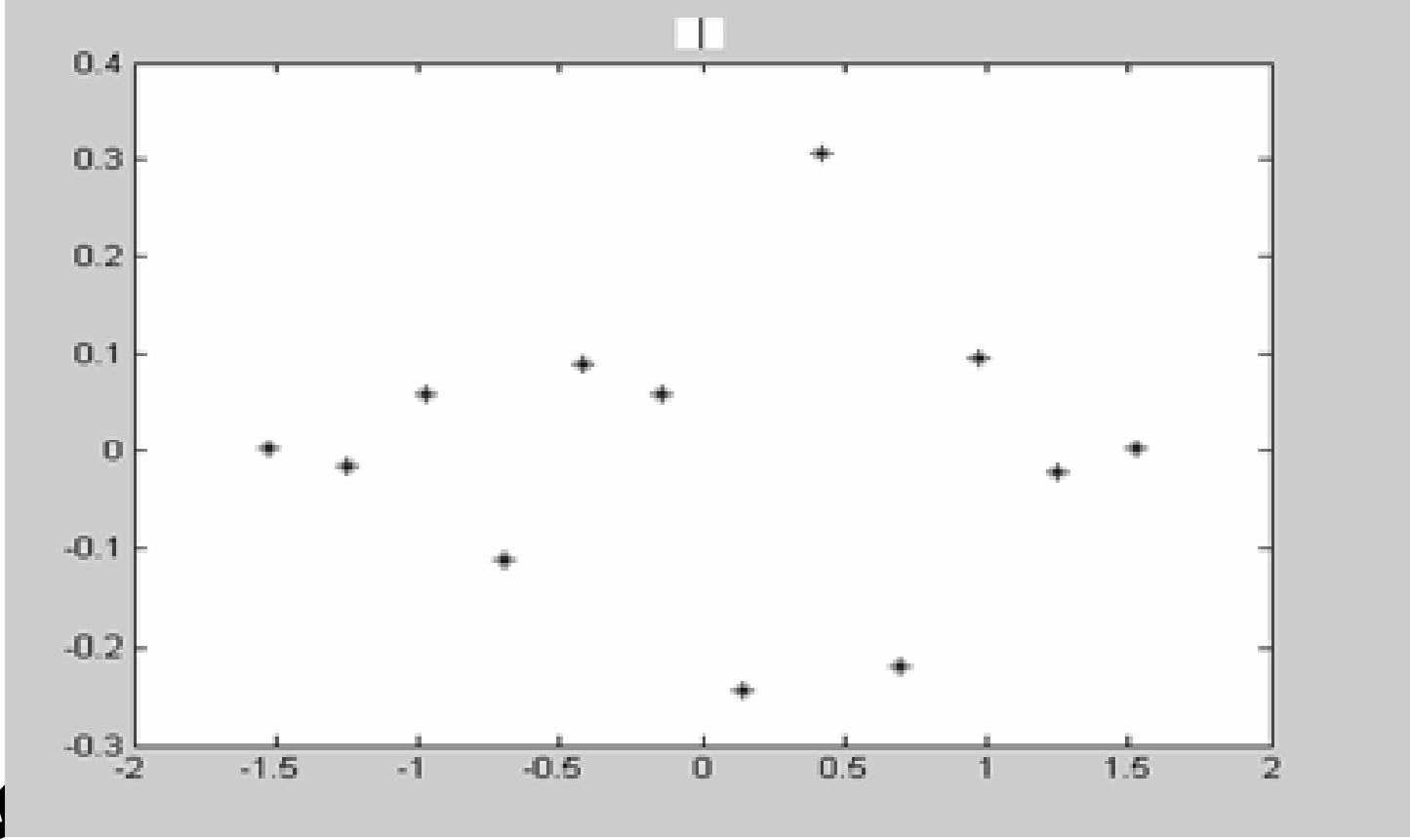
Nu ne putem declara mulțumiți de rezultat și vom încerca să mărim gradul polinomului de aproximare, să zicem 9:

```
» ps=polyfit(Xs,Ys,9)
    ps =
    Columns 1 through 5
    -1.5418    1.0532    8.0954   -4.3763  -15.4076
     Columns 6 through 10
     4.6069   13.0780    0.4928   -4.2654   -1.0409
» n=2;y2=ps(1)+ps(2)*Xs.^(n-1);
for n=3:9 y2=y2+ps(n)*Xs.^(n-1);end
»  y2;
»  plot(Xs,y2,'k',Xs,Ys,'ko')
```

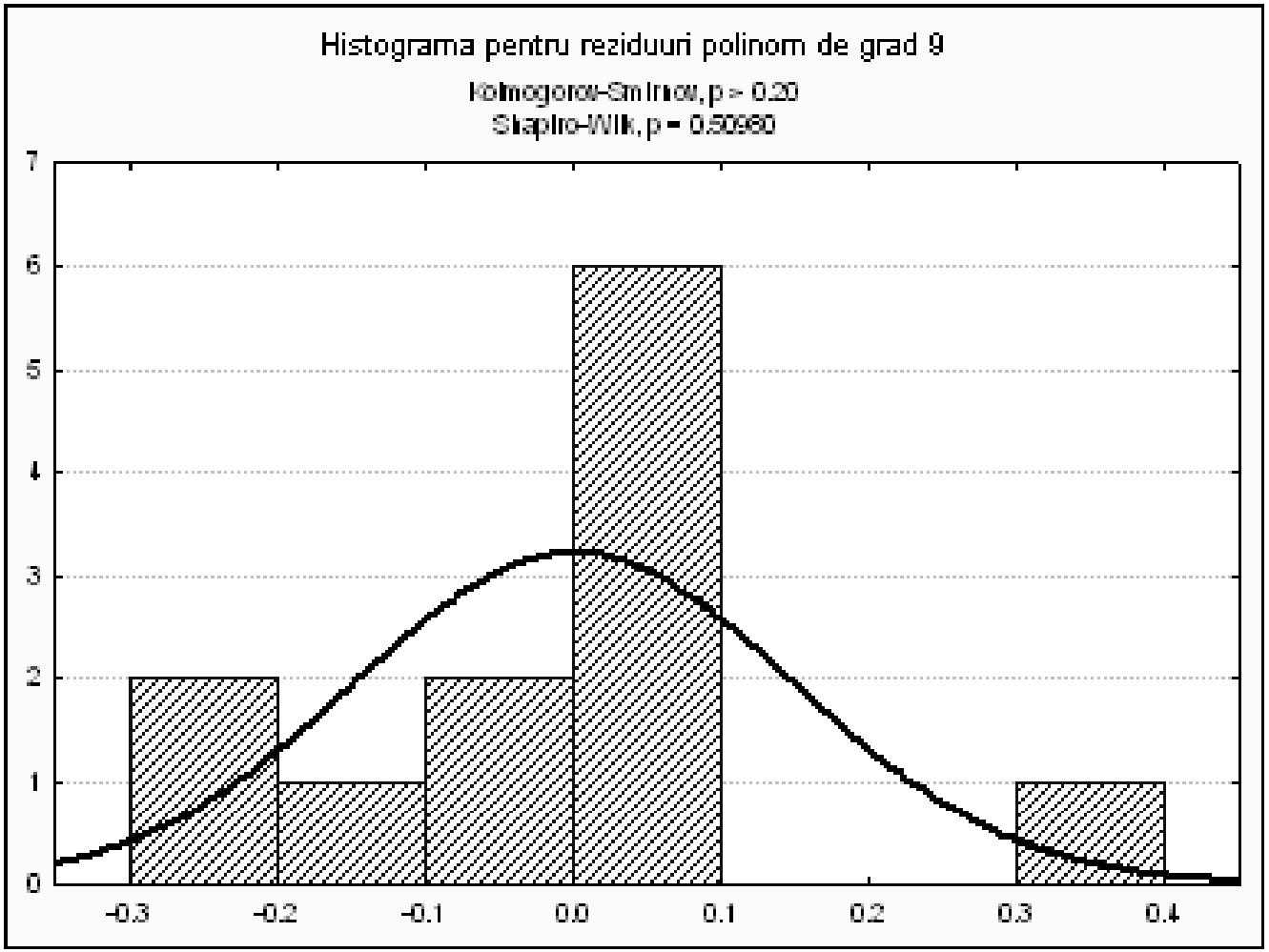



10
0
-10
-20
-30
-40
-50
-60
-70
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

» poll=polyval(ps,Xs,Ys); rez=poll-Ys; plot(Xs,rez,'*')



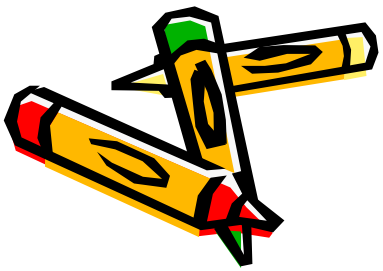
histoarama reziduurilor

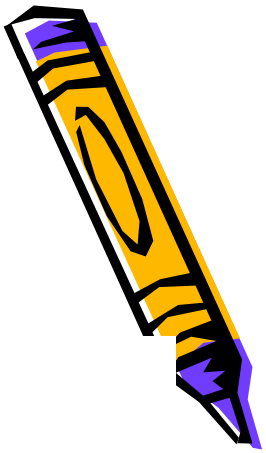


aproximarea prin combinatii de exponentiale

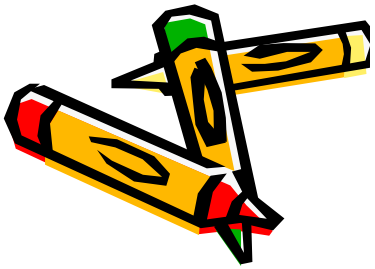
Considerăm că este necesar să amintim că în MATLAB este posibilă reprezentarea grafică în coordonate logaritmice sau semi-logaritmice, prin utilizarea funcțiilor

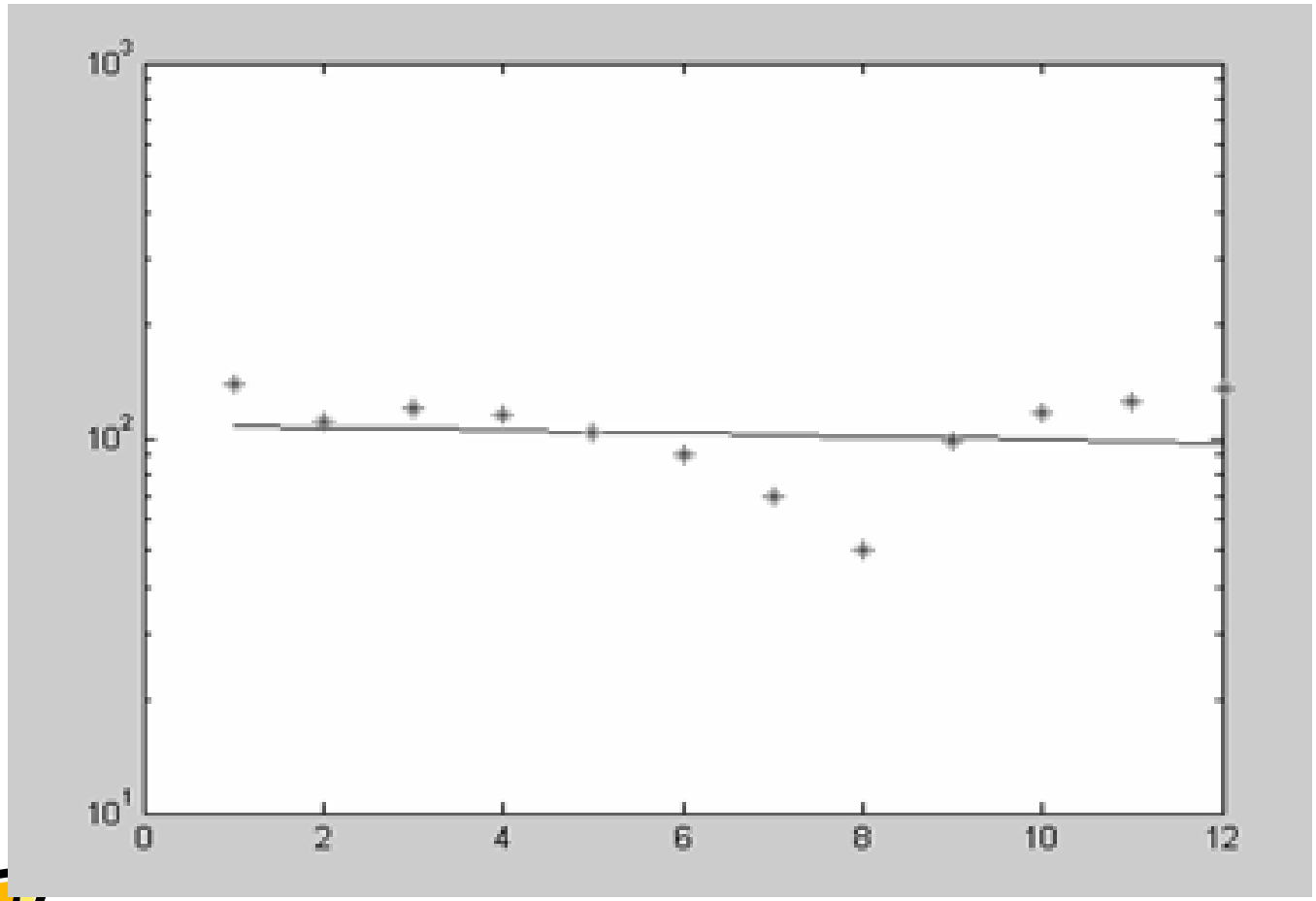
- loglog(x,y) scalează ambele axe utilizând logaritmul în baza 10;
- semilogx(x,y) scalează logaritmic numai aza Ox;
- semilogy(x,y) scalează logaritmic numai aza Oy.



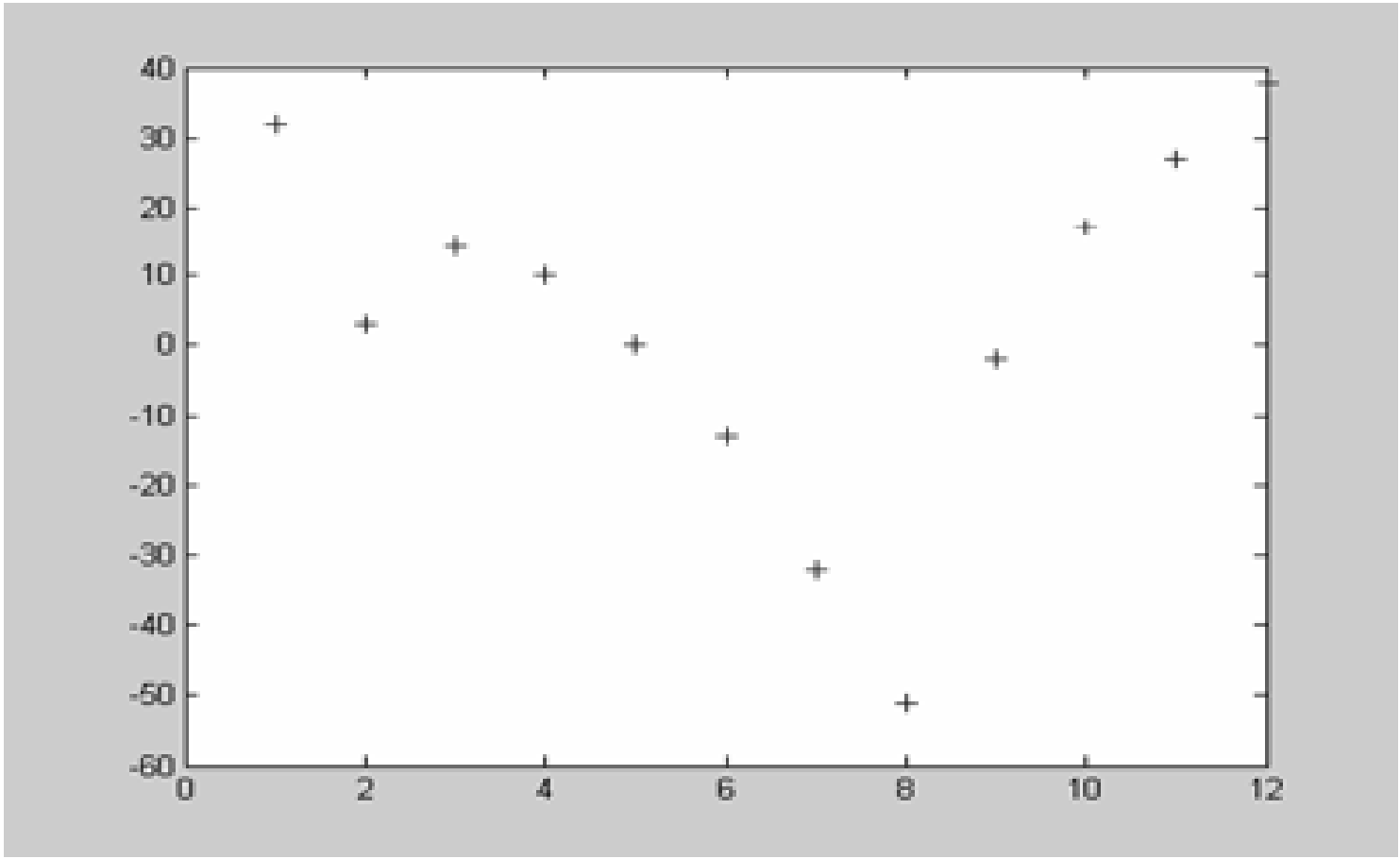


```
» logp1=polyfit(Xs,log10(Y),1);
    logp1 =
              -0.0150    2.0105
» logpred1=10.^polyval(logp1,Xs)
logpred1 =
  Columns 1 through 5
  107.9731  106.9455  105.9276  104.9194  103.9208
  Columns 6 through 10
  102.9317  101.9521  100.9817  100.0206   99.0686
  Columns 11 through 12
   98.1257   97.1918
» semilogy(X,logpred1,X,Y,'*r')
```

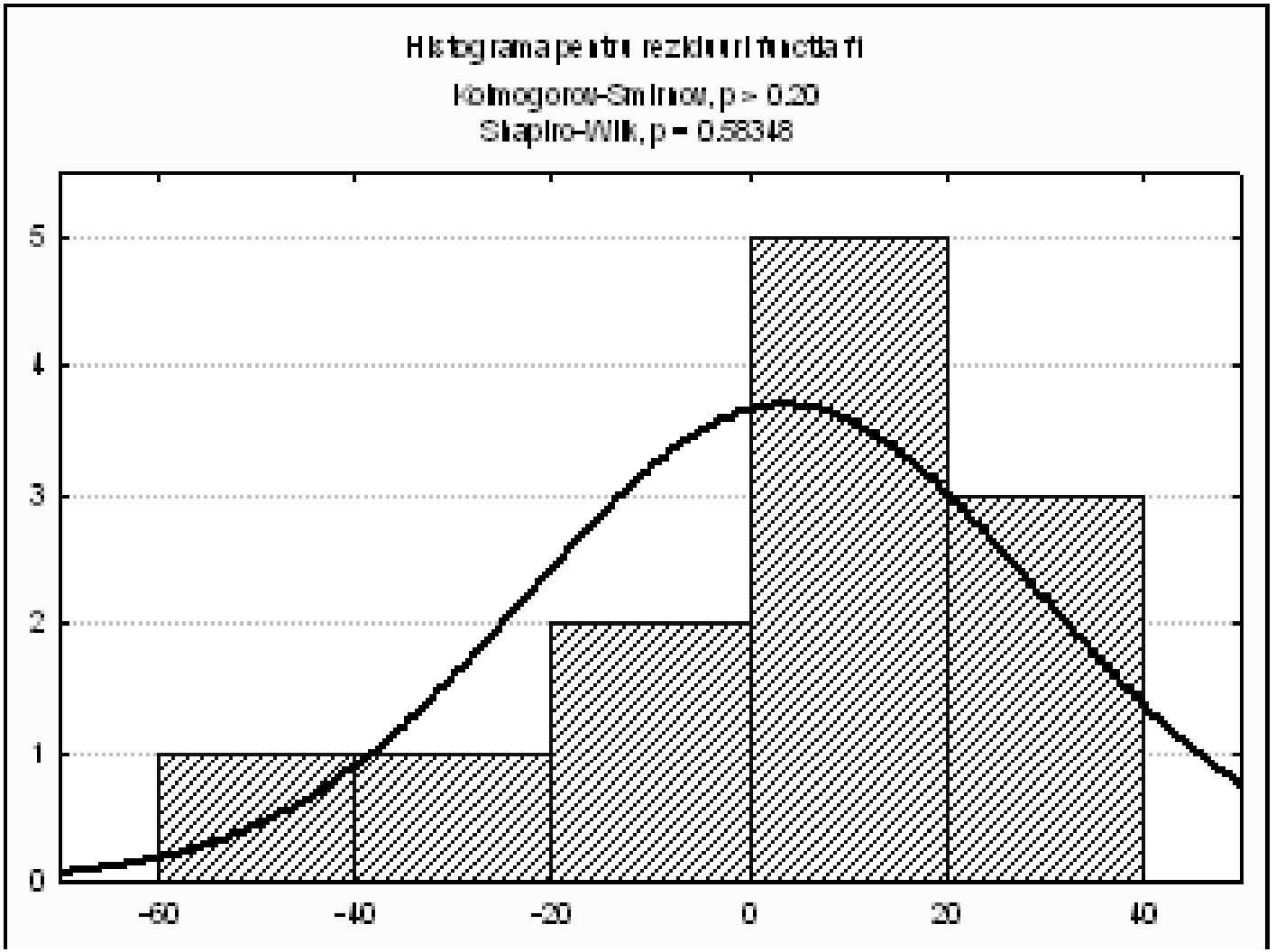



10^3
10^2
10^1
0
2
4
6
8
10
12

```
»r=Y-logpred1; plot(Xs,r,'+')
```

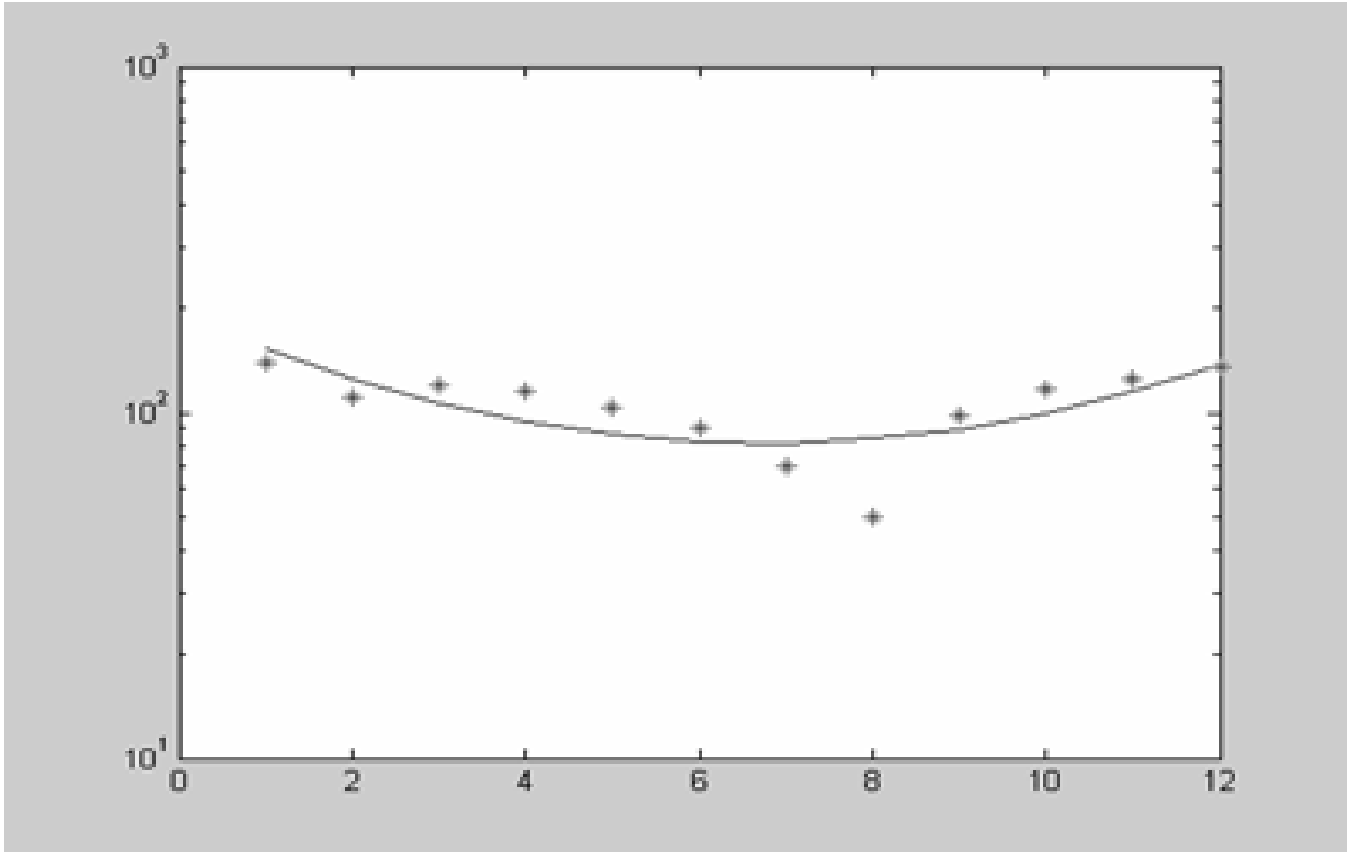


histograma reziduurilor



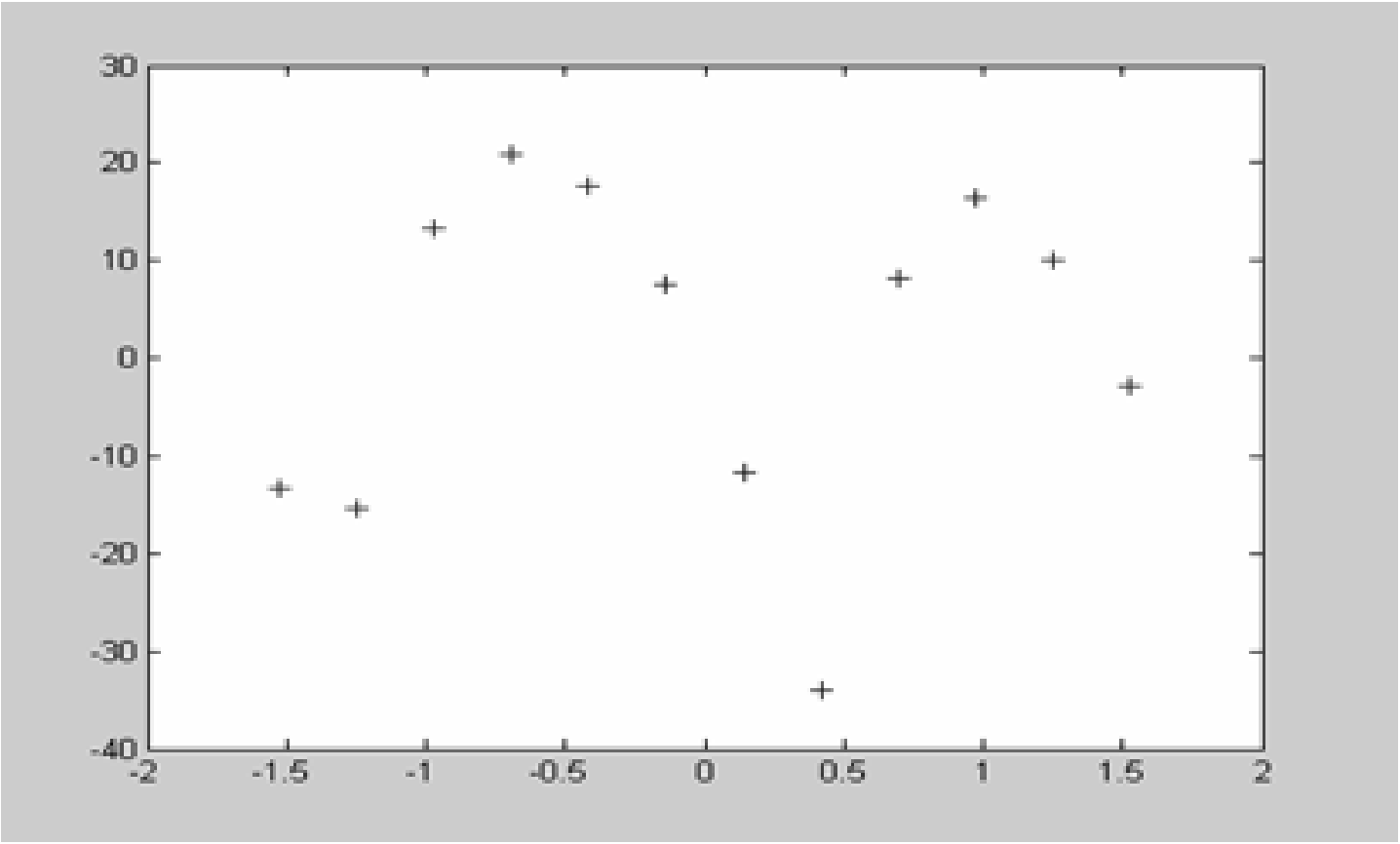
Vom încerca altă variantă de funcție

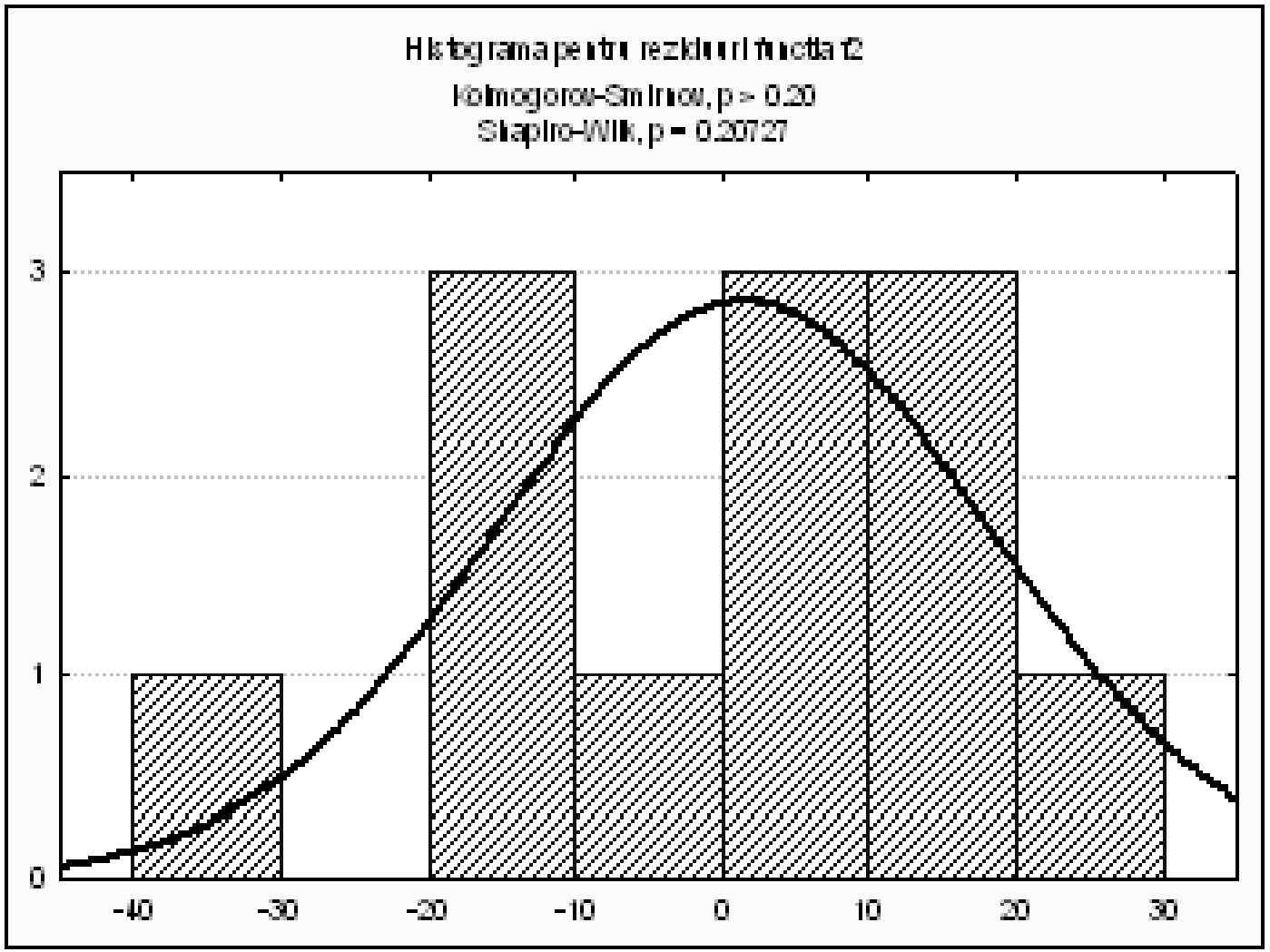
```
» logp2=polyfit(Xs,log10(Y),2);
»logpred2=10.^polyval(logp2,Xs);
»semilogy(X,logpred2,X,Y,'*')
```

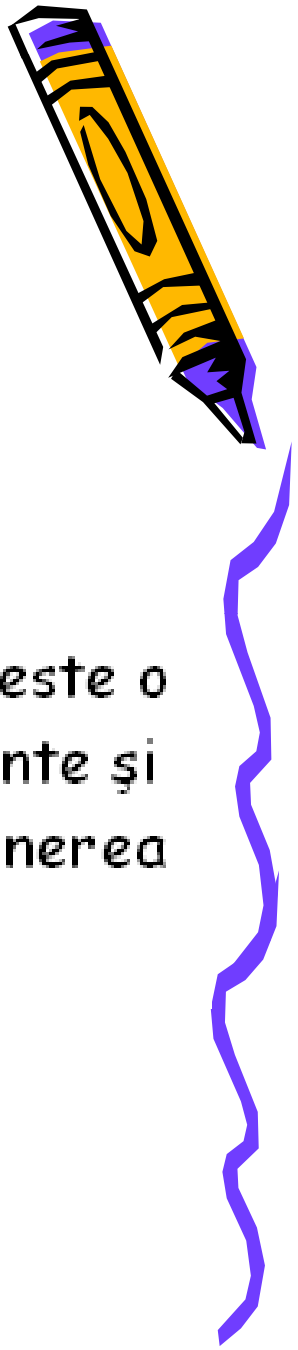


Reprezentăm grafic reziduurile precum şi histograma lor.

»r=Y-logpred2; plot(Xs,r,'+')




Histograma pentru reziduurile functiei f2
Kolmogorov-Smirnov, p > 0.20
Shapiro-Wilk, p = 0.20727
3
2
1
0
-40
-30
-20
-10
0
10
20
30



Problema utilizării regresiei neliniare la un set de date este o procedură euristică, fiind necesare mai multe experimente şi folosirea cunoştinţelor specifice domeniului pentru obţinerea unor rezultate bune