

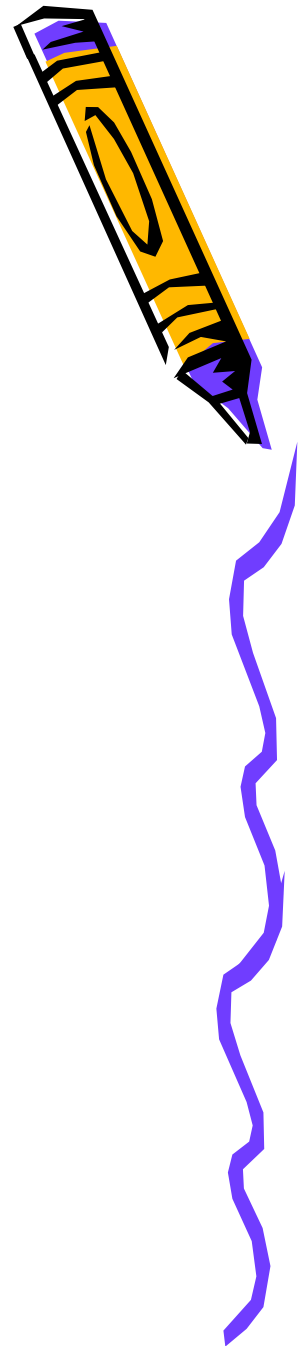


# Metode regresive

Marina Gorunescu  
mgorun@inf.ucv.ro



# Despre prognoză





*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.

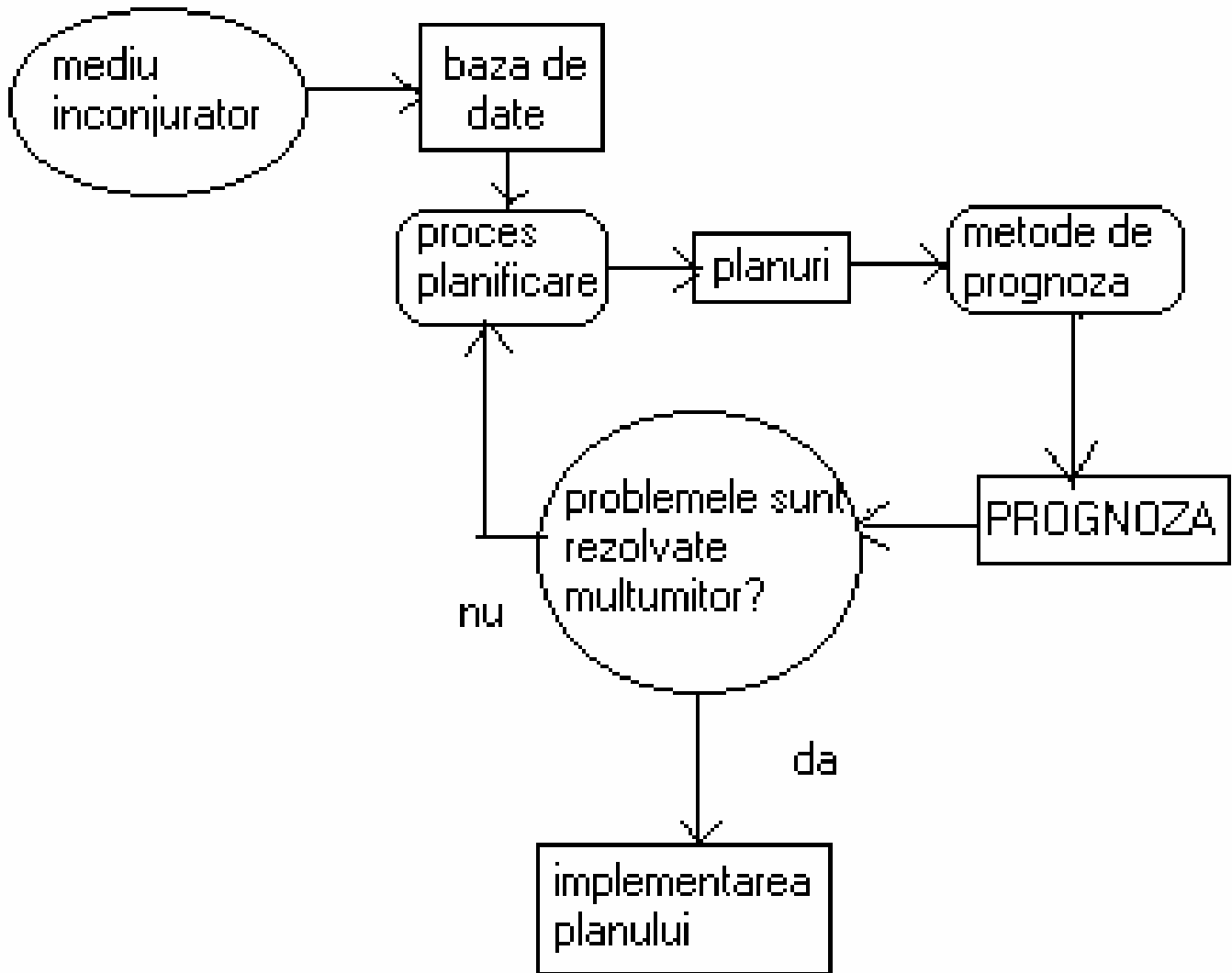




Cei ce se ocupă de planificare folosesc metodele prognozei pentru simularea și vizualizarea rezultatelor planurilor construite.

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







O asemenea abordare a problemei pare naturală, în practică, însă deseori se revizuieste 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.



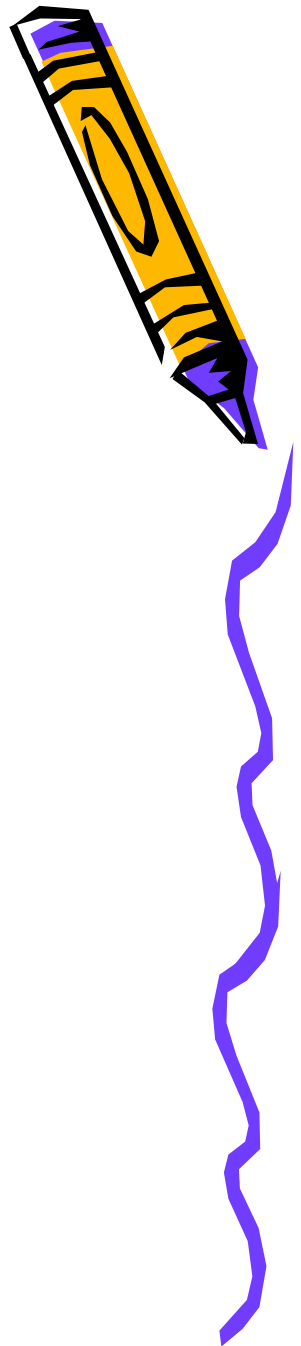
# aplicatii

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





# Regresia liniară





O problemă binecunoscută în multe arii de cercetare este aceea care presupune existența unui set de date privind două sau mai multe variabile aleatoare, scopul modelării fiind descrierea relației dintre ele în vederea prognozării valorilor uneia în raport cu valorile celeilalte sau celorlalte.





Această problemă se pune atunci când între variabilele aleatoare considerate există o legătură consistentă, bazată pe natura intimă a fenomenelor care stau la baza lor.





Este posibil ca, din punct de vedere formal, doar pe baza datelor numerice, acestea să pară corelate, de exemplu toate să aibă tendința de creștere în același timp, acest fapt nefiind însă susținut de natura fenomenelor în cauză.

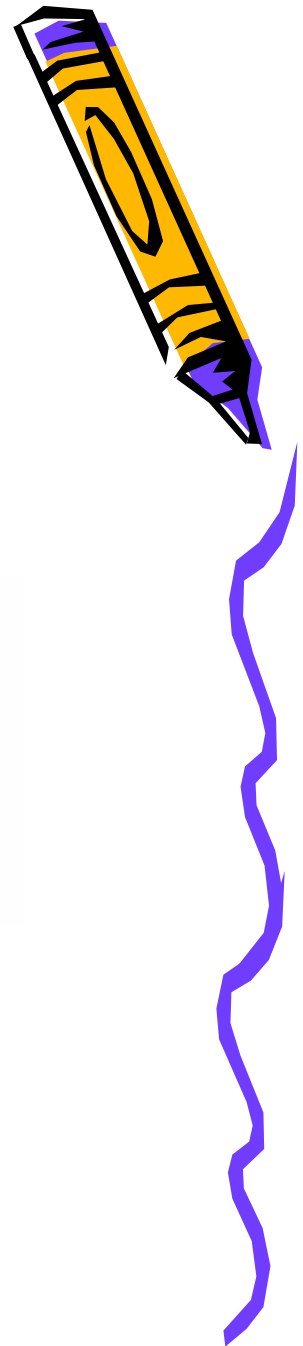


În concluzie, fără cunoașterea naturii intime a fenomenelor care stau la baza datelor, este hazardat de a întreprinde o analiză regresivă.



# Metoda regresiei

Metoda care se folosește pentru a descrie relația între valorile a două sau mai multe variabile aleatoare se numește *metoda regresiei* (noțiune introdusă de Pearson în 1908).



Dacă relația care stabilește legătura între variabila dependentă și variabilele independente este una liniară, se vorbește despre *regresia liniară*, în celălalt caz fiind vorba de *regresia neliniară* (polinomială, exponențială, logaritmică etc.).



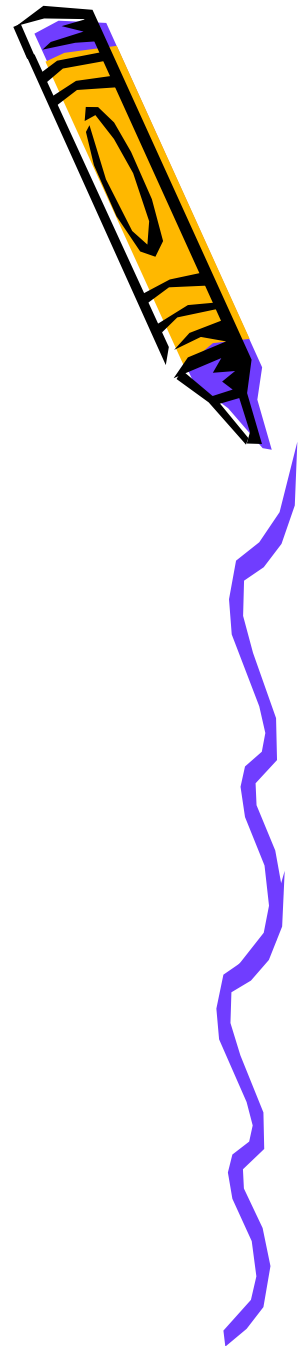


Vom prezenta pe scurt bazele regresiei liniare, începând prin a pune în evidență mecanismele cu care se pot evidenția legăturile între secvențe de date, provenind de la două sau mai multe variabile aleatoare.





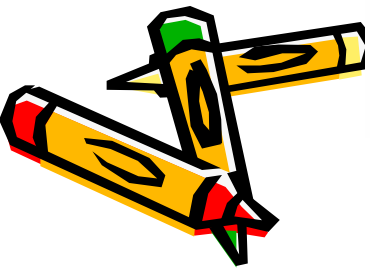
Cele mai multe studii statistice se referă la compararea a două sau mai multe grupuri de subiecți/obiecte sau la stabilirea unor legături existente între aceste grupuri.

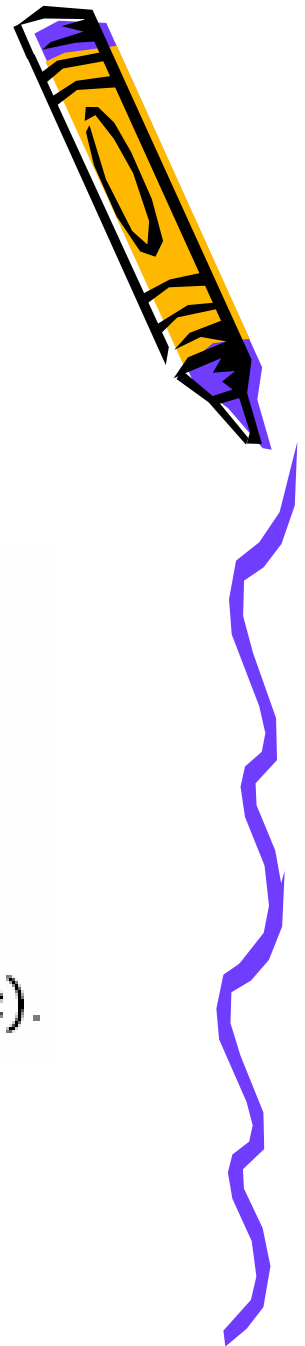


# Exemple

Analiza modului în care sunt corelate sau nu anumite valori medicale (colesterol, albumină, trigliceride etc.) în diferite cazuri, de exemplu pentru bărbați sau femei:

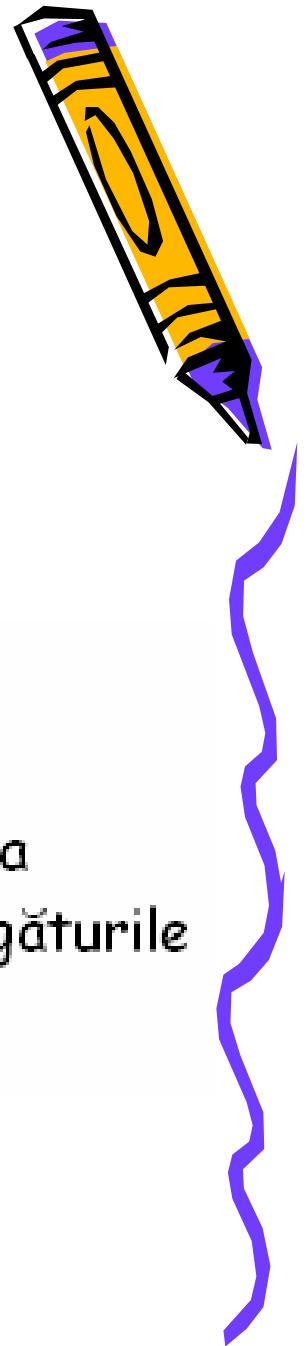
astfel, pe de o parte se identifică posibile legături între aceste caracteristici medicale la fiecare grup în parte și, pe de altă parte, se pun în evidență anumite deosebiri ce pot exista între grupuri, privitoare la studiul efectuat.



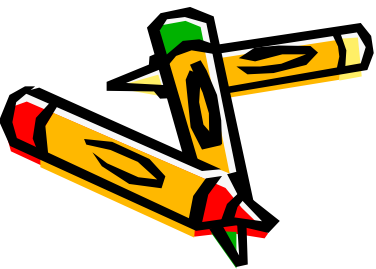


Un alt caz de astfel de analiză statistică, referitoare la stabilirea legăturii între două seturi de înregistrări, se aplică la stabilirea legăturii dintre înălțimea și greutatea unui individ pe baza analizei unui cuplu de serii statistice corespunzătoare înălțimii, respectiv greutății (pentru un eșantion semnificativ dintr-o populație).





În ambele exemple este deci vorba de descrierea, analizarea și compararea a două variabile statistice simultan (evident că se poate considera și descrierea statistică individuală, dar aceasta nu poate releva legăturile sau comparația între cele două seturi de date).



# Metoda diagramei de împrăștiere

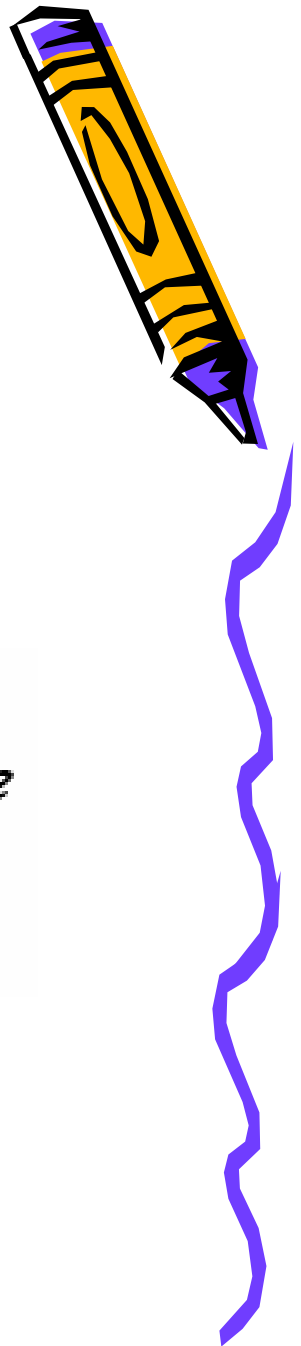
Să considerăm mai întâi cazul a două serii statistice  $\{x_i\}_{1 \leq i \leq n}$  și  $\{y_i\}_{1 \leq i \leq n}$  definite pe același lot de subiecți/obiecte.

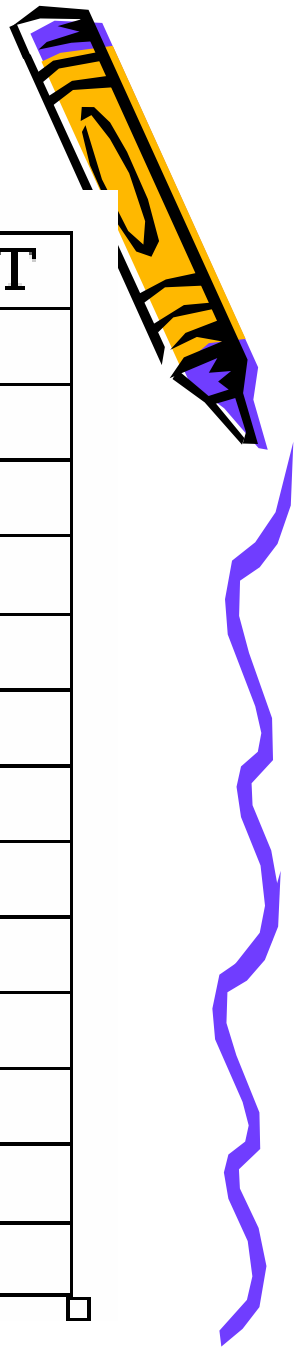
Plecând de la cele două serii, putem considera seria cuplurilor de observații  $\{(x_i, y_i)\}_{1 \leq i \leq n}$  definite de cele două variabile statistice pe același individ/obiect  $i$ . Cel mai obișnuit mod de reprezentare grafică al acestui cuplu de observații este cel folosind *norul* de puncte definit de reprezentarea bidimensională a punctelor  $(x_i, y_i)$  - așa numita *diagramă de împrăștiere*.



# Exemplu

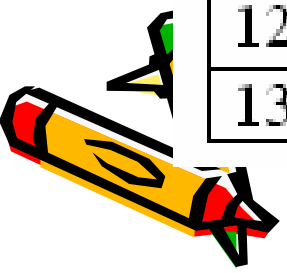
Considerăm o serie statistică formată din 25 observații privind două din principalele enzime serice: AST (*aspartate transaminase*) și gGT (*gamma glutamyl transferase*), prelevate de la un lot de 25 pacienți.





‡

patient	AST	gGT	patient	AST	gGT
1	22	11	14	30	7
2	24	40	15	16	24
3	10	10	16	12	16
4	32	6	17	33	27
5	24	20	18	33	9
6	53	15	19	9	5
7	24	9	20	21	11
8	58	34	21	17	10
9	13	22	22	54	15
10	30	10	23	32	8
11	22	4	24	19	5
12	29	13	25	23	4
13	18	3			



□



Diagrama ,norului' de împrăștiere a cuplurilor (AST, gGT)  
pentru lotul de 25 subiecți:

»  $x=[22\ 24\ 10\ 32\ 24\ 53\ 24\ 58\ 13\ 30\ 22\ 29\ 18\ 30\ 16\ 12\ 33\ 33\ 9$   
 $21\ 17\ 54\ 32\ 19\ 23];$

»  $y=[11\ 40\ 10\ 6\ 20\ 15\ 9\ 34\ 22\ 10\ 4\ 13\ 3\ 7\ 24\ 16\ 27\ 9\ 5\ 11\ 10\ 15$   
 $8\ 5\ 4];$

» `plot(x,y,'kO')`





diagrama imprastierii ( AST, gGT)

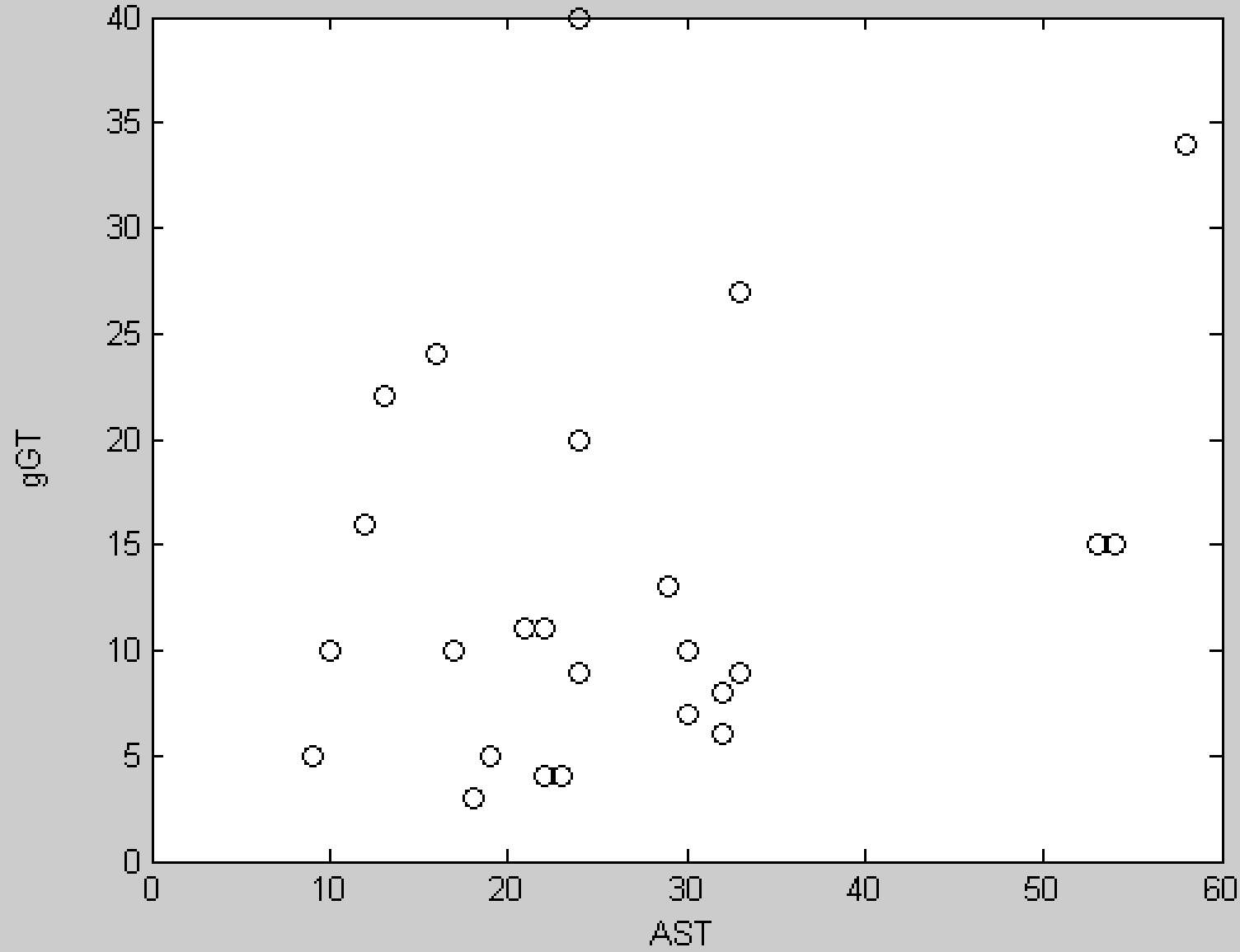




Diagrama împrăștierei se dovedește a fi un instrument util în descrierea statistică, putând produce informații importante privind legătura între cele două serii statistice. Este preludiul unei analize statistice analitice ulterioare.

Astfel, analizând forma norului, se pot deduce la prima vedere informații utile privind legăturile între variabile:





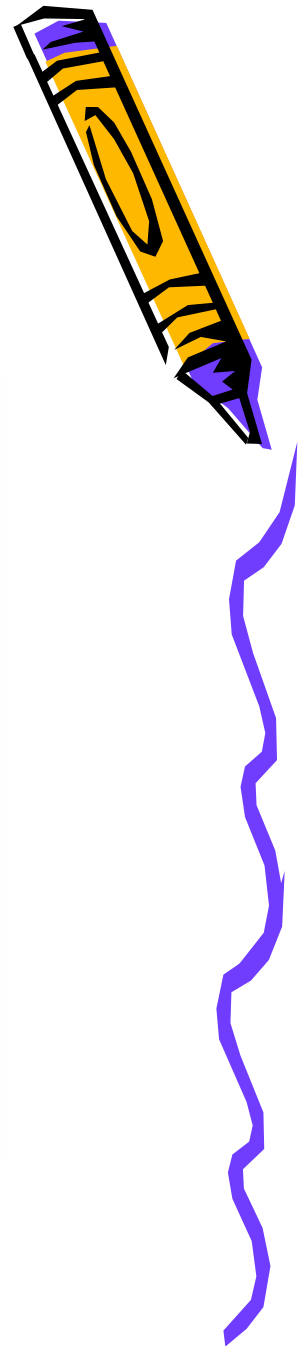
- norul are forma unei elipse mai mult sau mai puțin alungite, paralelogram alungit, figură geometrică alungită simetrică față de o axă, fapt ce implică o legătură liniară între variabile;
- norul are forma unui cerc sau pătrat, fapt ce implică independența variabilelor.

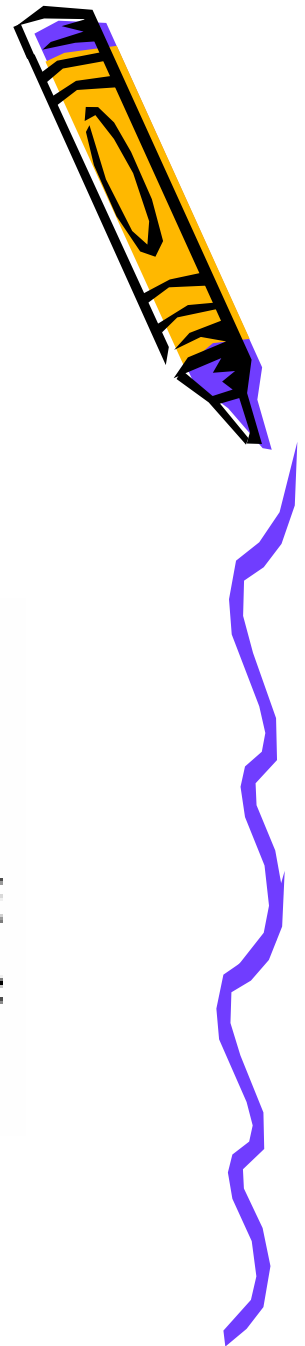


# metoda prezentarii numerice a datelor multiple

Prezentarea este făcută prin tabele, existând câteva reguli pentru obținerea unui efect semnificativ.

- este indicat ca datele de aceeași natură să fie puse pe coloane și nu pe linii, deoarece s-a observat că astfel pot fi citite și analizate vizual mai ușor.
- tabelele pot conține fie date neprelucrate, adică datele reale, neprocesate, atunci când volumul acestora nu este prea mare pentru observator, fie date prelucrate, rezultate ale procesării statistice.



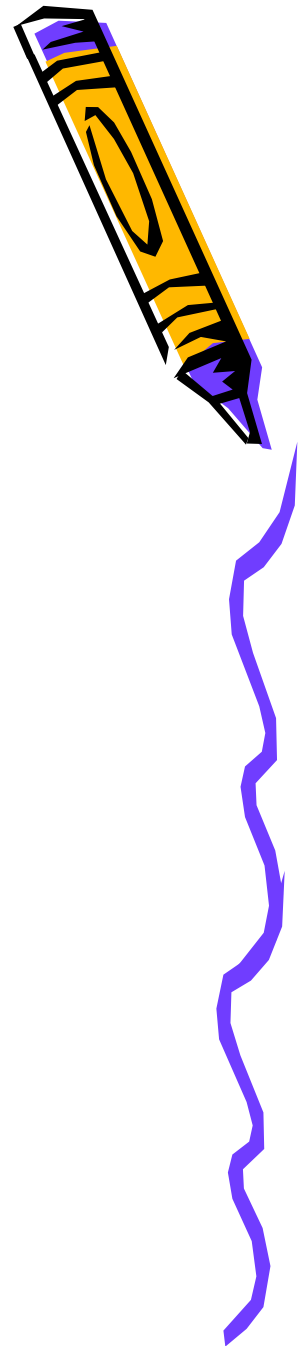


O mare parte a studiilor statistice uzuale se ocupă cu analiza relației între două variabile statistice ce corespund aceluiași grup de subiecți/obiecte. Cel mai cunoscut exemplu se referă la relația ce există între înălțimea și greutatea unui individ ce corespunde unor anumite standarde geografice, rasiale etc.



Pentru a o identifica, se studiază relația dintre cele două caracteristici măsurate pe indivizii dintr-un anumit lot.

Există două motive importante pentru care se efectuează un asemenea studiu:





1. Descrierea relației care ar putea exista între cele două variabile, analizând legătura între cele două serii de observații.

Concret, se analizează dacă tendința ascendentă a uneia implică o tendință ascendentă, descendentă sau nici o tendință a celeilalte;





- 
2. În ipoteza existenței unei legături reale între ele, identificată în prima instanță, să se poată prognoza valorile uneia în raport cu valorile celeilalte pe baza ecuației de regresie.

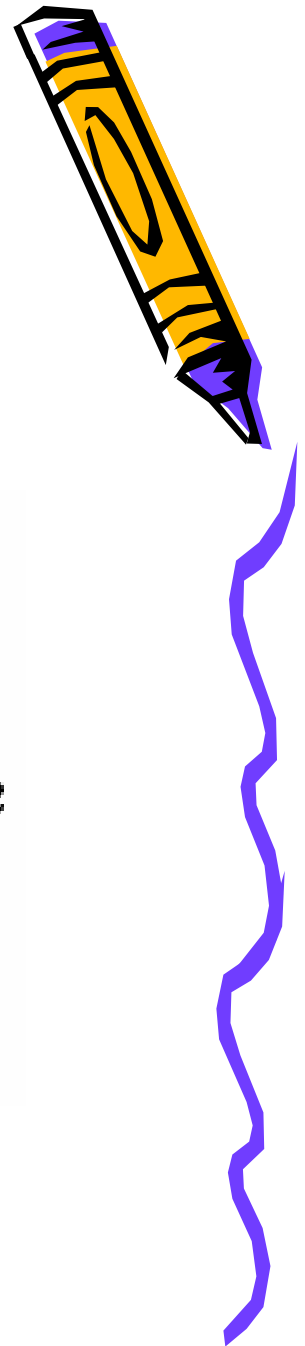




# Metoda corelației

Scopul final este *prognoza*, în condiția în care este posibilă.

Metoda prin care analizăm posibilele asociații între valorile a două variabile statistice continue prelevate de la același grup de subiecți, este cunoscută ca metoda *corelației* și are ca indice *coeficientul de corelație*.



# coeficientul de corelație

Coeficientul de corelație poate fi calculat pentru orice set de date, dar, pentru ca el să aibă relevanță statistică, trebuie îndeplinite două condiții majore:

1. cele două variabile să fie definite de același lot de subiecți/obiecte, cuplurile de date corespunzând aceluiași individ/obiect/instanță;





2. cel puțin una din variabile să aibă o repartiție aproximativ normală, ideal fiind ca ambele să fie normal repartizate.

Dacă datele nu au o repartiție normală (cel puțin una din variabile) se procedează fie la transformarea lor pentru normalizare, fie la considerarea unor coeficienți de corelație ne-parametrici.

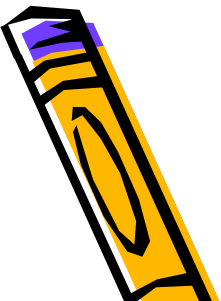




În afară de coeficientul de corelație se poate obține, în cazul când ambele variabile sunt aproximativ gaussiene, și intervalul de încredere corespunzător acestuia.

Prezentăm formulele matematice ce stau la baza calculării coeficientului de corelație și a intervalului de încredere corespunzător.





Să considerăm două serii statistice  $\{x_i\}_{1 \leq i \leq n}$  și  $\{y_i\}_{1 \leq i \leq n}$  corespunzătoare variabilelor statistice  $X$  și  $Y$ , generate de un grup de subiecți/obiecte.

Prin *coeficientul de corelație*  $r$  al celor două variabile, numit și *Pearson's r* vom înțelege numărul real  $r$ , cuprins între  $-1$  și  $1$ , definit de formula

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}};$$

(folosită în programele de computer):





Pentru calcule concrete (manuale, cu ajutorul calculatorului de buzunar) se folosește formula de mai sus, scrisă sub forma:

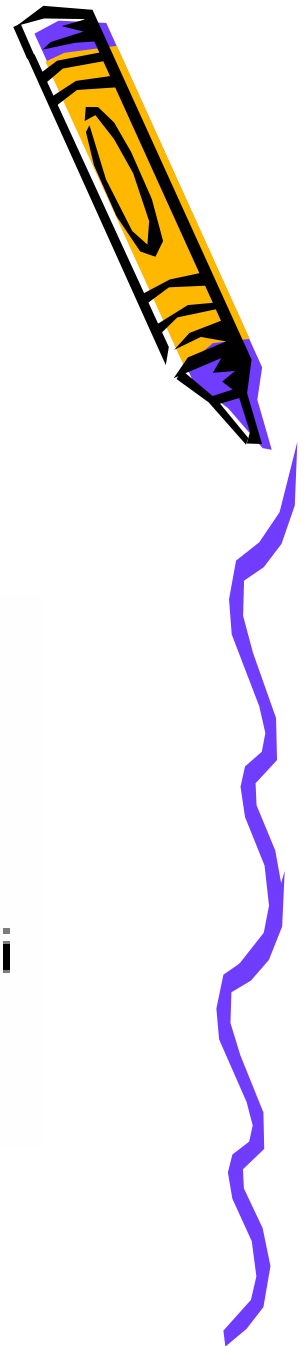
$$r = \frac{\sum_{i=1}^n x_i \cdot y_i - \frac{1}{n} \left( \sum_{i=1}^n x_i \right) \cdot \left( \sum_{i=1}^n y_i \right)}{\sqrt{\left( \sum_{i=1}^n x_i^2 - \frac{1}{n} \cdot \left( \sum_{i=1}^n x_i \right)^2 \right) \cdot \left( \sum_{i=1}^n y_i^2 - \frac{1}{n} \cdot \left( \sum_{i=1}^n y_i \right)^2 \right)}}$$

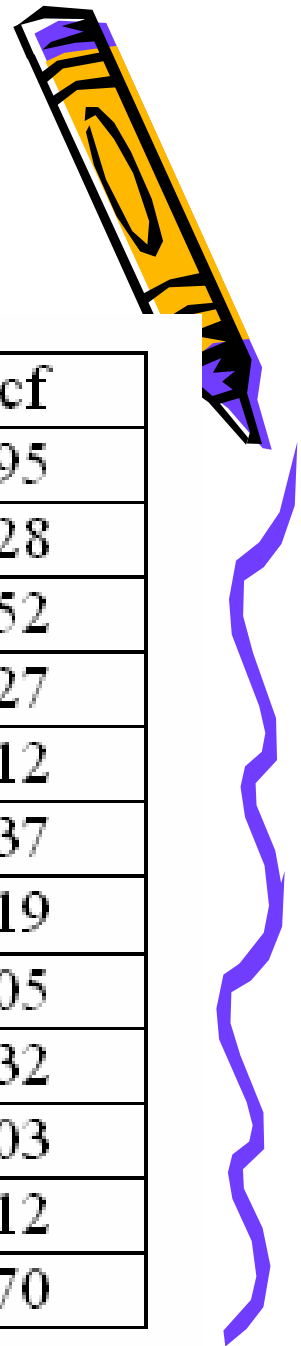


# Exemplu

Considerăm datele culese de la un lot de 24 pacienți, având diabet zaharat de tip I privind două variabile:

- $G(\text{mmol/l})$  - glucoza în sânge;
- $Vcf(\% / s)$  - viteza medie de contracție a ventriculului stâng, obținută prin eco - cardiografie.

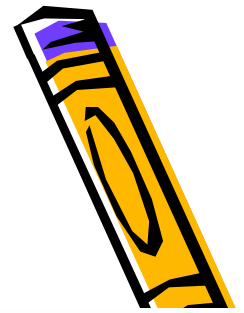




patient	G	Vcf	patient	G	Vcf
1	15.3	1.76	13	19.0	1.95
2	10.8	1.34	14	15.1	1.28
3	8.1	1.27	15	6.7	1.52
4	19.5	1.47	16	8.6	1.27
5	7.2	1.27	17	4.2	1.12
6	5.3	1.49	18	10.3	1.37
7	9.3	1.31	19	12.5	1.19
8	11.1	1.09	20	16.1	1.05
9	7.5	1.18	21	13.3	1.32
10	12.2	1.22	22	4.9	1.03
11	6.7	1.25	23	8.8	1.12
12	5.2	1.19	24	9.5	1.70







Vom calcula în *MATLAB* matricea coeficienților de corelație, folosind funcția *corrcoef*:

```
»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];
```

```
» A=[G' V'];
```

```
»corrcoef(A)
```

```
ans =
```

```
1.0000 0.4189
```

```
0.4189 1.0000
```

Coeficientul de corelație a celor două variabile este 0.4189.





În ceea ce privește construcția intervalului de încredere 95% pentru  $r$ , plecând de la faptul că variabila aleatoare:

$$z = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right)$$

este normal repartizată, rezultă că intervalul de încredere 95% pentru  $z$  are forma  $(z_1, z_2)$  unde:

$$z_1 = z - \frac{1.96}{\sqrt{n-3}}, \quad z_2 = z + \frac{1.96}{\sqrt{n-3}},$$

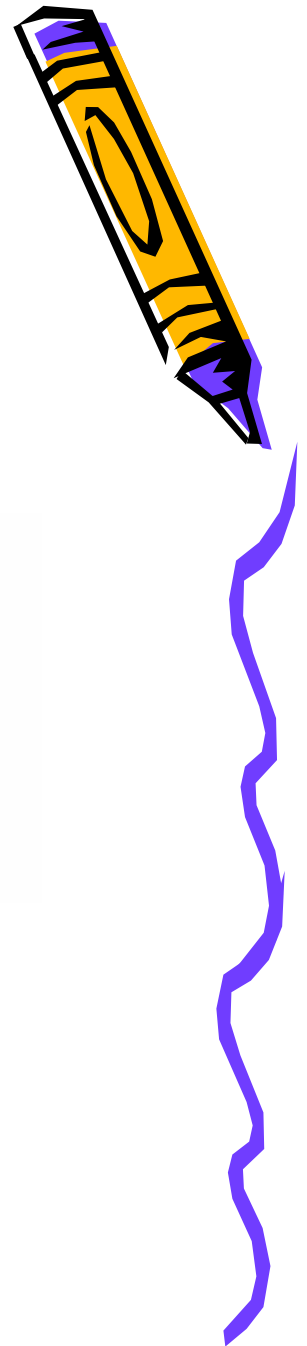




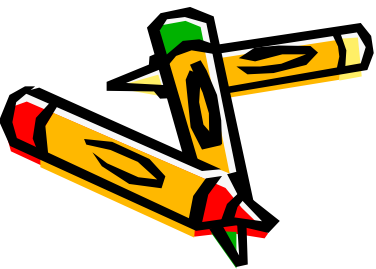
de unde rezultă că, aplicând transformarea inversă, obținem intervalul de încredere 95% pentru  $r$ , dat de:

$$\left( \frac{e^{2z_1} - 1}{e^{2z_1} + 1}, \frac{e^{2z_2} - 1}{e^{2z_2} + 1} \right)$$





Coeficientul de corelație  $r$  (Pearson) ia valori cuprinse între  $-1$  și  $+1$ , trecând deci și prin valoarea  $0$  care indică o asociație neliniară între cele două variabile (independență liniară):





- o valoare a lui  $r$  apropiată de  $-1$  indică o corelație negativă puternică, adică tendința unei variabile de a scădea puternic când cealaltă variabilă crește;
- o valoare a lui  $r$  apropiată de  $+1$  indică o corelație pozitivă puternică, adică tendința de creștere puternică a unei variabile atunci când și cealaltă variabilă crește.





Problema care se pune în acest context este stabilirea unui prag pentru  $r$  de la care să putem trage concluzia că cele două variabile sunt într-adevăr corelate.





În acest sens indicăm

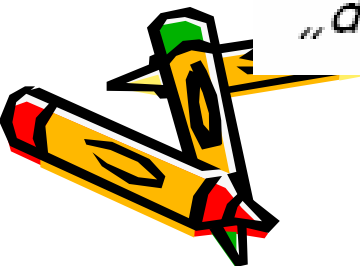
- fie un prag definit de inegalitatea

$$|r| \cdot \sqrt{n-1} \geq 3,$$

prag de la care se poate considera că legătura dintre cele două variabile este semnificativă,

- fie utilizarea nivelului de semnificație  $p$  asociat calculării coeficientului  $r$  :

*„dacă  $p < 0.05$  atunci legătura este semnificativă”.*



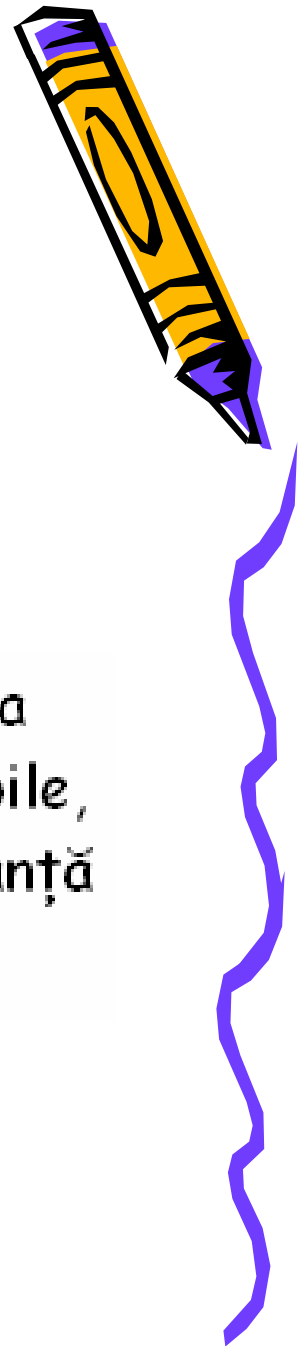


Un coeficient de corelație important nu implică totdeauna în mod necesar o legătură naturală, intrinsecă, între caracteristicile ce definesc cele două variabile statistice analizate.

De exemplu, în medicină, aceeași valoare redusă a coeficientului de corelație poate fi importantă în epidemiologie dar nesemnificativă din punct de vedere clinic







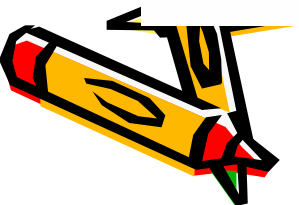
În concluzie, coeficientul de corelație este o măsură a legăturii liniare, „aritmetice”, dintre cele două variabile, care poate fi câteodată și întâmplătoare, fără relevanță reală.





Presupunând că legătura dintre cele două variabile, reliefată de coeficientul de corelație, nu este întâmplătoare, există trei posibile explicații:

- Variabila  $X$  influențează (cauzează) variabila  $Y$ ;
- Variabila  $Y$  influențează (cauzează) variabila  $X$ ;
- Ambele variabile  $X$  și  $Y$  sunt influențate de același fenomen din fundal, diferit de ele.





Prezentarea corelației dintre două variabile statistice trebuie să urmeze un anumit model:

- se prezintă mai întâi diagrama de împrăștiere a norului de puncte;
- când se prezintă coeficientul de corelație  $r$ , valoarea sa trebuie să aibă două zecimale și să fie însoțită de nivelul de semnificație  $p$  și de intervalul de încredere corespunzător, dacă este posibil.
- trebuie menționat și numărul de observații analizate.





*Covarianța* poate fi privită ca „momentul” corelației și, amintindu-ne și de formula sa probabilistă dată de:

$$\text{cov}(X, Y) = E[(X - E[X]) \cdot (Y - E[Y])] = E[XY] - E[X]E[Y],$$

observăm că ea este nulă dacă variabilele care generează cele două serii statistice sunt (liniar) independente.





# Teoremă

- Coeficientul de corelație  $r$  a două variabile aleatoare  $X$  și  $Y$  Ia valori în intervalul  $[-1, 1]$ , este nul dacă variabilele sunt independente și este egal cu  $\pm 1$  dacă și numai dacă variabilele  $X$  și  $Y$  verifică ecuația:

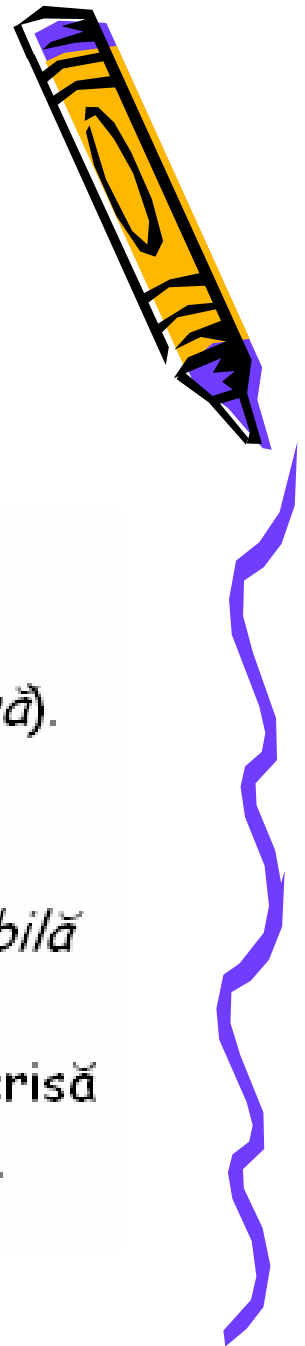
$$aX + bY = c \Leftrightarrow Y = aX + b, \quad a, b, c, A, B \in \mathbf{R}$$



# Regresia liniară

Modul de prezentare a legăturii liniare dintre două variabile (numerice), atunci când aceasta există, se numește *metoda regresiei liniare (regresia liniară)*. În acest scop, se consideră una dintre variabile ca *variabilă independentă* sau *variabilă predictor*, iar cealaltă variabilă ca *variabilă dependentă* sau *variabilă răspuns*.

Legătura liniară dintre cele două variabile este descrisă de o ecuație liniară, așa-numita *ecuația de regresie*.



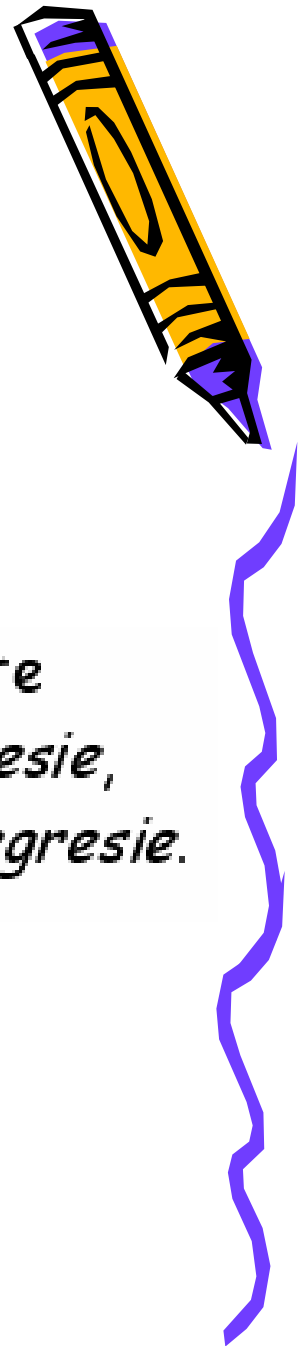
# variabila predictor, variabila raspuns

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



# dreapta de regresie

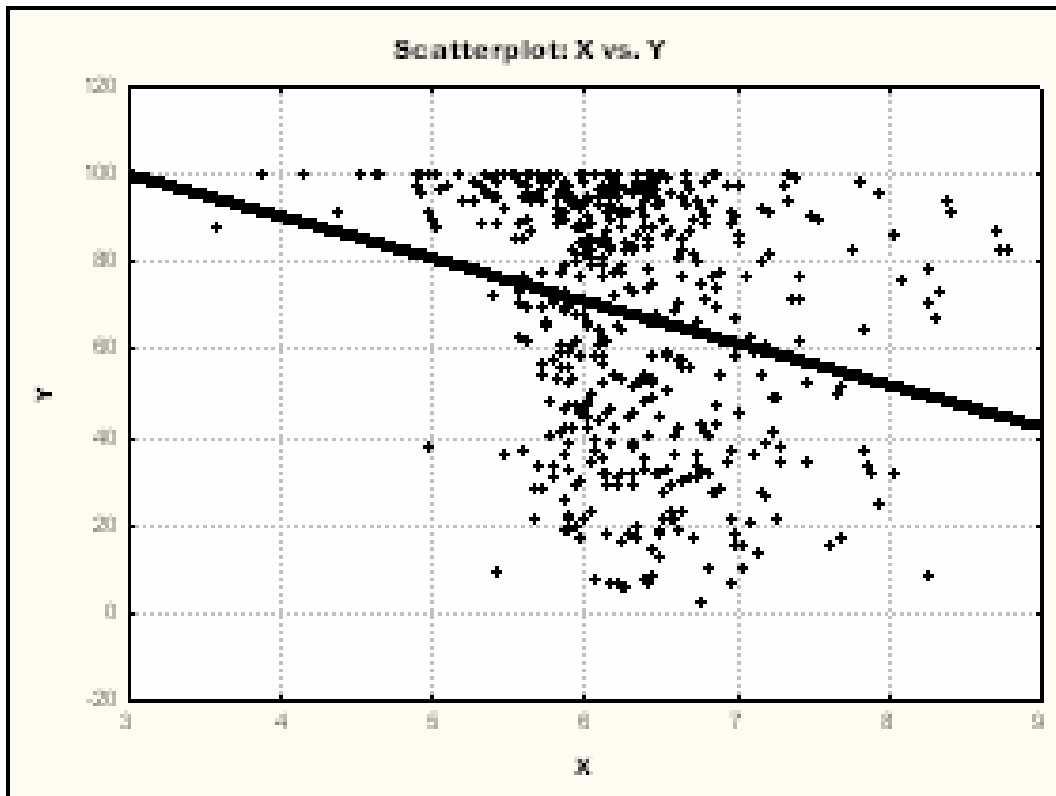
Legătura liniară dintre cele două variabile este descrisă de o ecuație liniară, *ecuația de regresie*, căreia îi corespunde geometric *dreapta de regresie*.





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





# metoda celor mai mici patrate



Ecuatia drepte de regresie se stabileste pe baza metodei "*celor mai mici patrate*".

Presupunem ca obtinem urmatorul tabel de valori, rezultat al masurarii unei marimi fizice  $f(x)$ .

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

astfel incat  $y_i = f(x_i), 0 \leq i \leq n$ ,

fiecare dintre valori fiind calculata cu o anumita eroare.





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

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 patratice

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

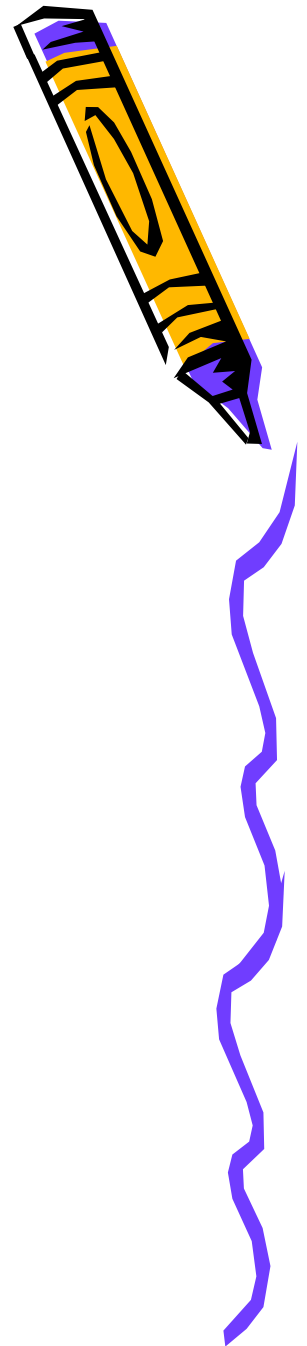
$$U(a, b) = \sum_{i=0}^n (f(x_i) - g(x_i))^2 = \sum_{i=0}^n (y_i - a - bx_i)^2$$





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





Rezolvând sistemul

$$\left\{ \begin{array}{l} \frac{\partial U}{\partial a} = -2 \sum_{i=0}^n (y_i - a - bx_i) = 0 \\ \frac{\partial U}{\partial b} = -2 \sum_{i=0}^n (y_i - a - bx_i) \cdot x_i = 0 \end{array} \right.$$





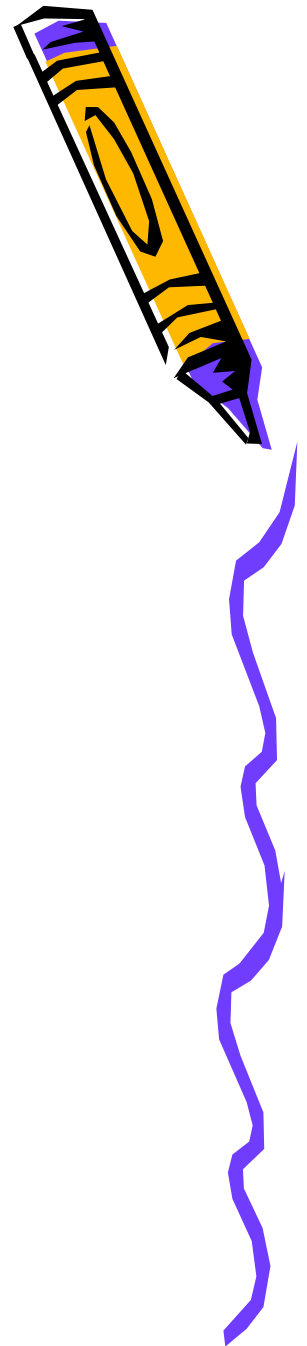
sistem echivalent cu:

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

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  $y = a_0 + b_0x$  liniarizează optim datele experimentale  $(x_i, y_i)$ ,  $0 \leq i \leq n$  și se numește *dreaptă de regresie* a lui  $y$  în raport cu  $x$ .

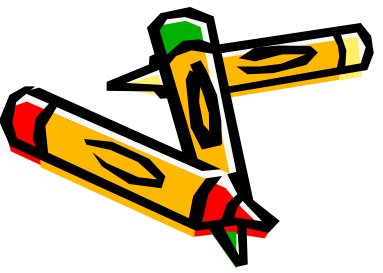
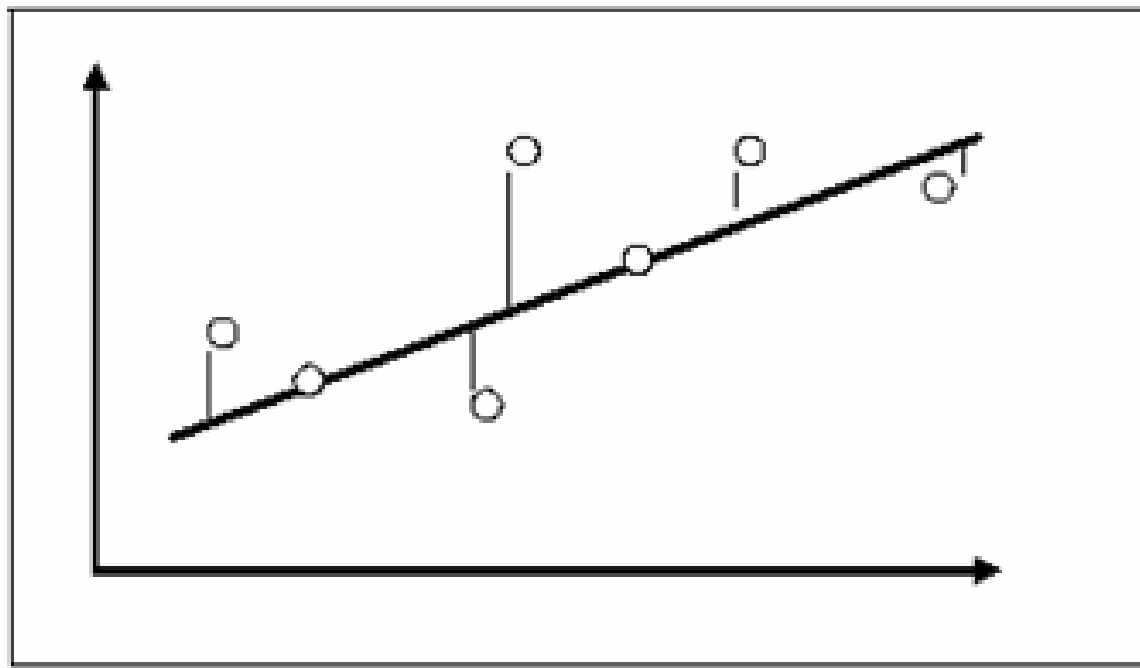
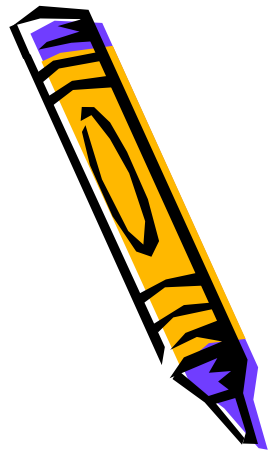




În cazul nostru, metoda constă în calcularea distanțelor (pe verticală) dintre punctele observate (reale) corespunzătoare cuplului de serii statistice și punctele (imaginare) de pe o anumită dreaptă (de regresie), ce trece prin mijlocul „norului” de puncte generate de cuplurile de date.

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 și care minimizează, distanța între date și ea (minimizând suma pătratelor distanțelor).

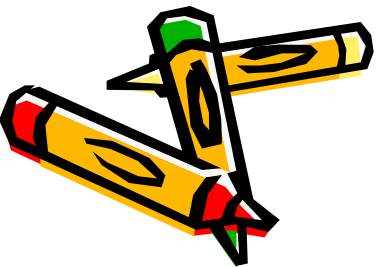




În final, obținem ecuația de regresie sub forma:

$$Y = a + bX$$

unde  $a$  se numește *interceptor* iar  $b$  se numește *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:

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





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

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

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

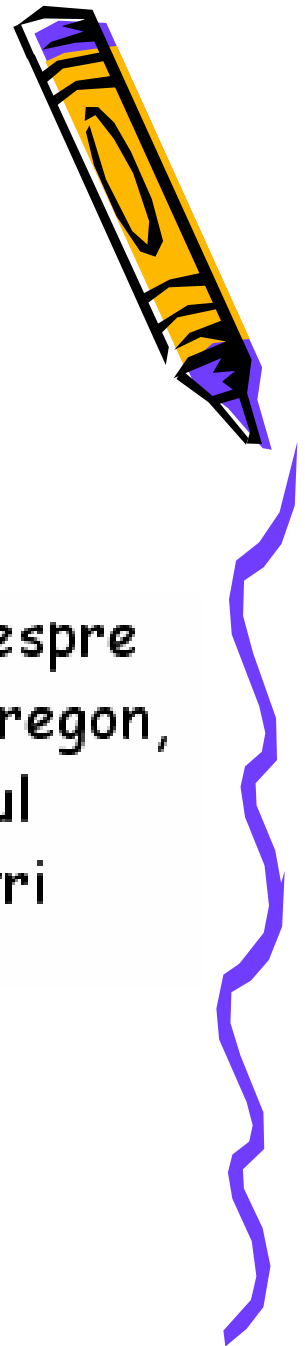
de unde obținem:

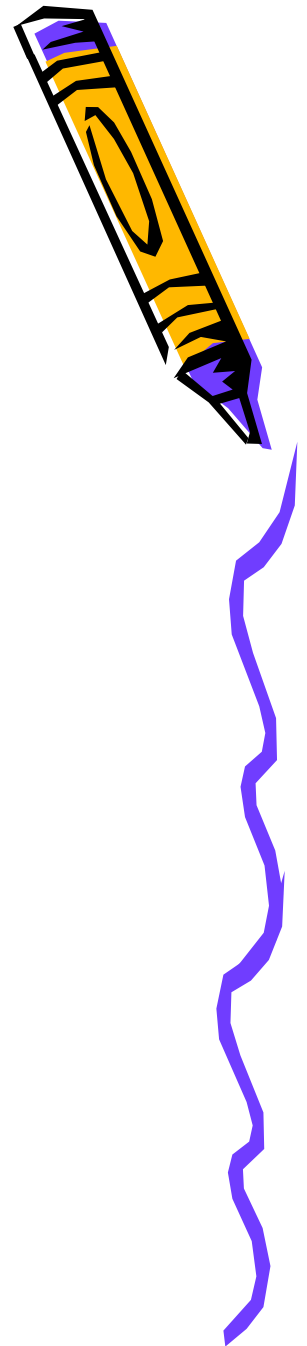
$$b = \frac{\sigma_{XY}}{\sigma_{XX}} .$$



# 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 pătrați)  $X$ .





	X = suprafața (mii de m <sup>2</sup> )	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

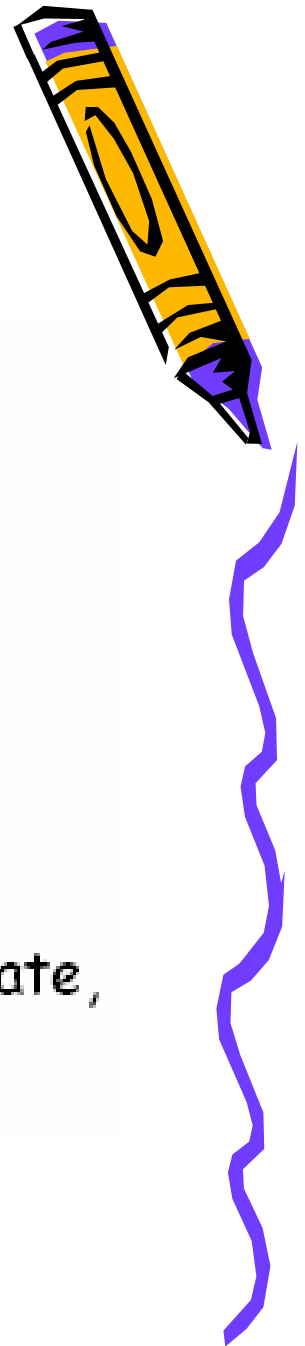


```
»X=[1.683 1.708 1.922 2.053 2.269];  
»Y=[259.9 259.9 269.9 270.0 285.0];  
»A=[X' Y'];  
» corrcoef(A)
```

```
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 și o desenăm în același cadran cu diagrama împrăștierii:

```
»b=(X*Y'-5*mean(X)*mean(Y))/((norm(X))^2-5*(mean(X))^2)
```

```
b =
```

```
40.8001
```

```
»a=mean(Y)-b*mean(X)
```

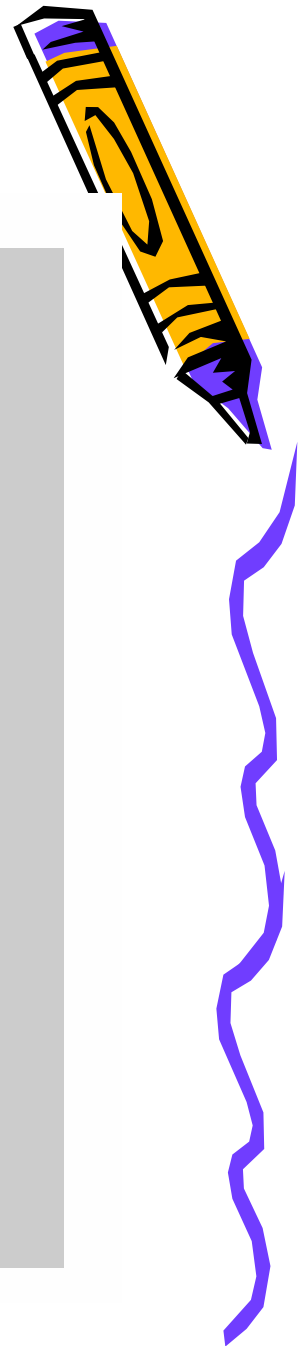
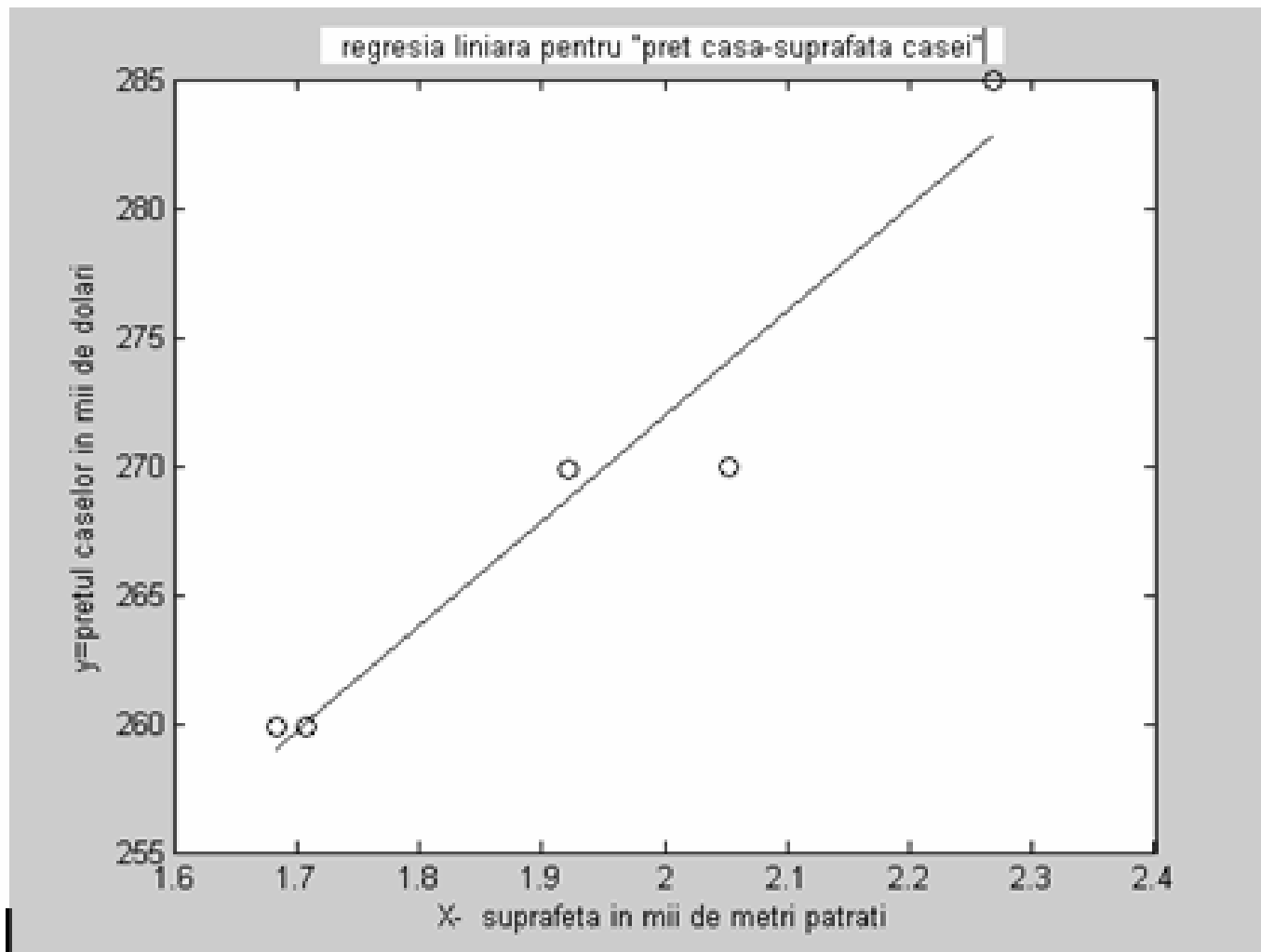
```
a =
```

```
190.3182
```

```
»plot(X,a+b*X);hold on
```

```
» plot(X,Y,'ko');hold off
```





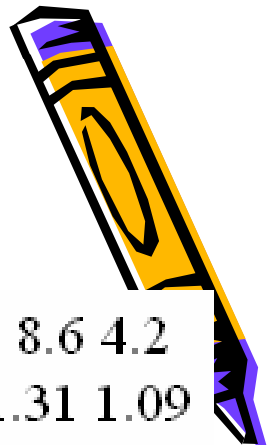


Am prezentat date culese de la un lot de 24 pacienți, având diabet zaharat, privind două variabile:

- $G$  (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
```





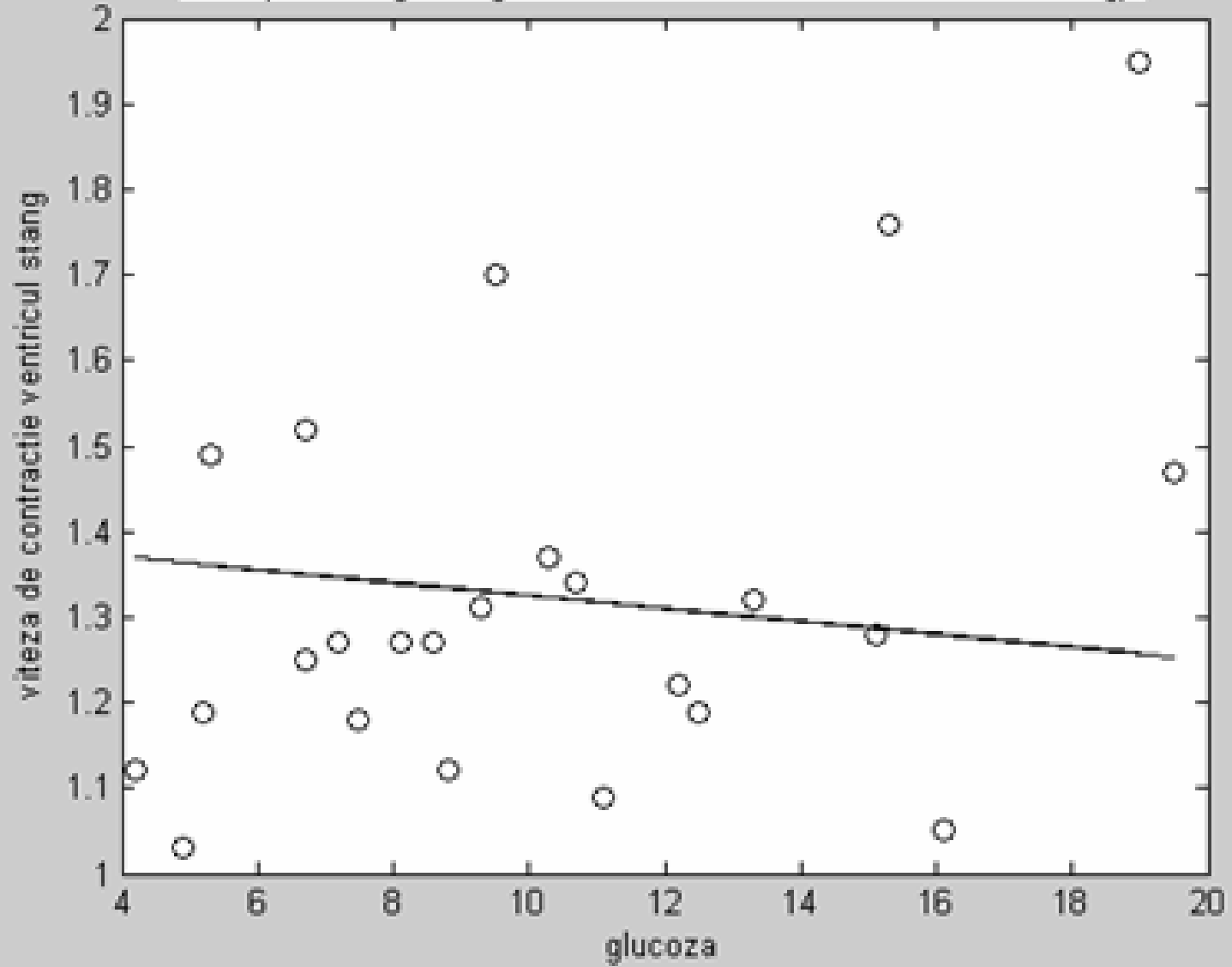
Să desenăm dreapta de regresie și diagrama împrăștierii:

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





dreapta de regresie glucoza/ viteza de contractie ventricul stang



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

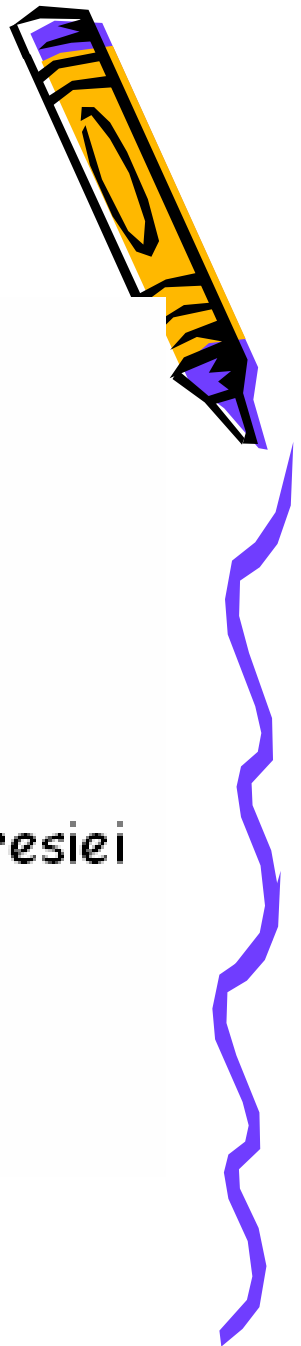
Presupunem că valoarea glucozei în sânge pentru un pacient este 15.1.

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

$$\gg V1 = a + b \cdot 15.1$$

$$V1 =$$

$$1.2871$$





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.

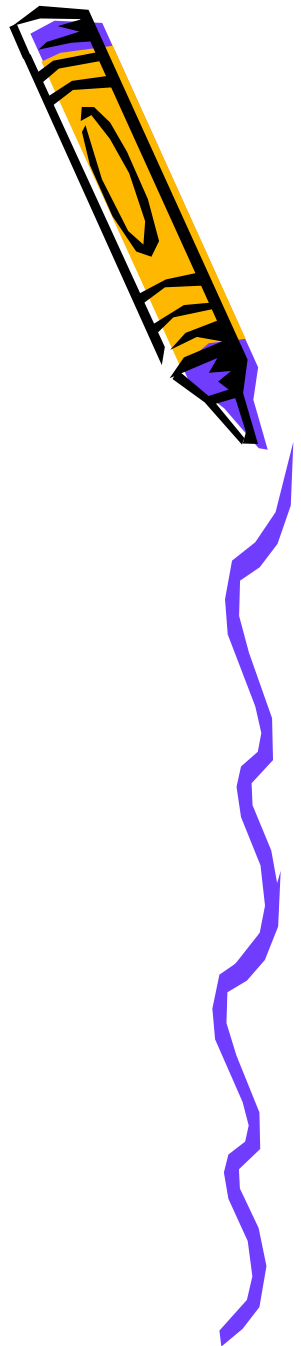




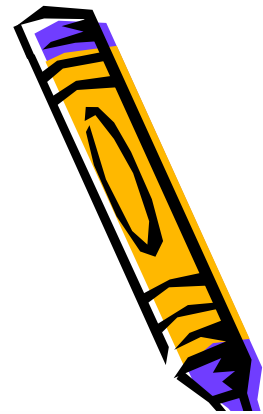
Dacă, de exemplu, cele două drepte de regresie au aproximativ aceeași pantă (sunt paralele), atunci putem considera diferența pe axa verticală ( $Y$ ) ca fiind diferența între mediile variabilei  $Y$  între cele două grupuri, observație ce este apoi urmată de o testare a semnificației statistice a diferenței. O asemenea analiză statistică face parte dintr-un studiu statistic mai vast, care se numește *analiza covarianțelor*.



# Regresia polinomială



# 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:

$$Y = a + b_1X + b_2X^2 + \dots + b_nX^n,$$

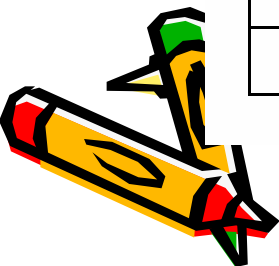
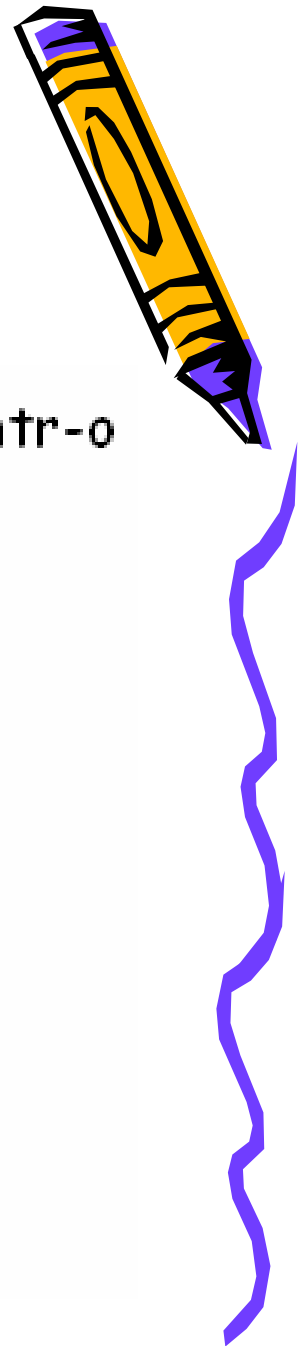
unde, ca de obicei,  $X$  este variabila predictoare iar  $Y$  este variabila prognozată, adică variabila răspuns.



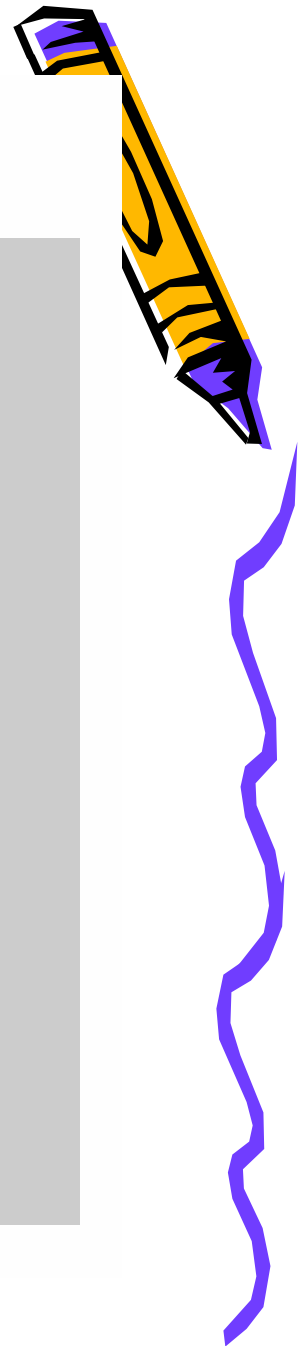
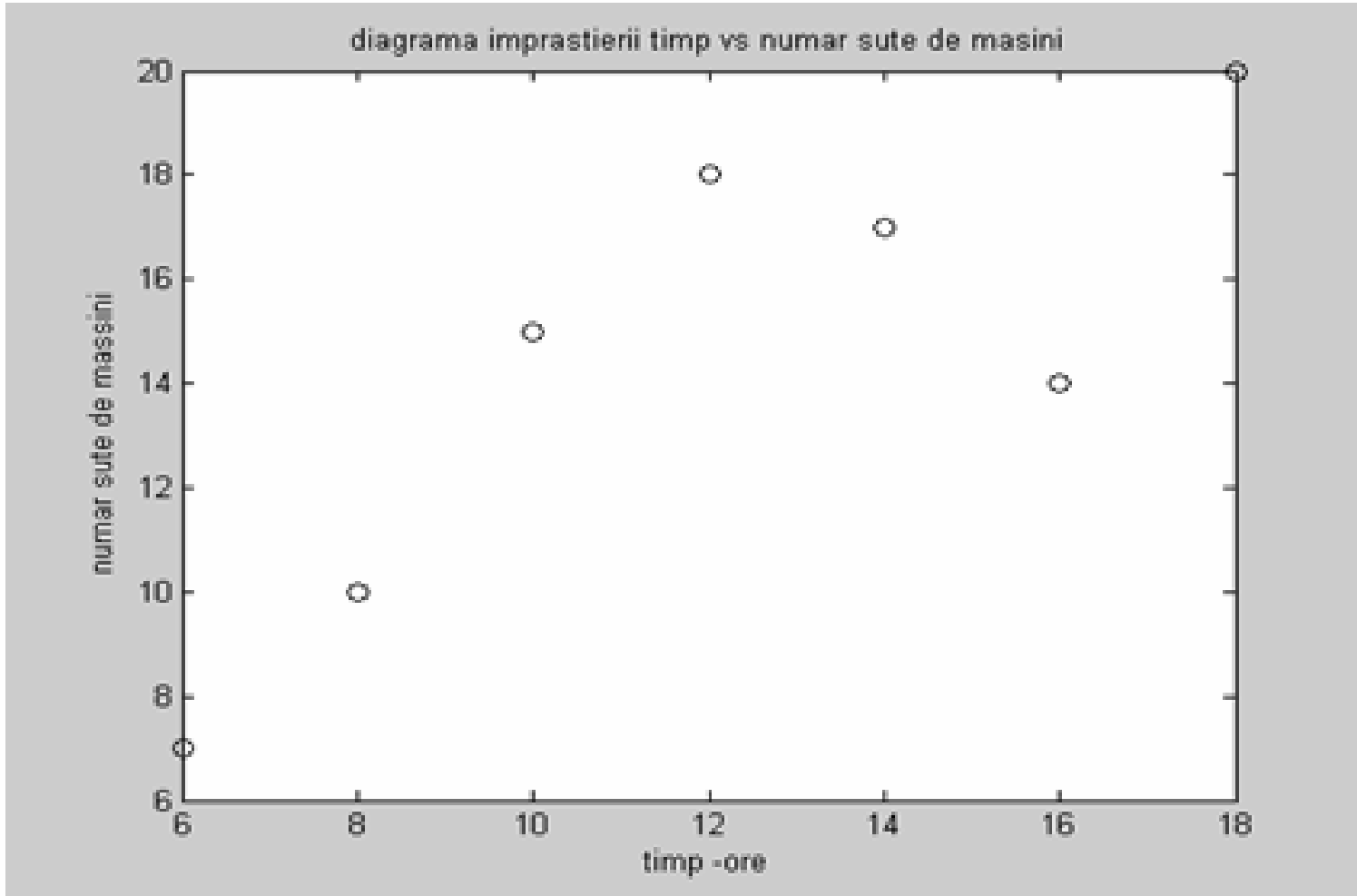
# exemple

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')
```



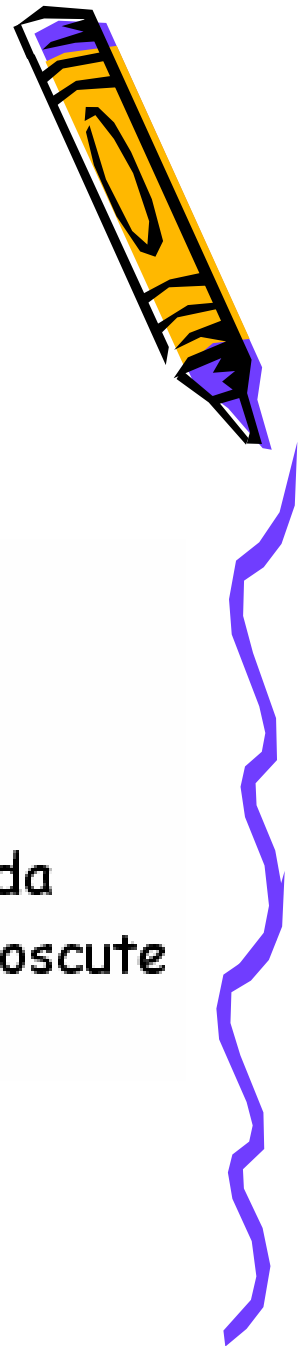


Diagrama împrăștierei sugerează că datele ar putea fi modelate ca o funcție polinomială

$$y = a_0 + a_1t + a_2t^2$$

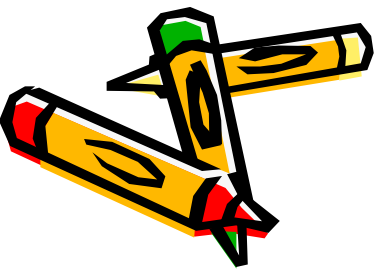
Coeficienții necunoscuți vor fi calculați folosind metoda celor mai mici pătrate. Vom rezolva sistemul de necunoscute

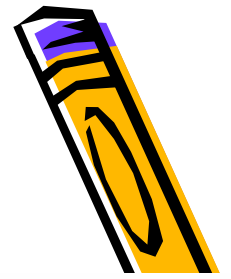
$a_0, a_1, a_2$ :





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





```
» 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 mașini vor trece prin intersecție  
în intervalul 21h-23h.

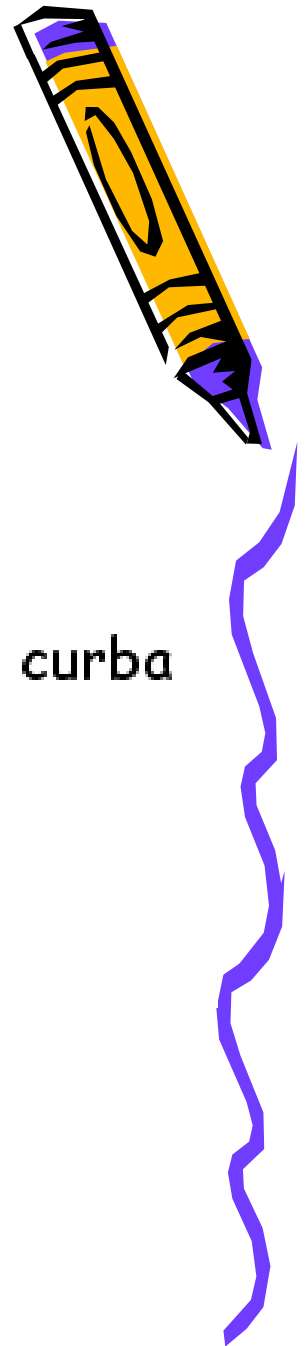
» `syms t`

» `f23=subs(f,t,23)`

`f23 =`

`13.7411`

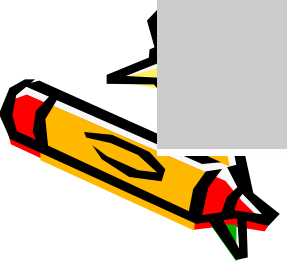
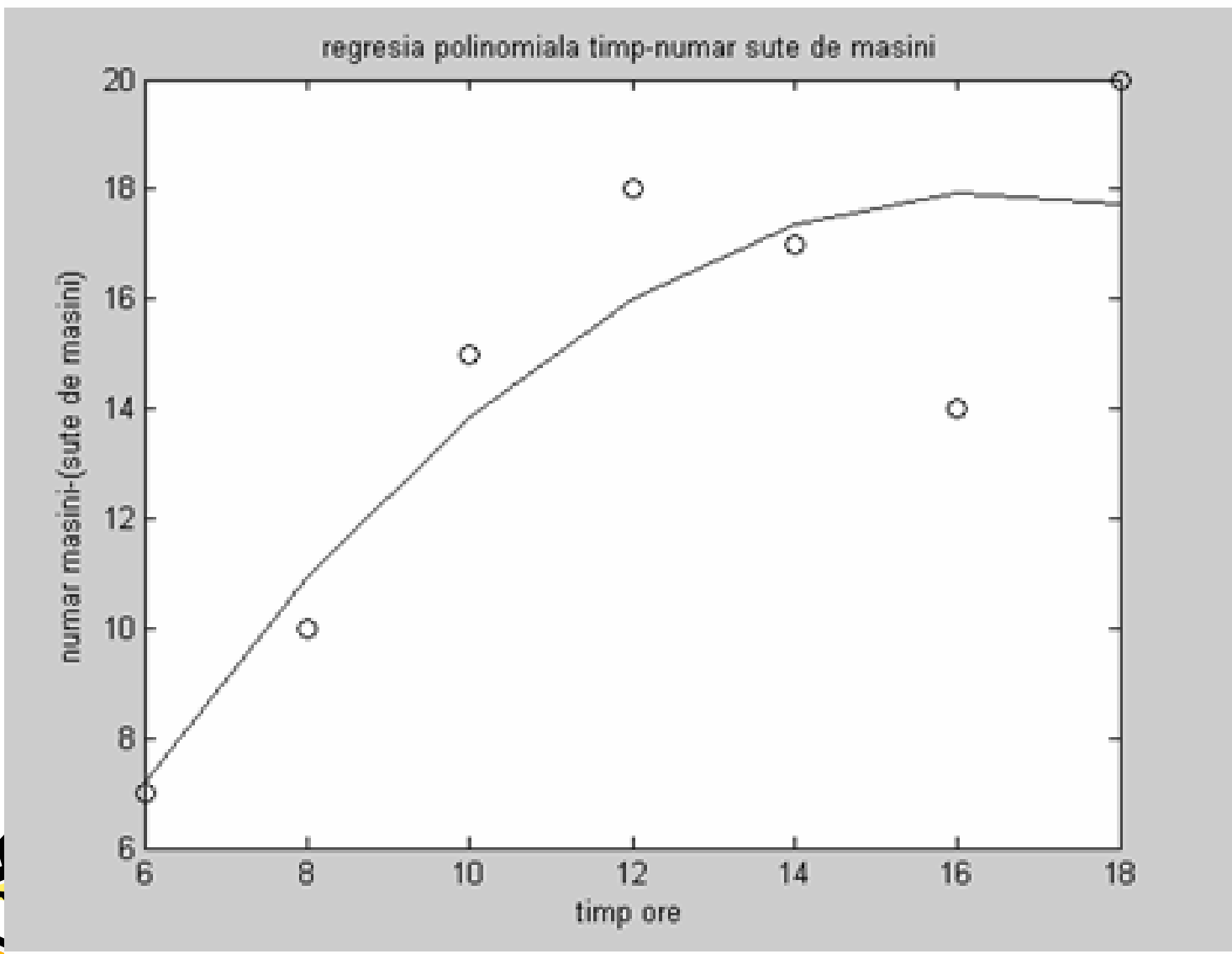




În final desenăm atât diagrama împrăștierei, cât și curba de regresie polinomială în același cadran:

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







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:





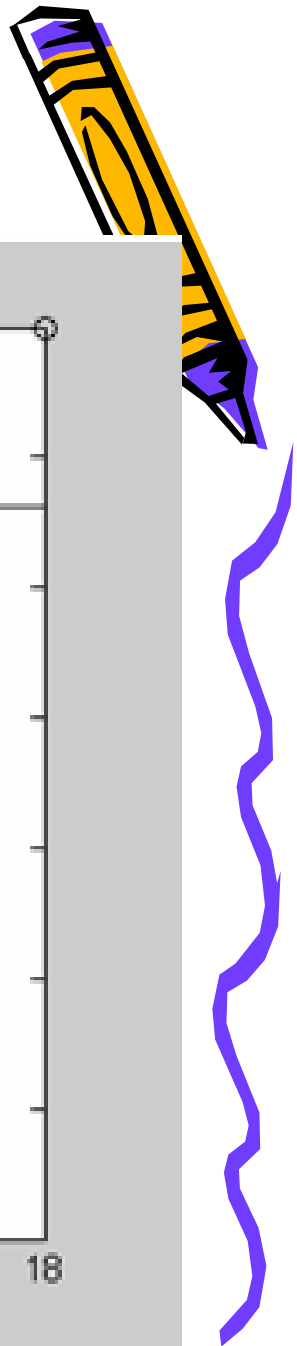
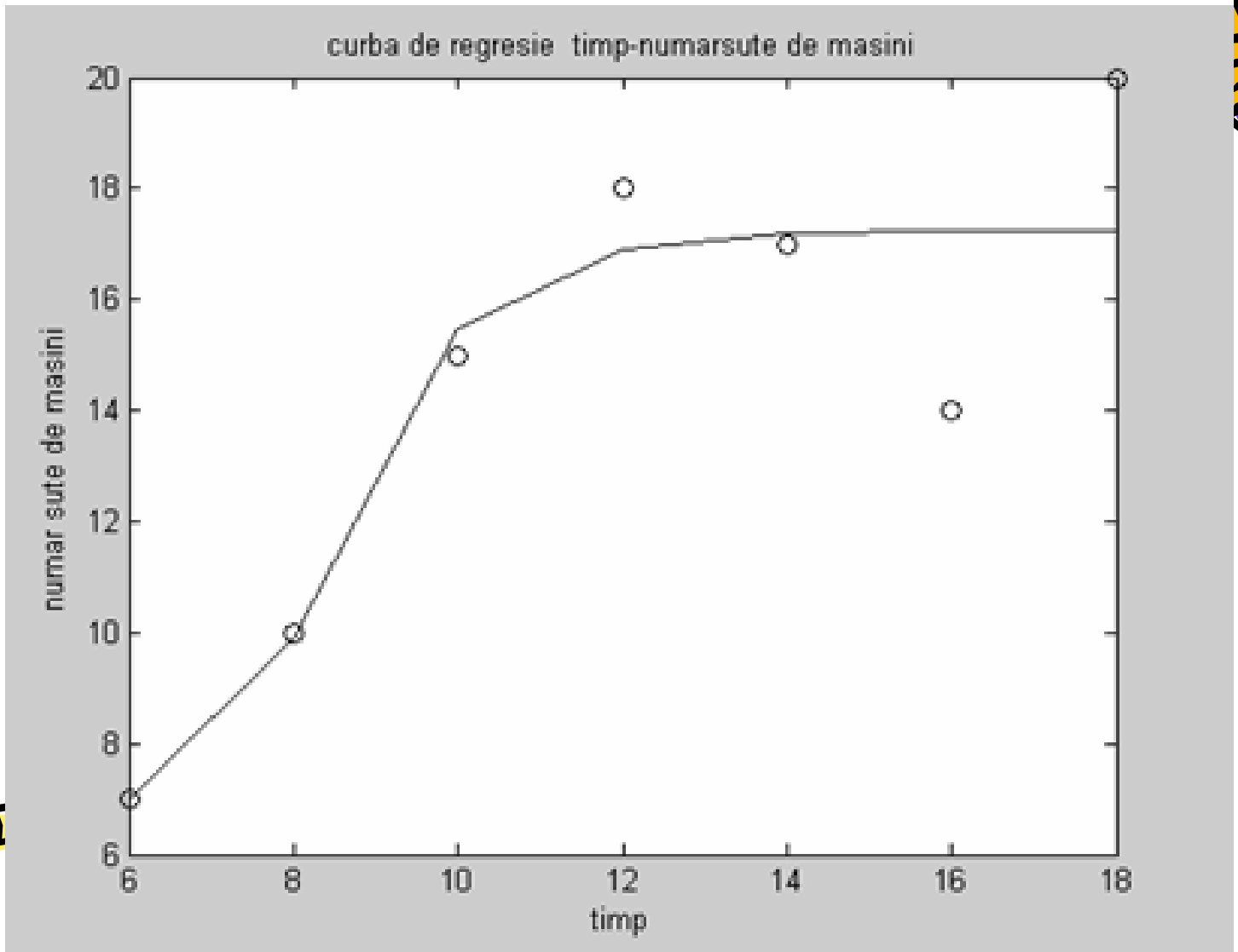
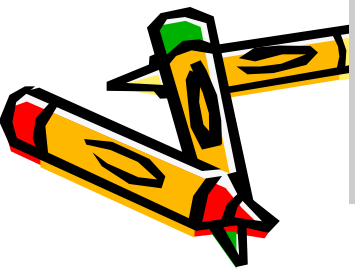
```
» t=6:2:18;n=[7 10 15 18 17 14 20];  
» A=[1 exp(-t(1)) t(1)^*exp(-t(1));1 exp(-t(2)) t(2)^*exp(-t(2));  
1 exp(-t(3)) t(3)^*exp(-t(3));1 exp(-t(4)) t(4)^*exp(-t(4));  
1 exp(-t(5)) t(5)^*exp(-t(5));1 exp(-t(6)) t(6)^*exp(-t(6));  
| 1 exp(-t(7)) t(7)^*exp(-t(7))];  
» a=A\n'
```

a =

```
1.0e+004 *  
0.0017  
4.9000  
-0.8856
```

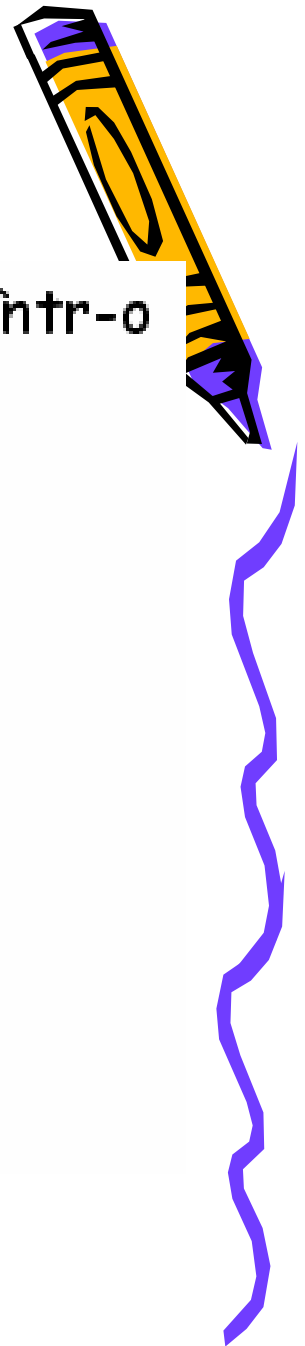
```
» plot(t,a(1)+a(2)^*exp(-t)+a(3)^*t.^*exp(-t));hold on  
» plot(t,n,'kO');hold off
```

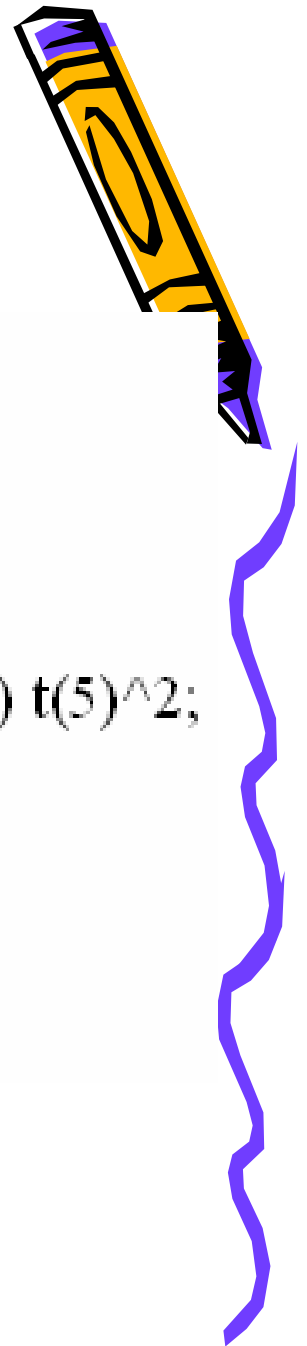




Considerăm situația numărului de pacienți internați într-o secție a unui spital pe parcursul a șase luni, situație reflectată în următorul tabel.

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



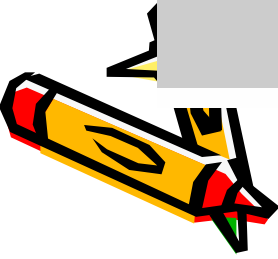
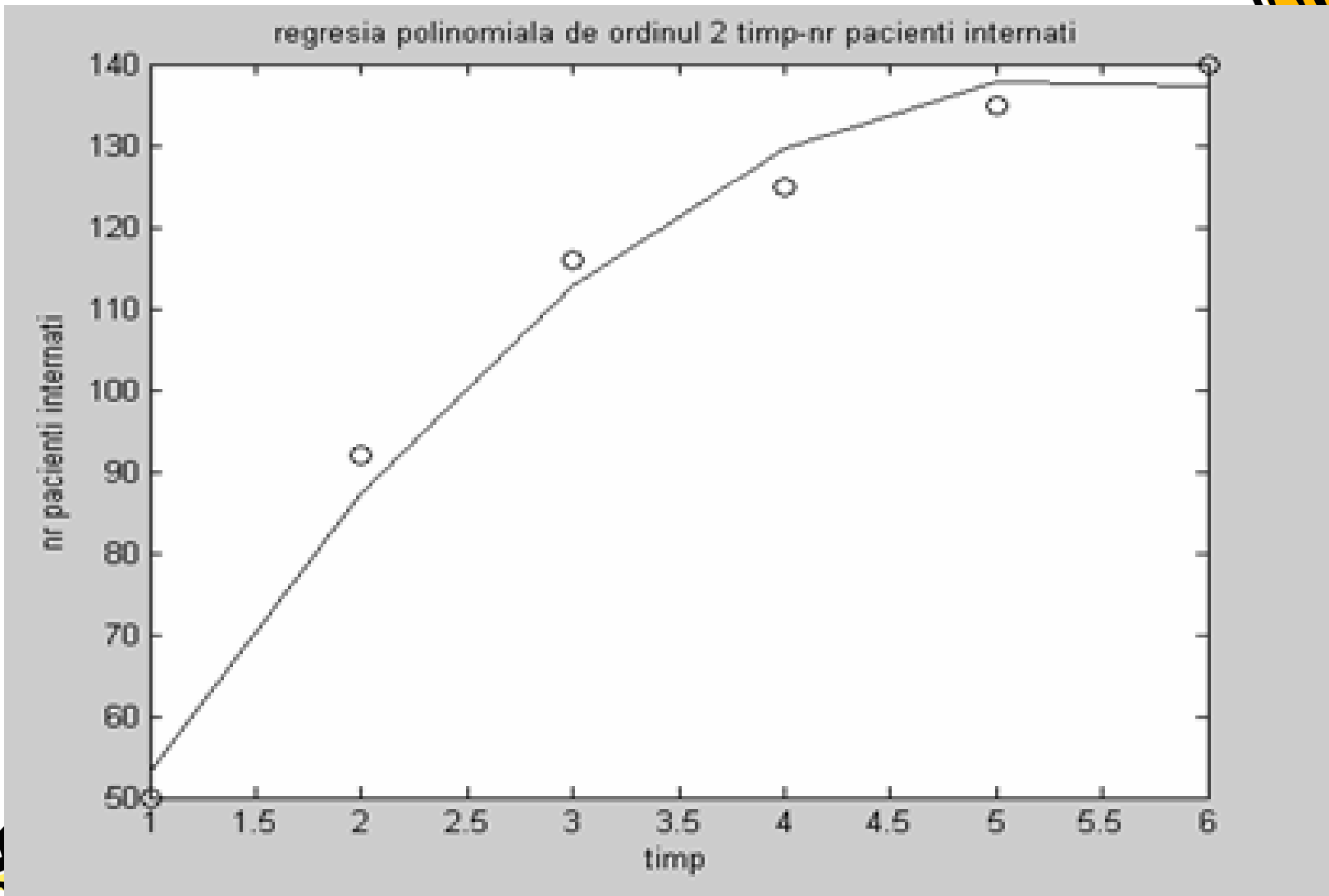
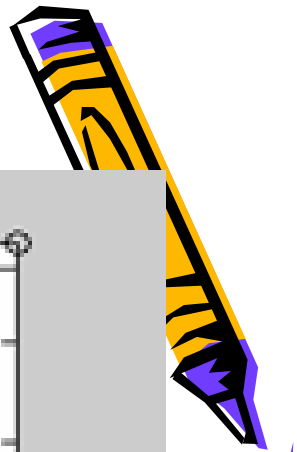


Pentru început, vom determina diagrama împrăștierei și curba de regresie polinomială:

- » `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`





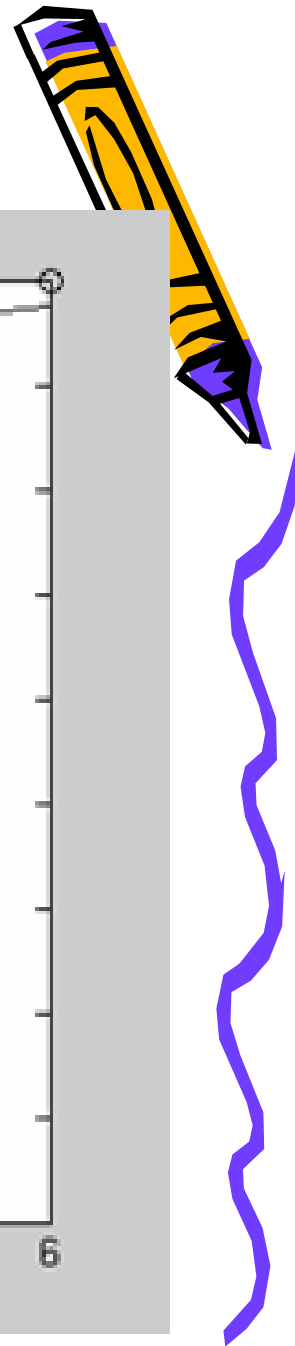
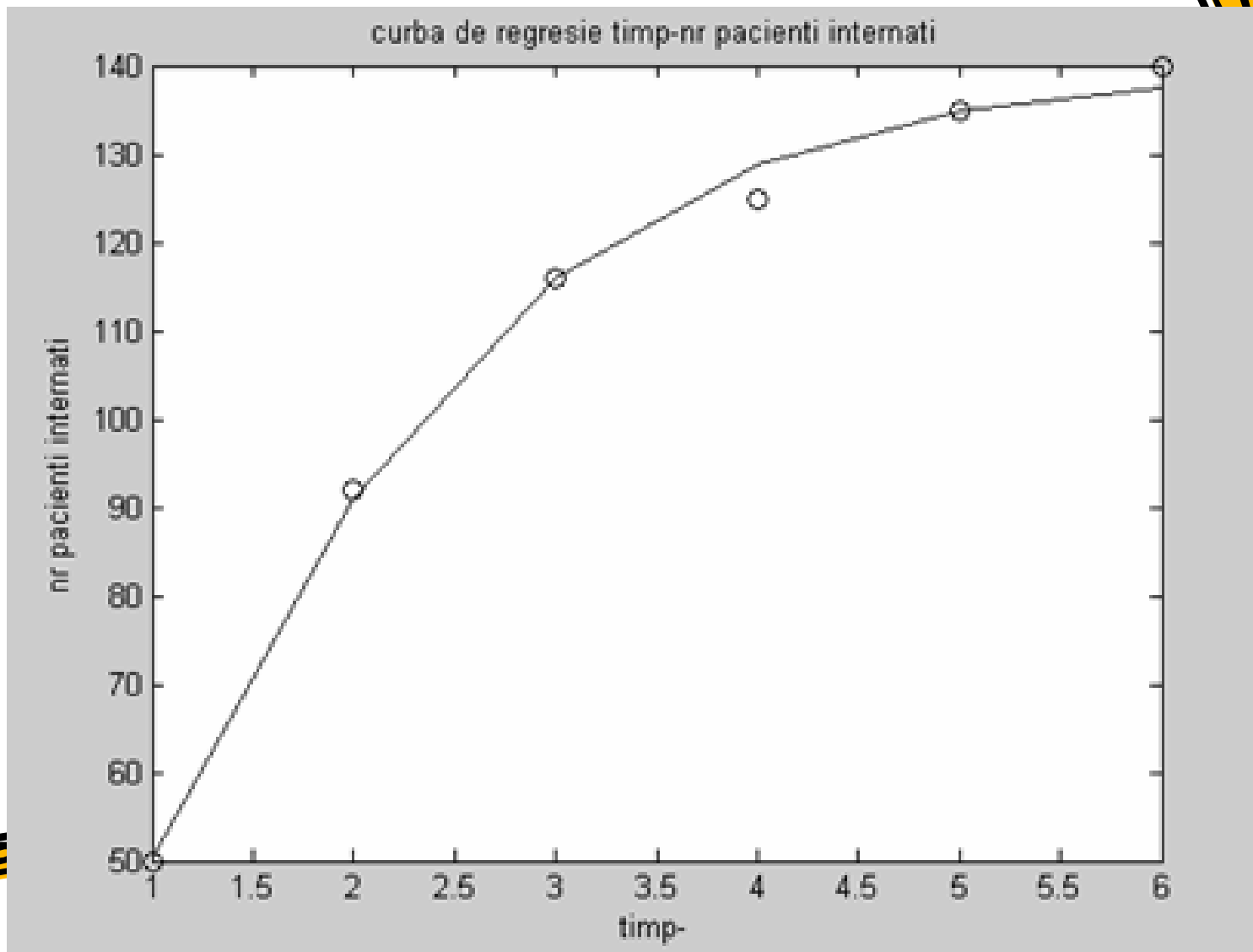




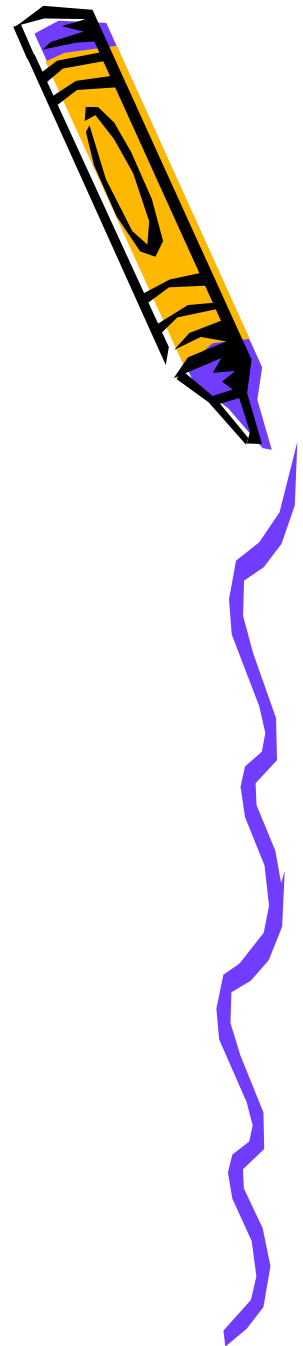
Rezolvăm aceeași problemă utilizând „linear in the parameters regression“:

- » `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`





# Regresia liniară multiplă



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  $X_1, X_2, \dots, X_n$ , 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 ș.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 și 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.





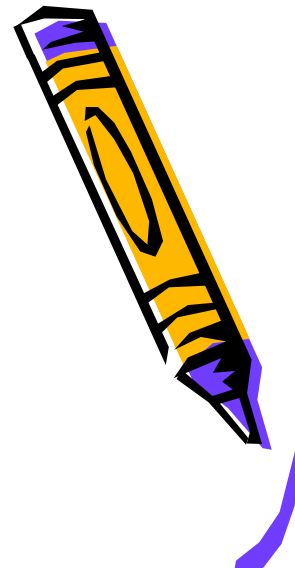
Deducând ecuația de regresie pe baza datelor culese de la un lot de pacienți cu cancer, se poate prognoza timpul pe care îl mai are de trăit un anumit pacient nou, introducând în ecuație valorile particulare ale acestuia pentru fiecare variabilă predictivă în parte.





Disponem, pe de o parte, de  $n$  variabile predictive sau explicative și, pe de altă parte, de o variabilă prognozată, explicată, numită și 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  $X_1, X_2, \dots, X_n$  și  $Y$  trebuie să verifice următoarele două condiții:





- legătura între variabila răspuns  $Y$  și variabilele predictive  $X_i$  trebuie să fie una liniară, așa cum arată și 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:

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

unde  $b_1, b_2, \dots, b_n$  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”.



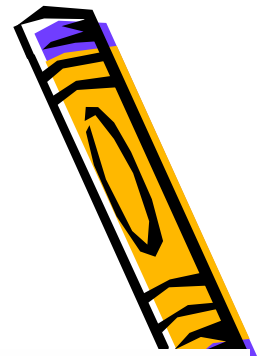
# exemplu



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 anumită diferență de preț pentru loturile de 4000 m<sup>2</sup> și cele de 6000 m<sup>2</sup> și altă diferență de preț între loturile de 24000 m<sup>2</sup> și cele de 26000 m<sup>2</sup>.







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 <sup>2</sup> )	categoria lotului (mii de m <sup>2</sup> )	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 și categoria lotului nu implică independența variabilelor *suprafața casei și categoria lotului*.

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





»  $B = [1 \ x1(1) \ x2(1); 1 \ x1(2) \ x2(2); 1 \ x1(3) \ x2(3); 1 \ x1(4) \ x2(4); 1 \ x1(5) \ x2(5);$   
 $1 \ x1(6) \ x2(6)]$

»  $a = B \setminus y'$

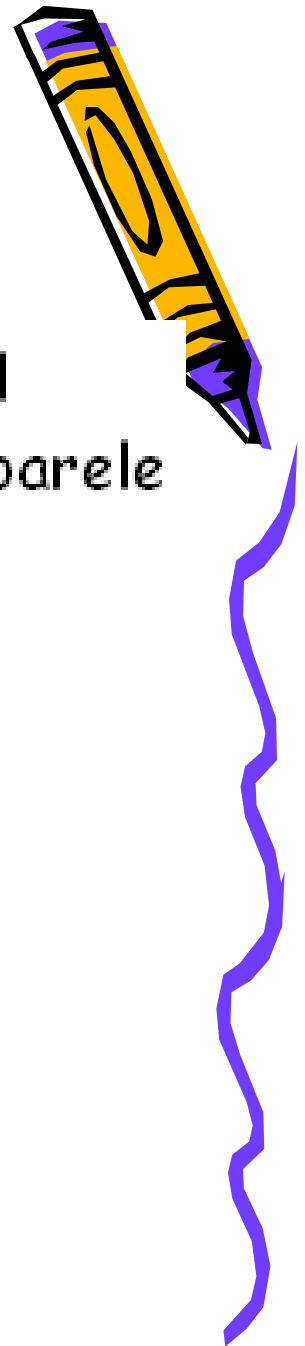
$a =$

122.3570

61.9756

7.0911

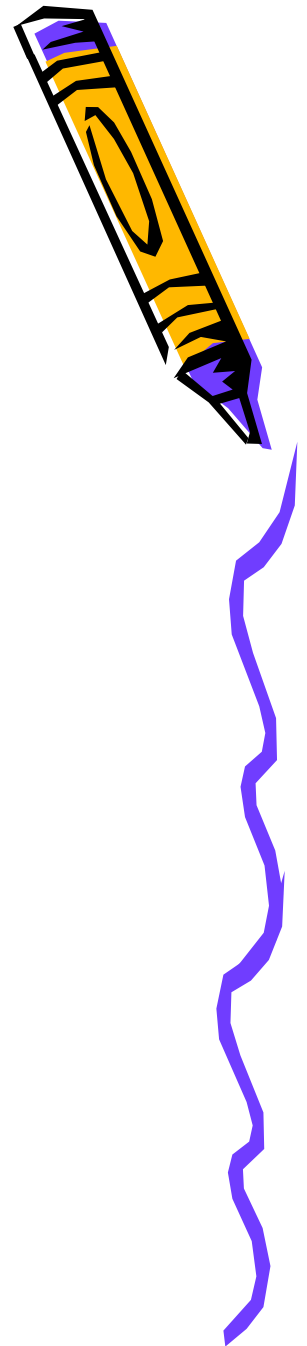




Folosind rezultatul obținut, putem prognoza prețul proprietăților, de exemplu să îl calculăm în următoarele cazuri:

- suprafața casei este de  $1805 \text{ m}^2$  și lotul de casă este din categoria 2;
- suprafața casei este de  $1200 \text{ m}^2$  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
```

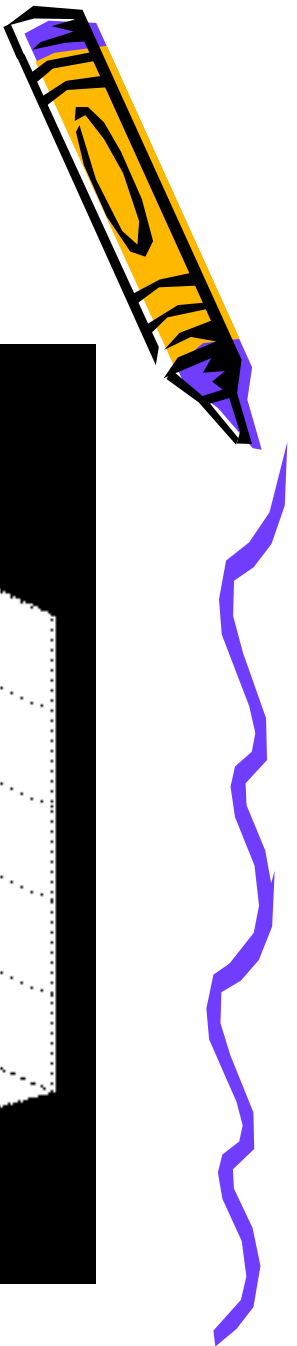
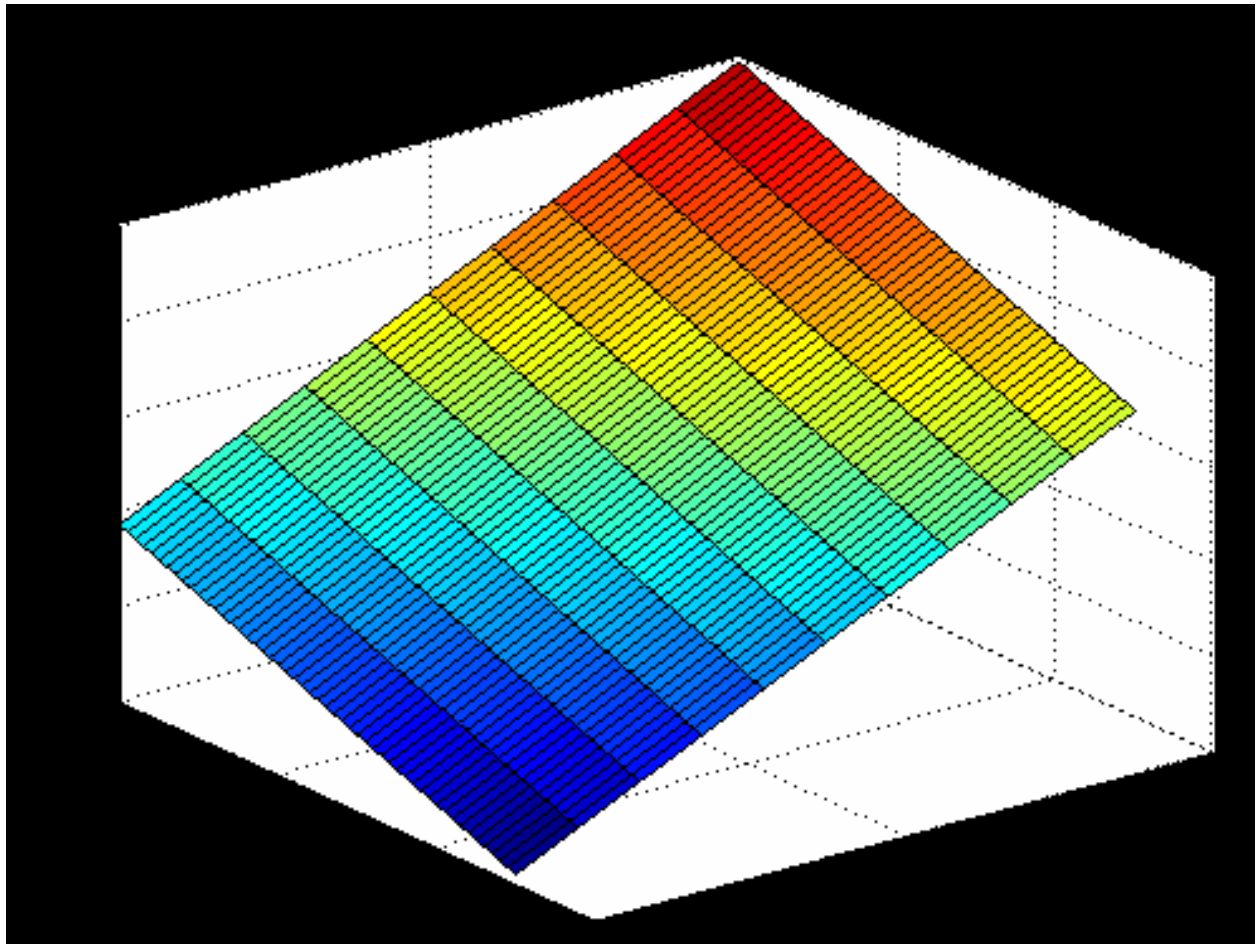




```
» [x1,x2]=meshgrid(1.5:.1:2.5,1:.1:6);  
» f=a(1)+a(2)*x1+a(3)*x2;  
» surf(x1,x2,f)
```







În cazul regresiei liniare multiple apare problema ierarhizării predictorilor în scopul păstrării numai a aceluia 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" directă



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 și 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" inversă

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.

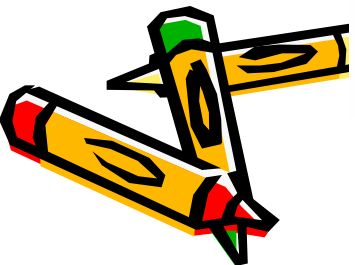


# exemplu

Analiza regresivă multiplă utilizată în predicția *indexului* de rezistență a mușchiului respirator  $PE_{max}$  (exprimat în  $cm H_2O$ )

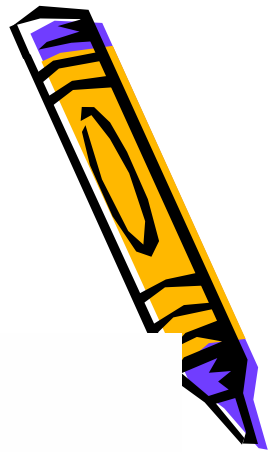
în funcție de variabilele predictoare:

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



- volumul respirator 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 ( $PE_{max}$ ).



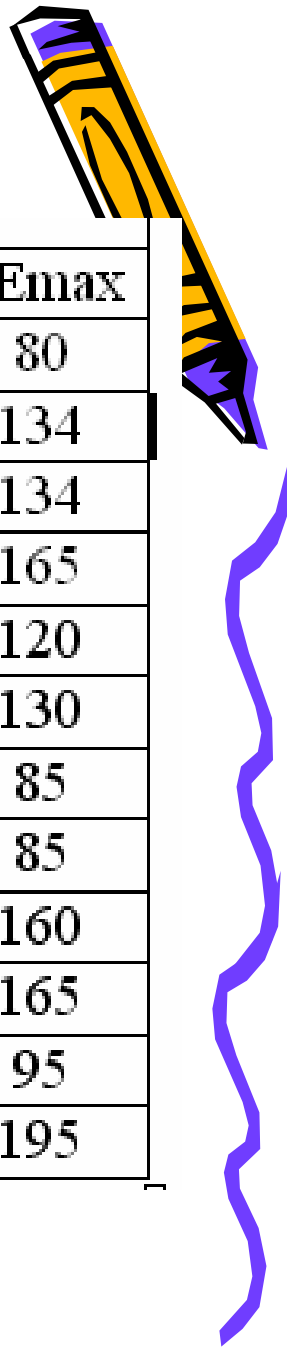


F

Vârsta	Sex	H	G	BMP	FEV <sub>1</sub>	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 <sub>1</sub>	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" directă -exemplu



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	<b>0.6329</b>
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  $P_{\max}$ .

( $p = 0.0001$ ).

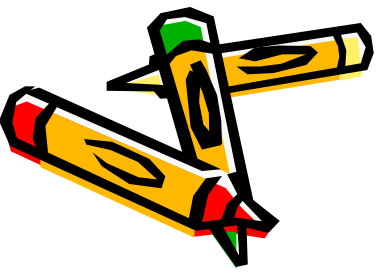
Calculăm:

- regresia simplă, adică  $P_{\max}$  funcție liniară de  $G$
- reziduurile modelului (diferența între valoarea prognozată și cea reală):





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







Alegem, dintre variabilele independente rămase, aceea care are cea mai mare corelație (ignorând semnul) cu reziduurile modelului determinat ( $R_1$ )

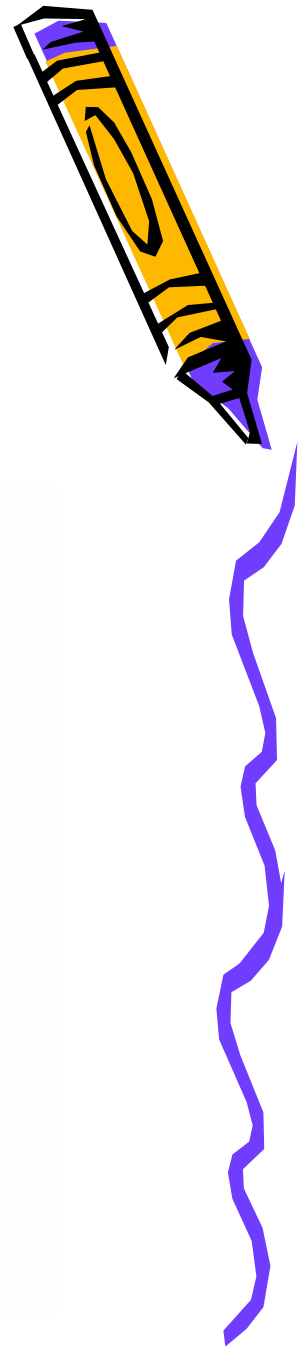
- »  $A1=[V' \ s' \ H' \ \text{BMP}' \ \text{FEV}' \ \text{RV}' \ \text{FRC}' \ \text{TLC}' \ \text{R1}']$ ;
- » `corrcoef(A1)`





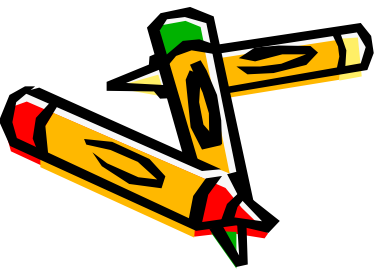
```
ans =  
Columns 1 through 5  
1.0000 -0.1671 0.9261 0.3778 0.1982  
-0.1671 1.0000 -0.1675 -0.1376 -0.4461  
0.9261 -0.1675 1.0000 0.4408 0.2389  
0.3778 -0.1376 0.4408 1.0000 0.4394  
0.1982 -0.4461 0.2389 0.4394 1.0000  
-0.5519 0.2714 -0.5695 -0.5824 -0.6988  
-0.6394 0.1836 -0.6243 -0.4344 -0.6848  
-0.4694 0.0242 -0.4571 -0.3649 -0.3936  
0.0517 -0.2177 0.0215 -0.2532 0.1328
```

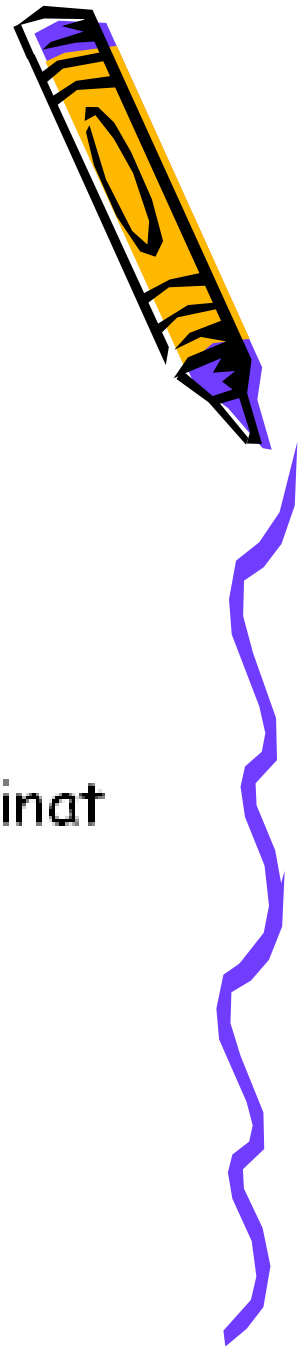




### Columns 6 through 9

-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	<b>-0.2532</b>
-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:



```

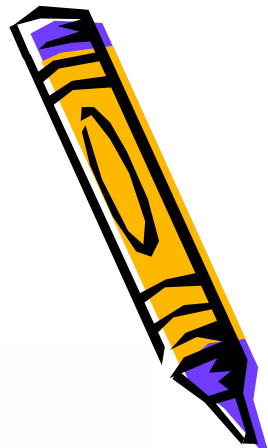
»B2=[1 G(1) BMP(1);1 G(2) BMP(2);1 G(3) BMP(3);
1 G(4) BMP(4);1 G(5) BMP(5);1 G(6) BMP(6);
1 G(7) BMP(7);1 G(8) BMP(8);1 G(9) BMP(9);
1 G(10) BMP(10);1 G(11) BMP(11);1 G(12) BMP(12);
1 G(13) BMP(13);1 G(14) BMP(14) ;1 G(15) BMP(15);
1 G(16) BMP(16);1 G(17) BMP(17);1 G(18) BMP(18);
1 G(19) BMP(19);1 G(20) 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'];
» corcoef(A2)

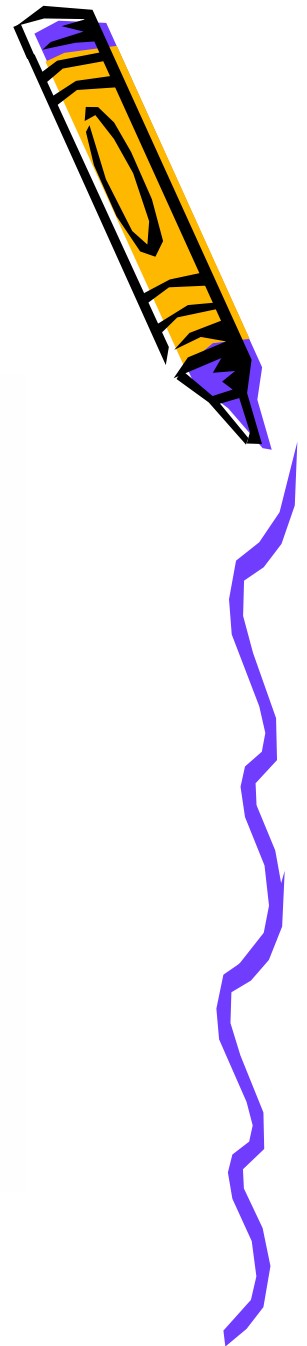
```



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

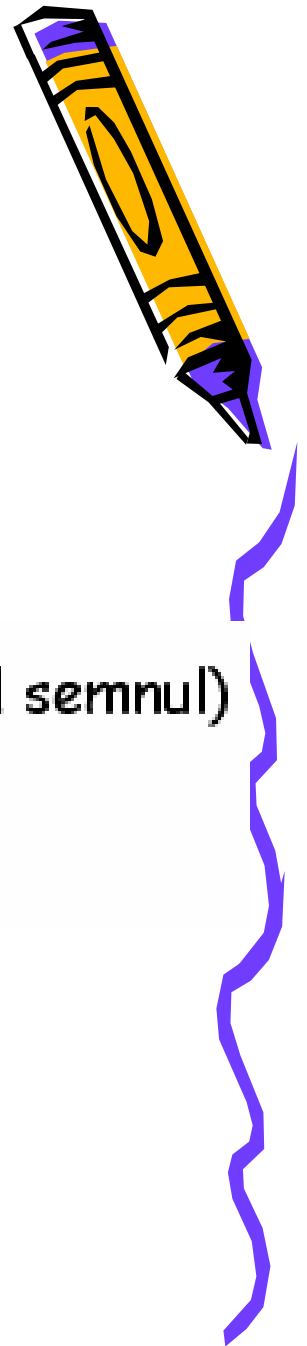




### Columns 6 through 8

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



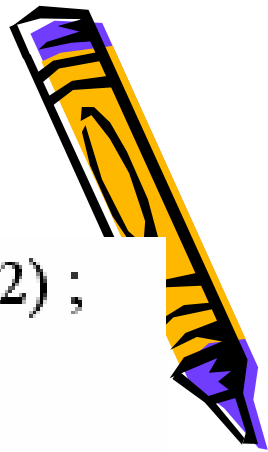


Predictorul FEV are cea mai mare corelație (ignorând semnul) cu reziduurile modelului determinat ( $R^2$ ) ( $p = 0.0210$ ).  
Repetăm procedeul:





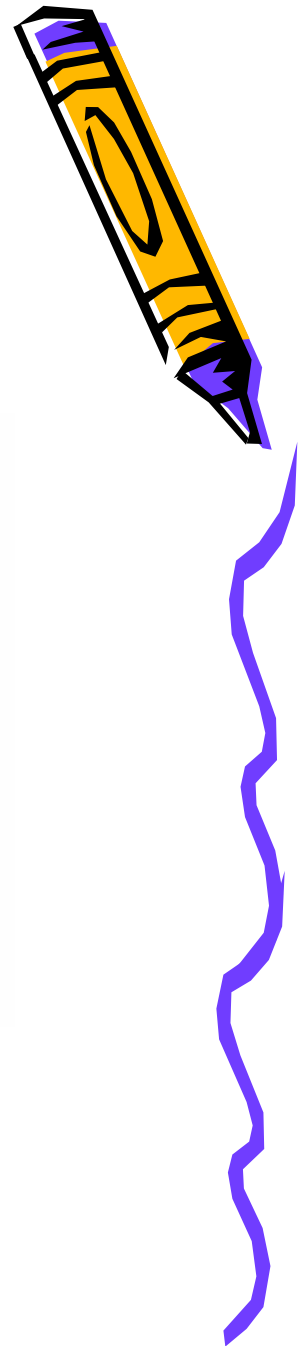
»B3=[1 G(1) BMP(1) FEV(1) ;1 G(2) BMP(2) FEV(2) ;  
1 G(3) BMP(3) FEV(3) ;1 G(4) BMP(4) FEV(4) ;  
1 G(5) BMP(5) FEV(5) ;1 G(6) BMP(6) FEV(6) ;  
1 G(7) BMP(7) FEV(7) ;1 G(8) BMP(8) FEV(8) ;  
1 G(9) BMP(9) FEV(9);1 G(10) BMP(10) FEV(10) ;  
1 G(11) BMP(11) FEV(11) ;1 G(12) BMP(12) FEV(12) ;  
1 G(13) 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)
```

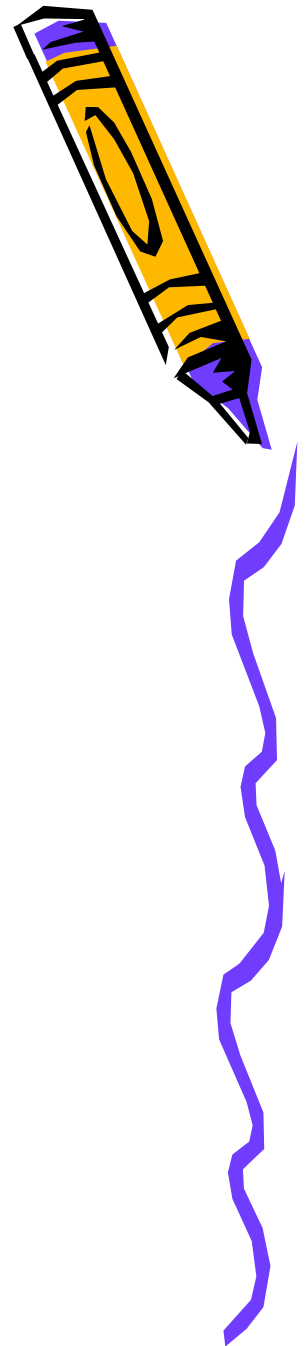




Columns 1 through 5

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:

$$PE_{\max} = 120.8813 + 1.6093 G - 1.2540 BMP + 0.7150 FEV$$



# "regresia pas cu pas" inversă



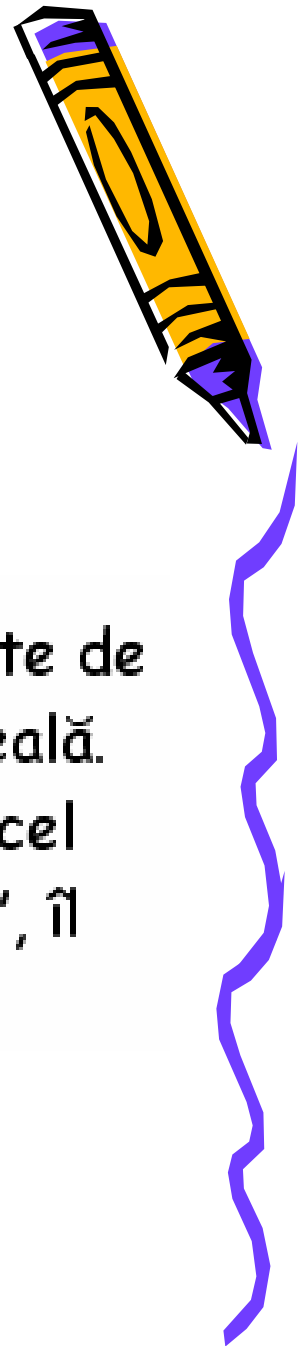
Datele utilizate la construcția modelului „regresia pas cu pas inversă”, sunt prezentate în tabelul de mai jos.

	estimare	eroare standard (b)	t-test	p
<u>Interceptor (b<sub>0</sub>)</u>	63.54564	12.70163	5.002952	0.000046
G	1.18671	0.30086	3.944453	0.000646

Ecuția de regresie este:

$$\text{PE}_{\text{max}} = 63.55 + 1.19 \cdot G.$$





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.

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

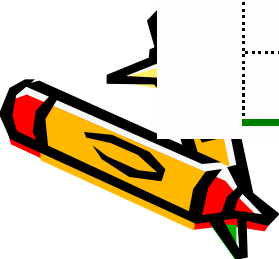


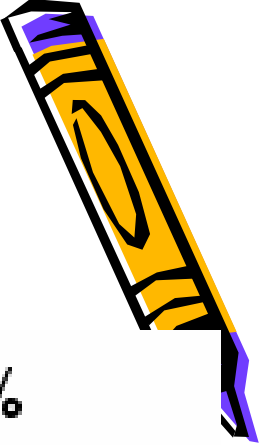




vom lua în considerație toate cele nouă variabile explicative. Rezultatul aplicării acestei metode este tabelul următor.

	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

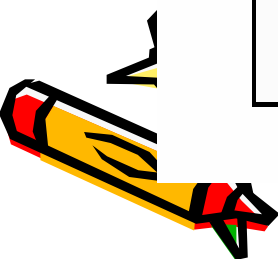
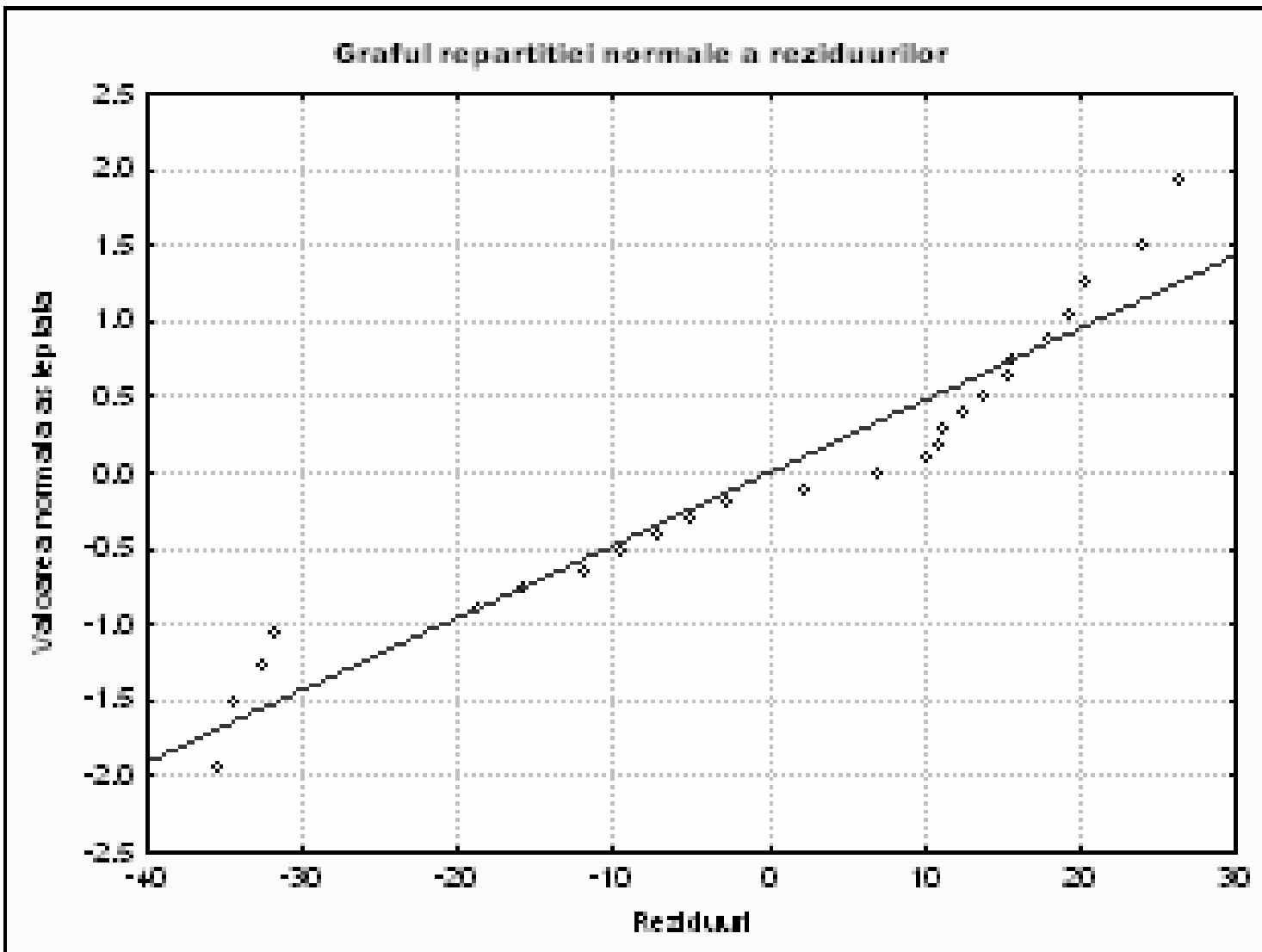




Nicio variabilă predictivă nu trece pragul de 5% al nivelului de semnificație  $p$ .

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.







Ecuția de regresie corespunzătoare este dată de formula:

$$PE = 137.23 - 1.79 \text{ BMP} + 2.87 \text{ G} + 1.49 \text{ FEV}_1 + 0.17 \text{ RV} - 2.47 \text{ V} - 0.3 \text{ H} - 0.16 \text{ FRC} + 0.11 \text{ TLC} - 1.38 \text{ S}$$





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  $PE_{max}$ , prin introducerea în ecuația de regresie a valorilor sale individuale.

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



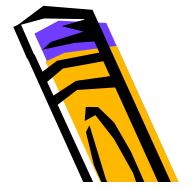
# exemple



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 codifica fiecare pacient ca un vector cu patru elemente,

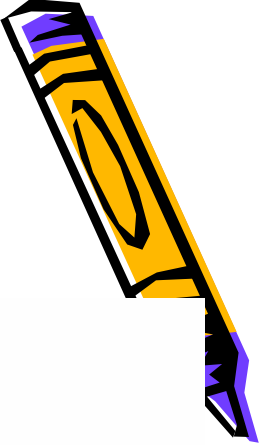
$$\mathbf{x} = (x_1, x_2, x_3, y).$$

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: I = 0,35, II = 0,3, III = 0,25, IV = 0,1),

- variabila  $Y$  -gaussiană de medie 36 și deviație standard 12.





Scopul analizei regresive efectuate:

de a stabili legătura între variabilele predictive  $X_1$ ,  $X_2$  și  $X_3$  și variabila dependentă  $Y$ .

Concret, analizăm legătura între factorii predictivi: diametrul mediu al tumorii, vârstă și stadiu și 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.



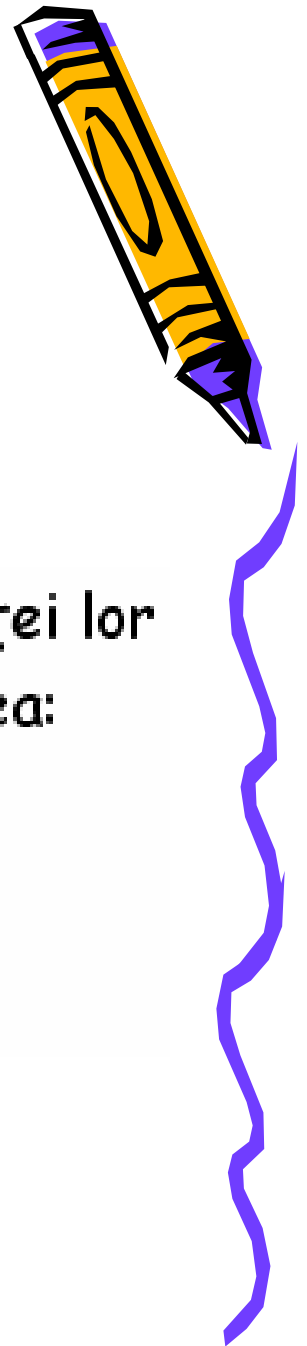
# rezultatele analizei regresive



	Beta standardizat	Coeficient b	Nivel sennificație p
Interceptor		24.425	0.00018
Diametru mediu	0.045	0.046	0.0647
Vârștă	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.





Ecuția de regresie liniară multiplă este dată de:

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

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 și este în stadiul III și 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)

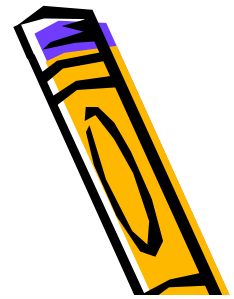




. Variabilele predictive în acest caz sunt următoarele:

- $X_1$  - 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 școli elementare din vecinătatea locuințelor (localitatea Edgewood), care se iau în considerare la cumpărarea unei case.





Primul model regresiv propus este de forma:

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

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:

$$Y = b_0 + b_1X_1 + \dots + b_4X_4 + b_5X_3X_4 + b_6X_5 + b_7X_5^2 + b_9D_7 + \\ + b_{10}D_8 + b_{11}D_9 + b_{12}D_{10} + b_{13}D_{11} + b_{14}D_{12}$$

care este validat din punctul de vedere al analizei reziduurilor.





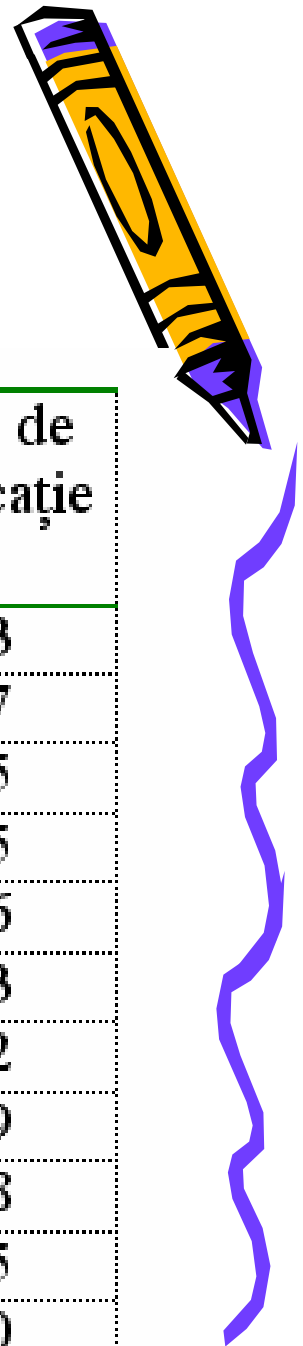
Ultimul model propus, neliniar și acesta, de ecuație

$$Y = b_0 + b_1X_1 + \dots + b_4X_4 + b_5X_3X_4 + b_6X_5 + b_7X_5^2 + b_9D_7 + \\ + b_{10}D_8 + b_{11}D_9$$

se dovedește 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 <sub>1</sub>	56.719	27.974	0.047
X <sub>2</sub>	9.917	3.438	0.005
X <sub>3</sub>	-98.156	42.666	0.025
X <sub>4</sub>	-78.910	27.752	0.006
X <sub>3</sub> X <sub>4</sub>	30.390	11.878	0.013
X <sub>5</sub>	3.301	3.169	0.302
X <sub>5</sub> <sup>2</sup>	1.641	0.733	0.029
X <sub>6</sub>	13.119	8.285	0.118
D <sub>7</sub>	27.424	10.988	0.015
D <sub>8</sub>	67.062	16.822	0.000
D <sub>9</sub>	47.273	14.844	0.002





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

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





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 eșantion mic dar având multe variabile predictive.

Ca regulă generală, numărul de predictorii nu trebuie să depășească valoarea  $n/10$ , unde  $n$  este volumul eșantionului.

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

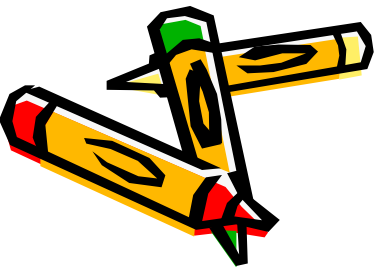


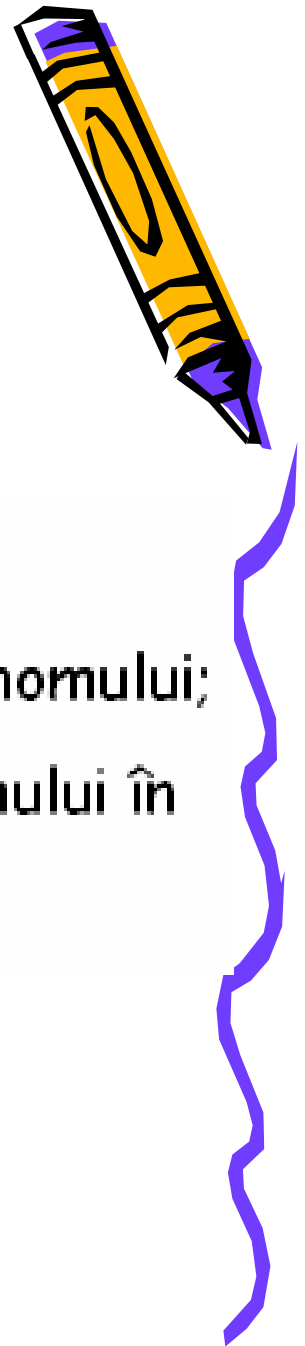
# studiu de caz cu MATLAB



Aproximarea unui set de date  $\{(x_i, y_i), 1 \leq i \leq n\}$  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  $\{x_i, 1 \leq i \leq n\}$ .



# exemplu

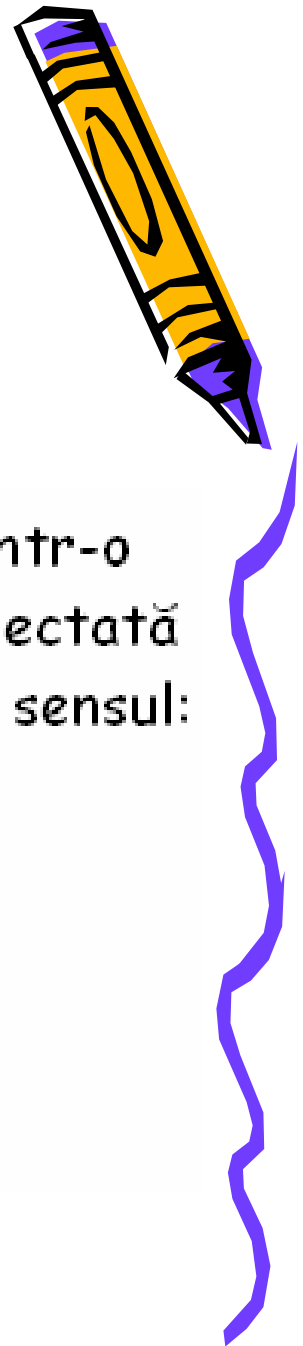
Considerăm situația numărului de pacienți internați într-o secție a unui spital pe parcursul unui an, situație reflectată în următorul tabel, unde am codificat lunile anului în sensul:

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



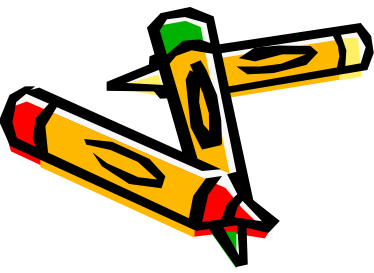
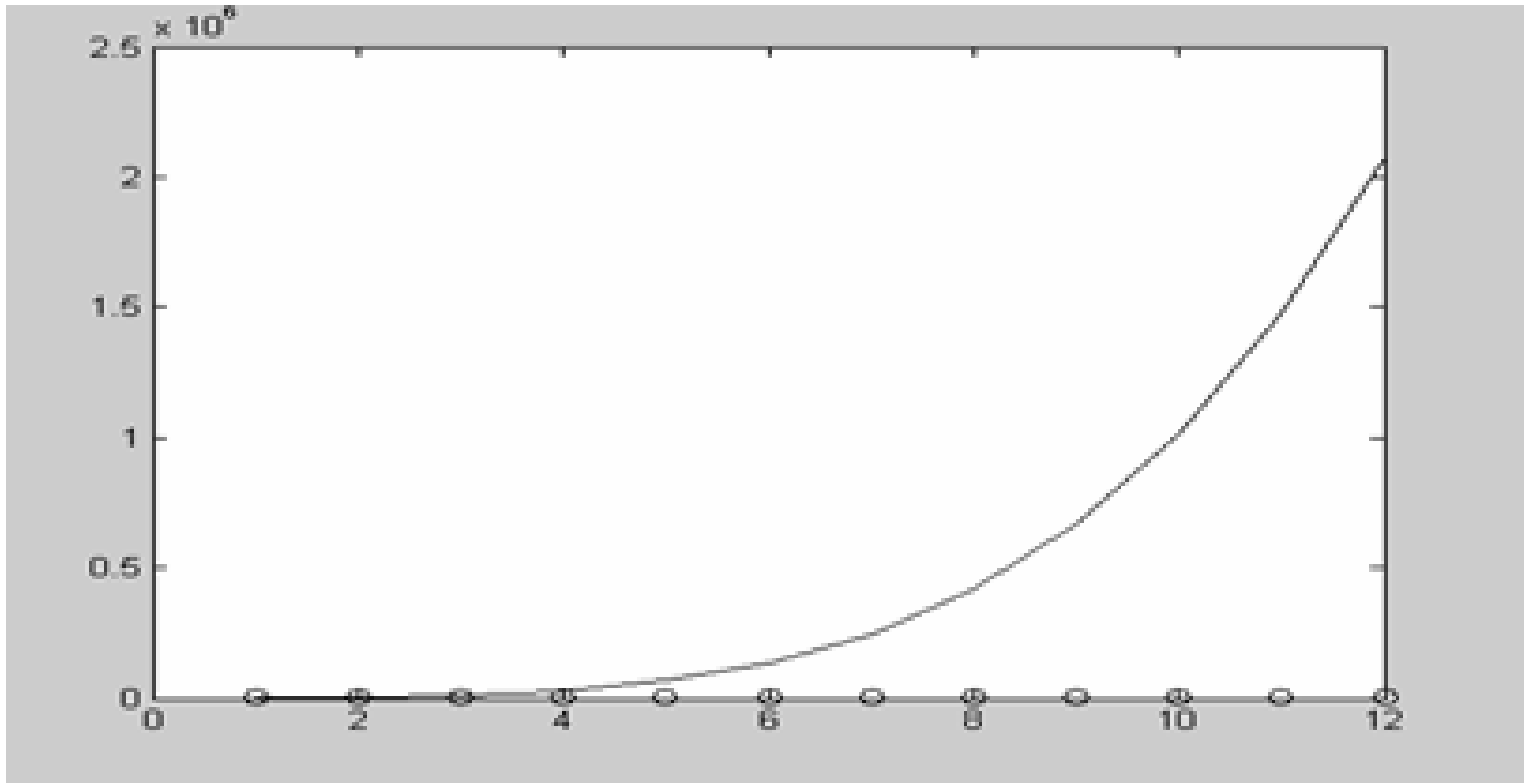




Să desenăm în același sistem de axe setul de date și polinomul de gradul 4 ce-l aproximează:

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





Rezultatul nu este mulțumitor, datorită valorilor mari ale ordonatelor.

În acest caz este necesară normalizarea datelor.

»  $X_s = (X - \text{mean}(X)) ./ \text{std}(X)$

$X_s =$

Columns 1 through 5

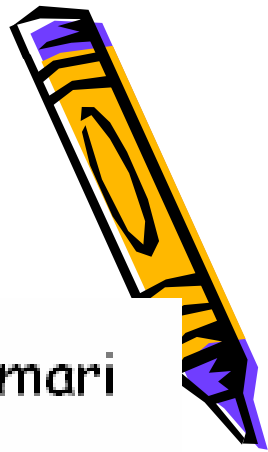
-1.5254 -1.2481 -0.9707 -0.6934 -0.4160

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)
```

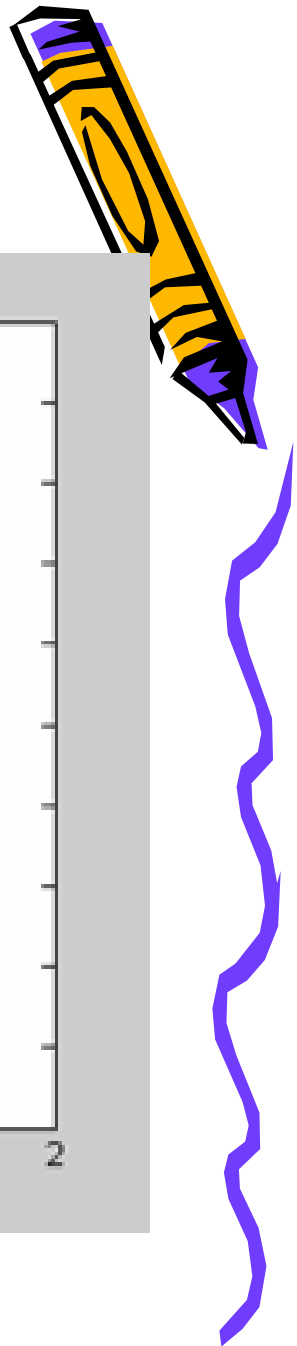
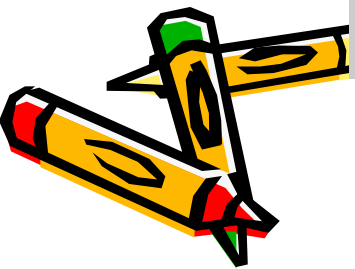
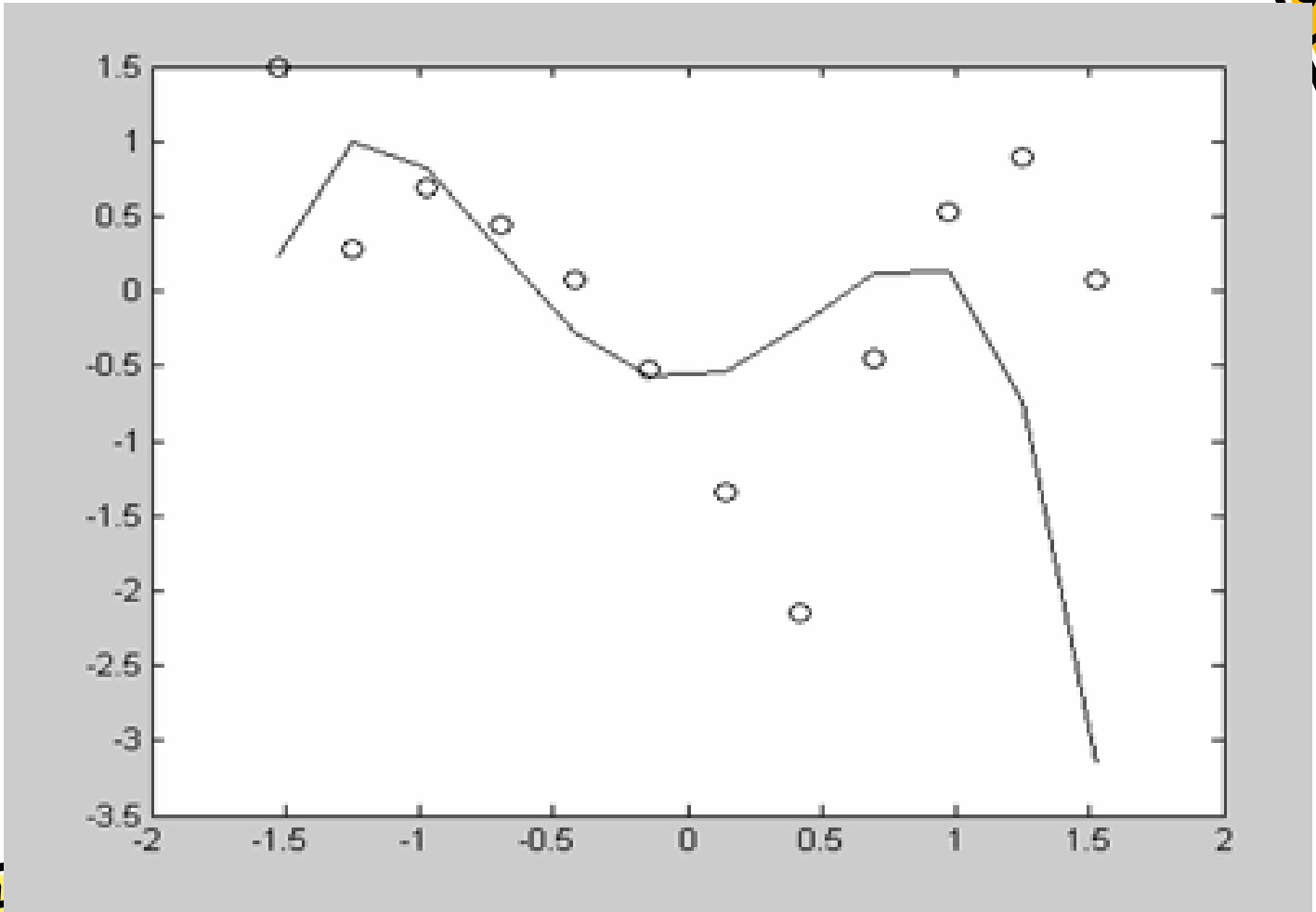
```
ps =
```

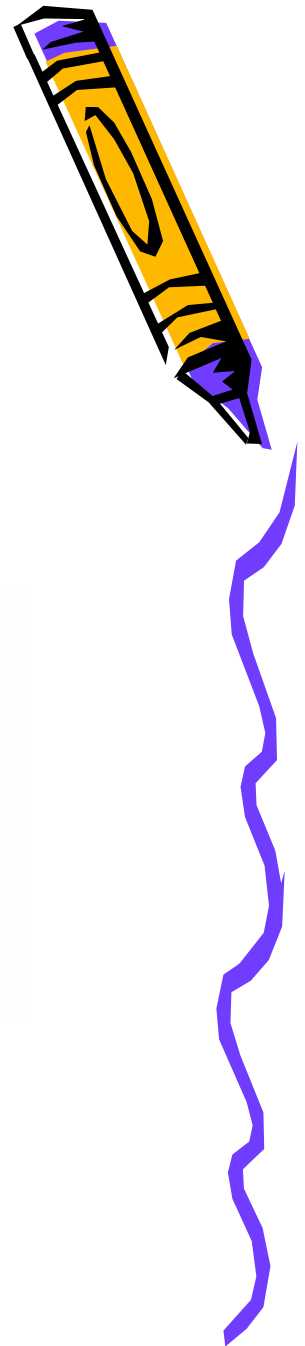
```
    -0.5961    0.1525    2.1616   -0.5425   -1.0883
```

```
» 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')
```







Reziduurile, fiind diferențele între valorile prognozate și 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

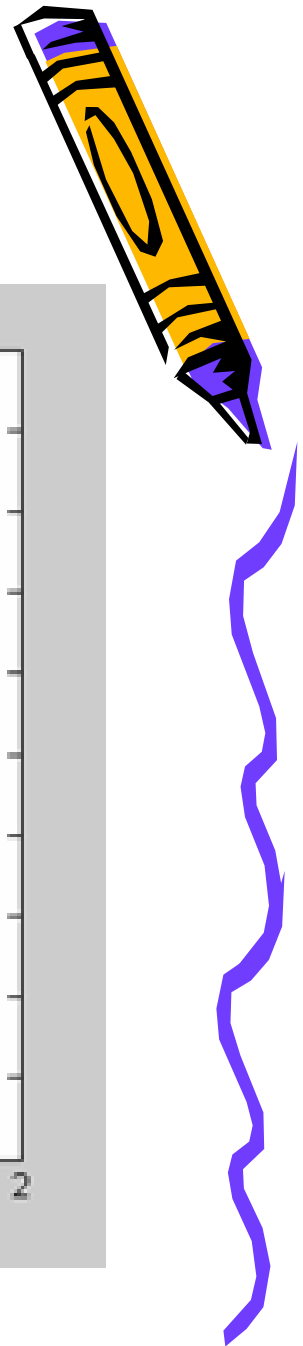
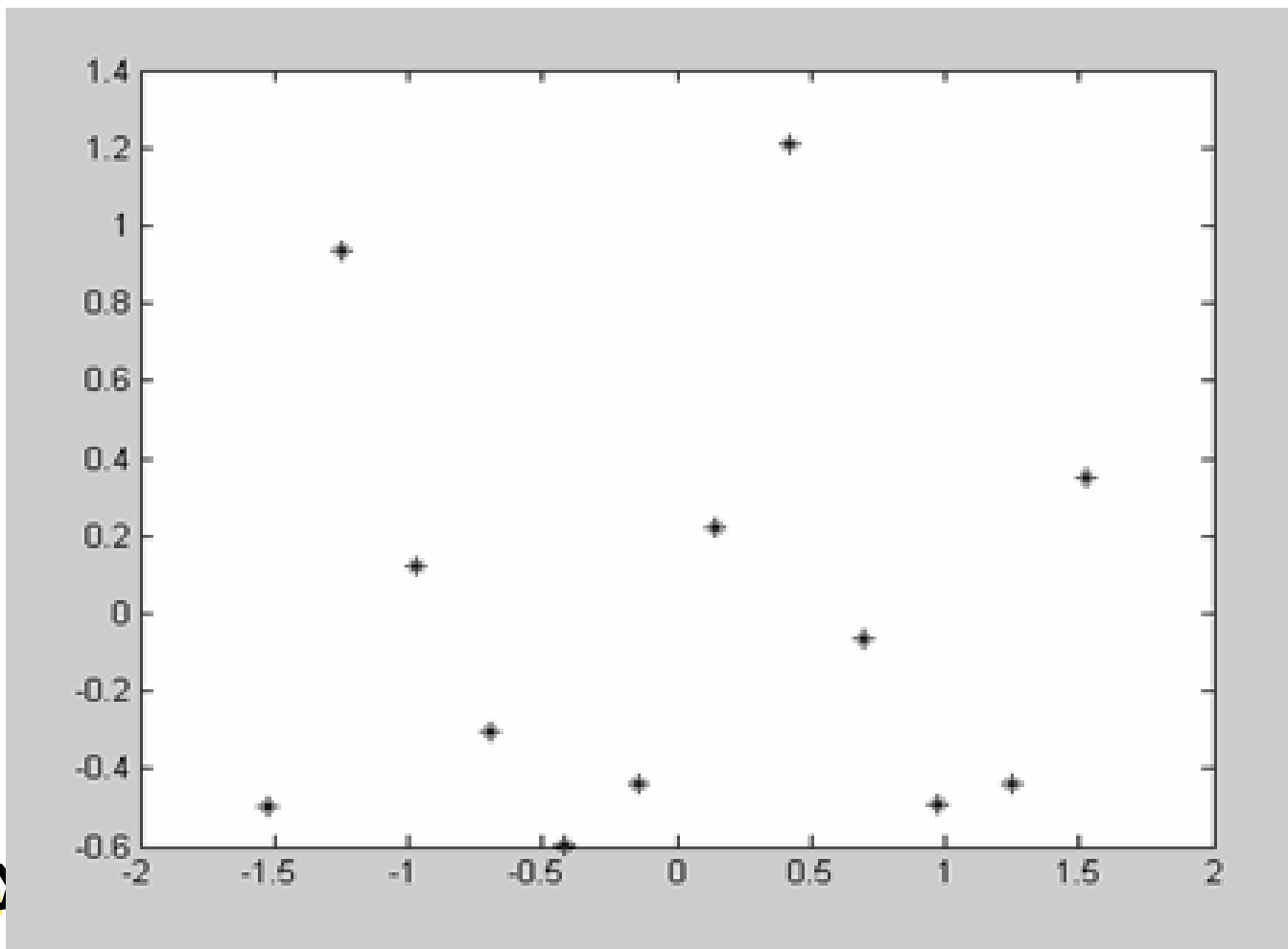
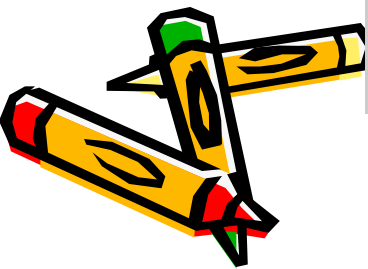
-0.4411 0.2211 1.2079 -0.0623 -0.4921

Columns 11 through 12

-0.4377 0.3500

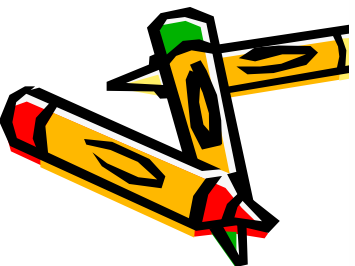
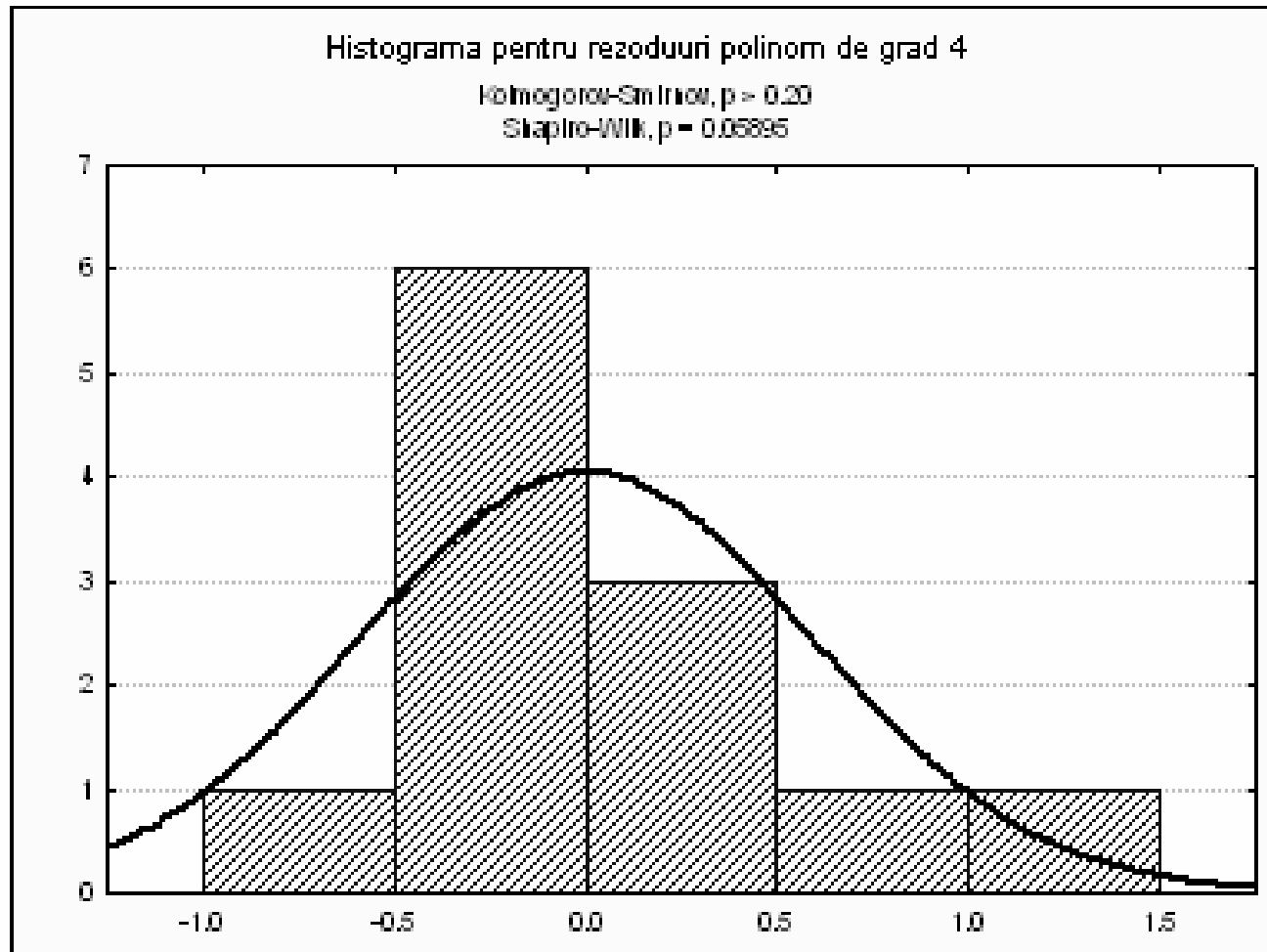
» plot(Xs,rez,'\*')







# histograma reziduurilor

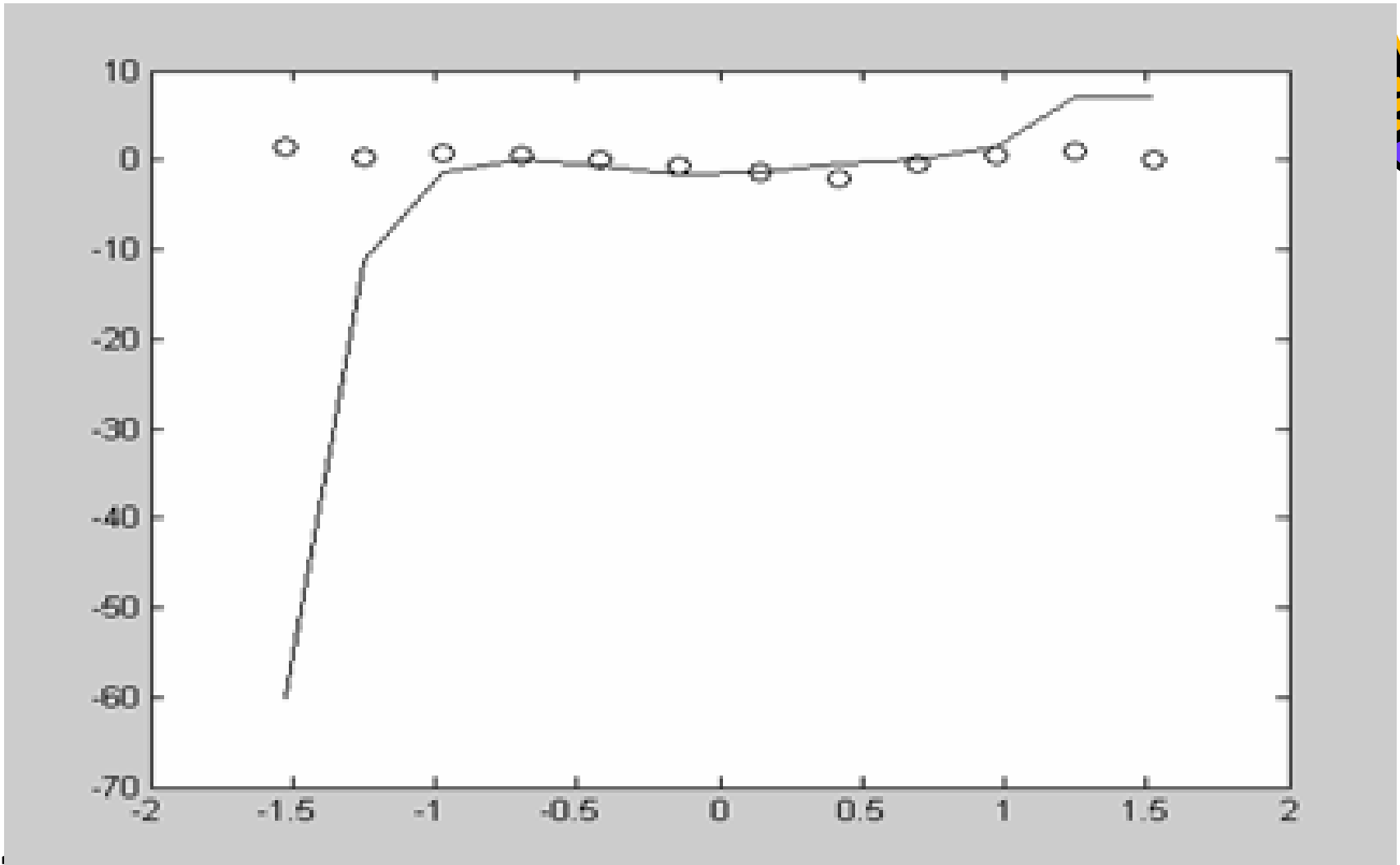




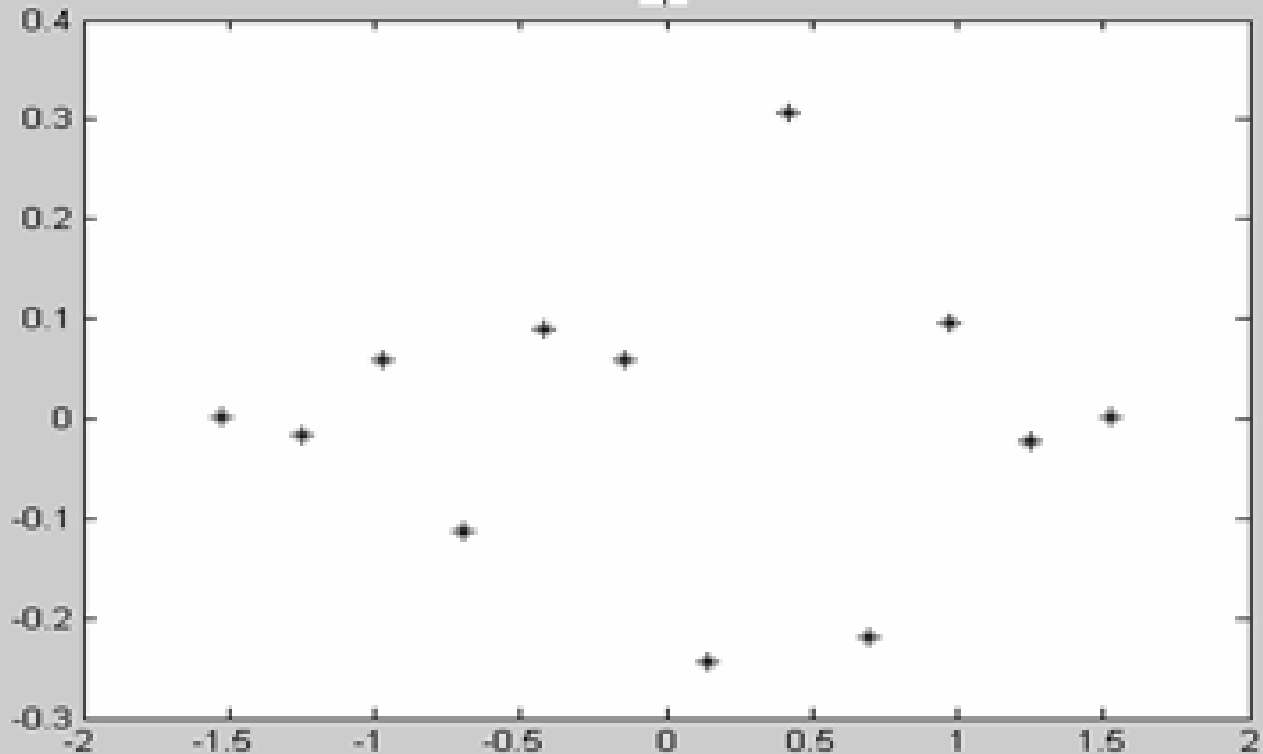
Nu ne putem declara mulțumiți de rezultat și vom încerca să mărim gradul polinomului de aproximare, să zicem 9:

```
» 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')
```

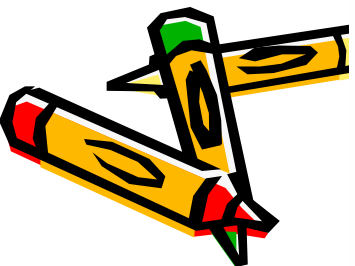
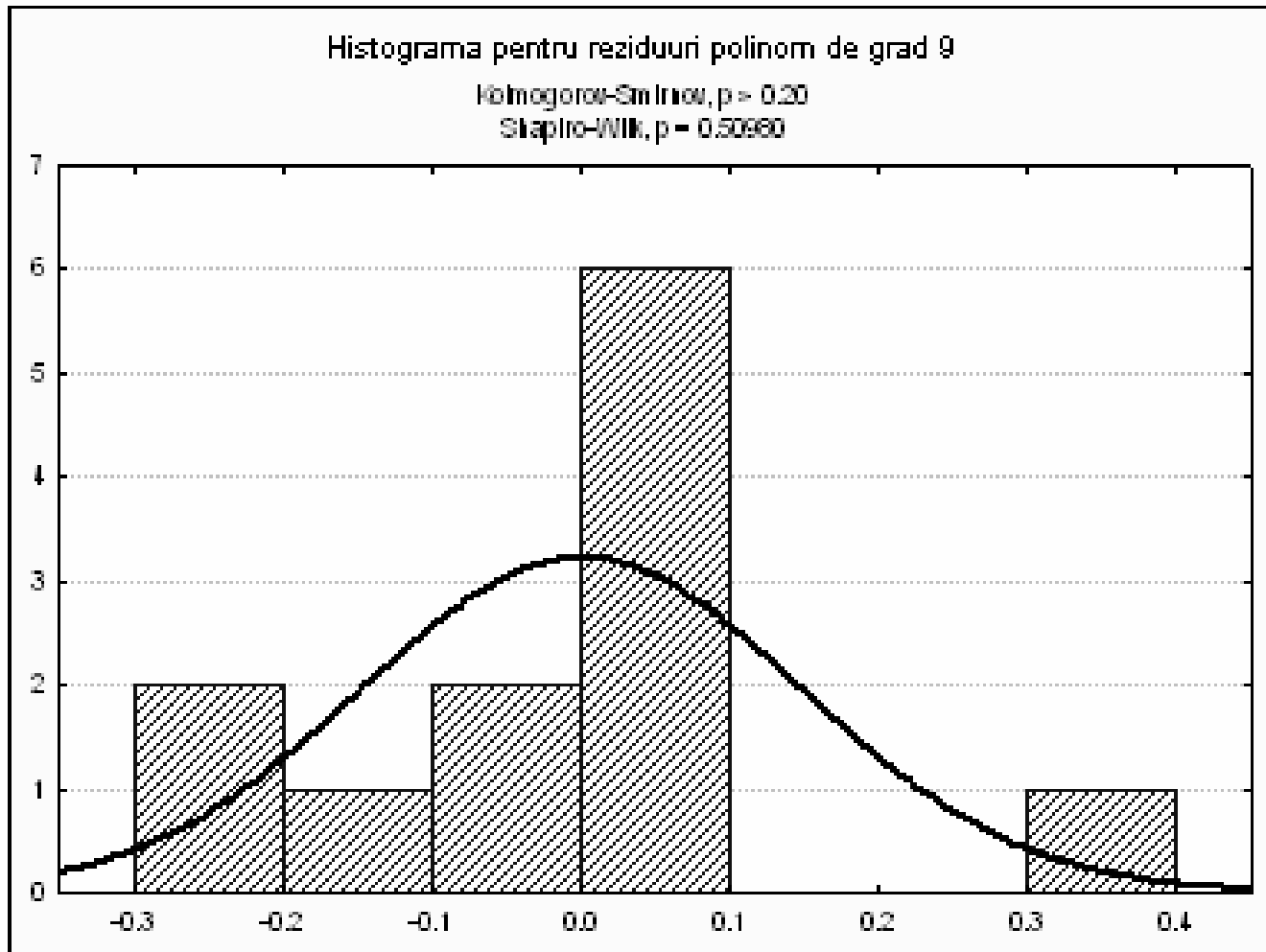
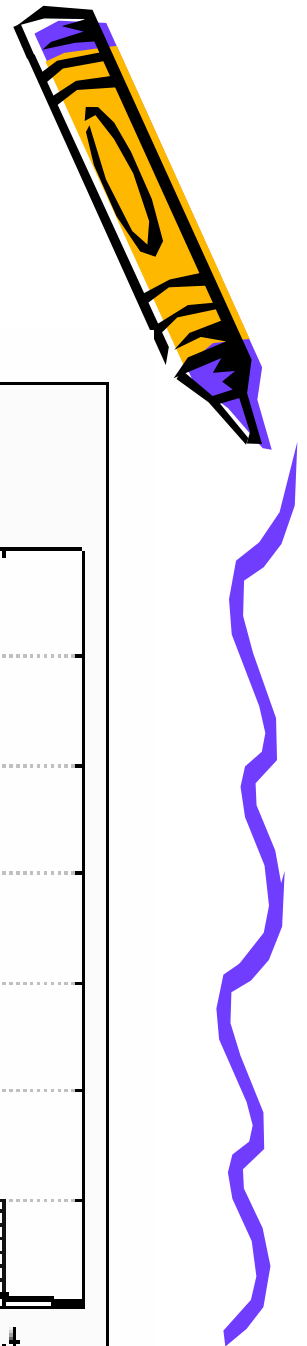




```
» pol1=polyval(ps,Xs,Ys); rez=pol1-Ys; plot(Xs,rez,'*')
```



# histoarama reziduurilor



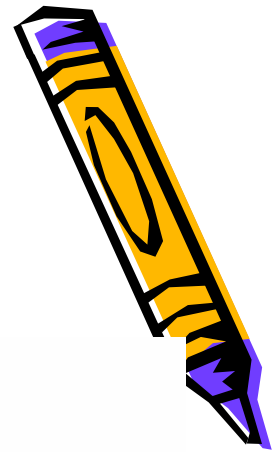
# aproximarea prin combinatii de exponentiale



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

- $\text{loglog}(x,y)$  scalează ambele axe utilizând logaritmul în baza 10;
- $\text{semilogx}(x,y)$  scalează logaritmice numai axa  $Ox$ ;
- $\text{semilogy}(x,y)$  scalează logaritmice numai axa  $Oy$ .





```
» 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
```

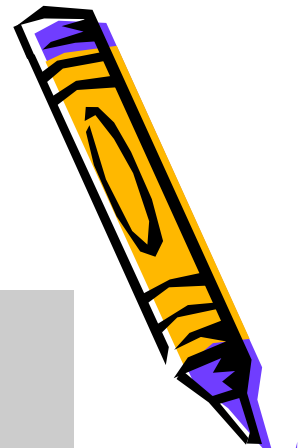
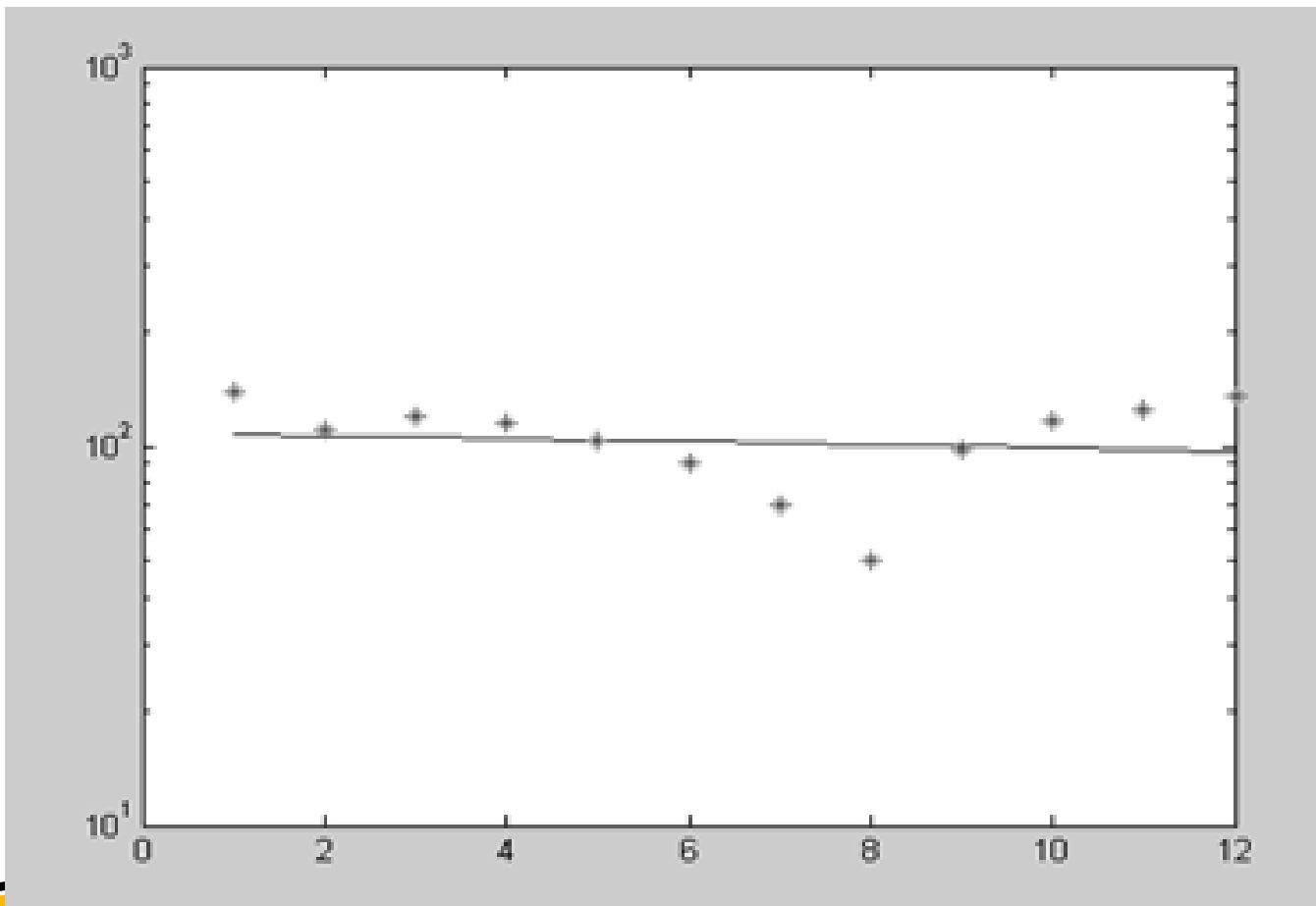
```
102.9317 101.9521 100.9817 100.0206 99.0686
```

```
Columns 11 through 12
```

```
98.1257 97.1918
```

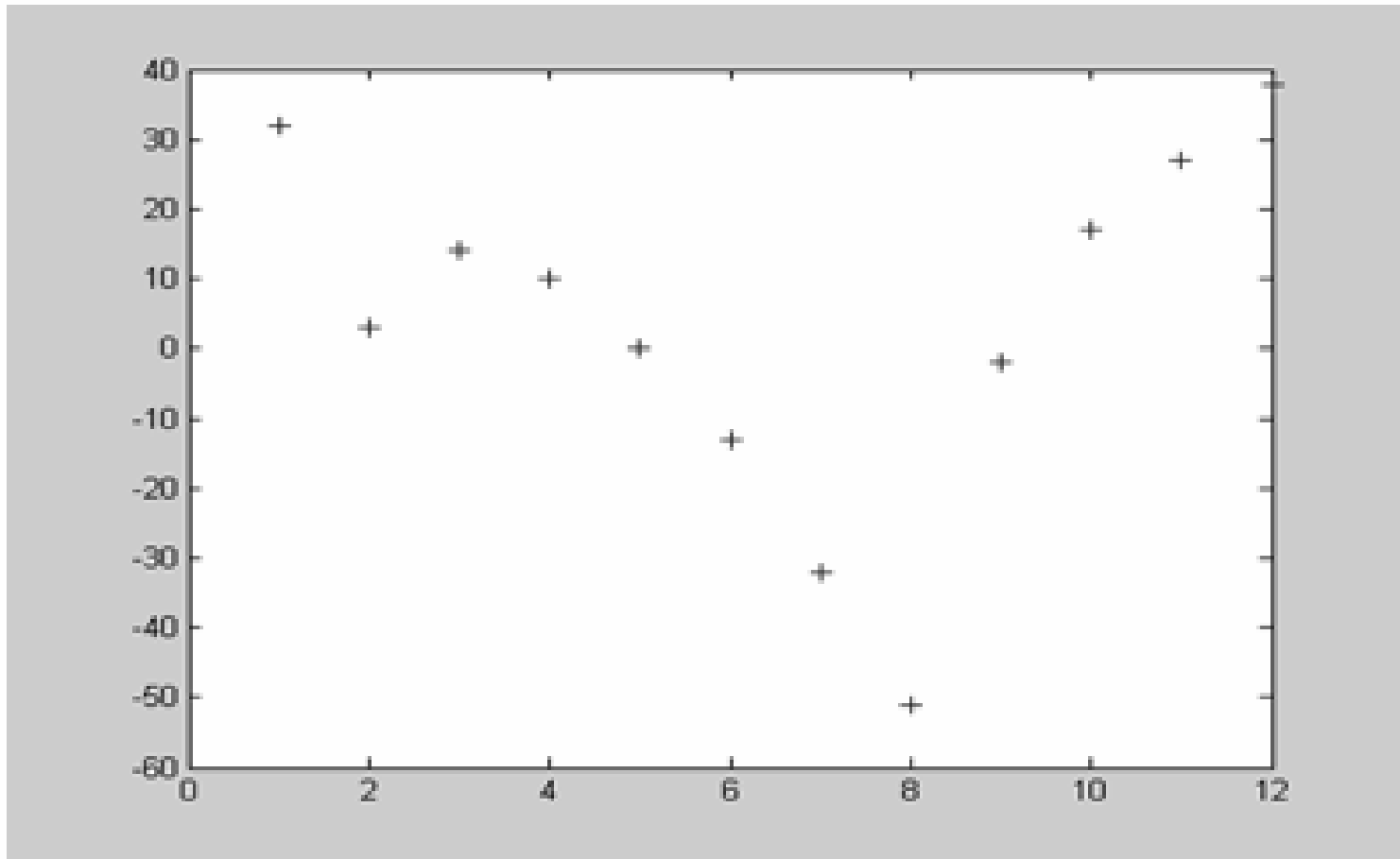
```
» semilogy(X,logpred1,X,Y,'*')
```

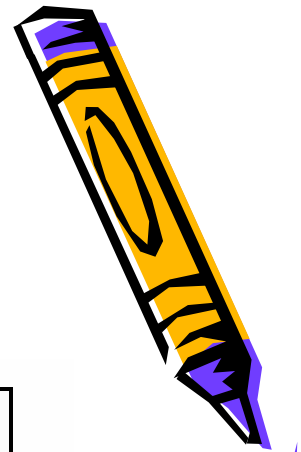




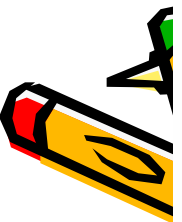
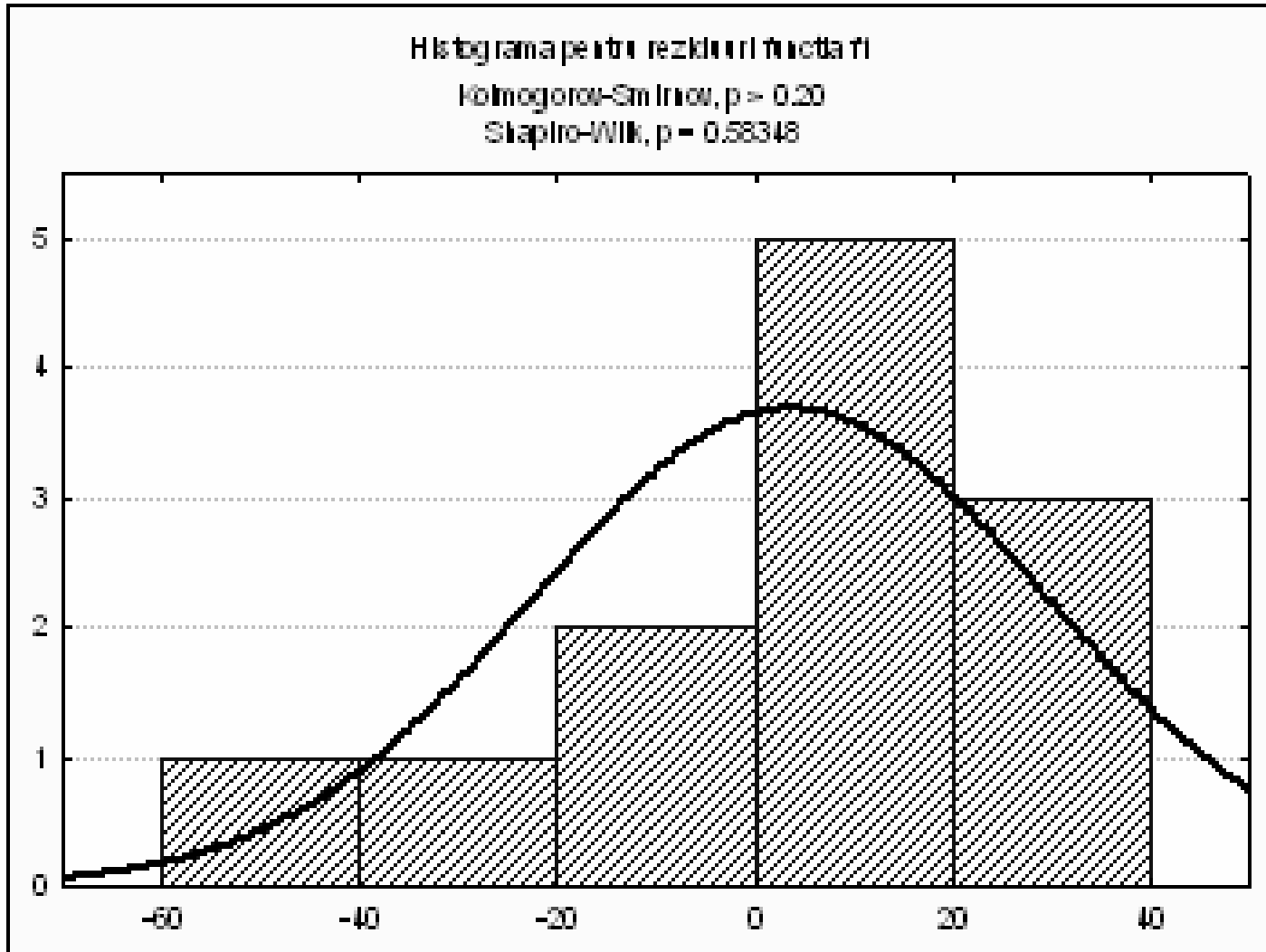


```
»r=Y-logpred1; plot(Xs,r,'+')
```



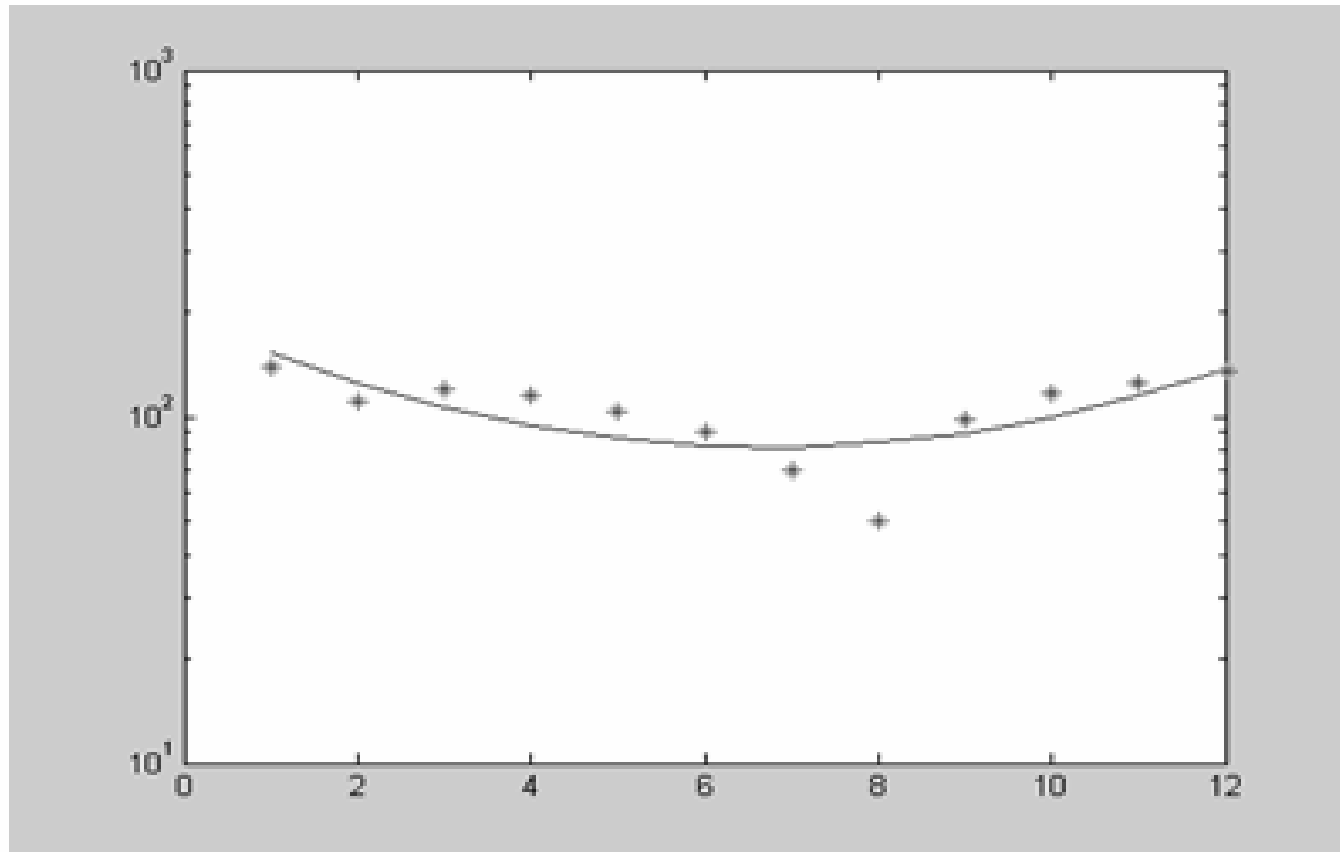


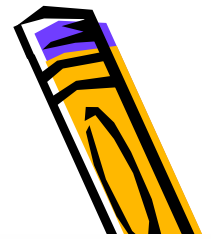
# histograma reziduurilor



Vom încerca altă variantă de funcție

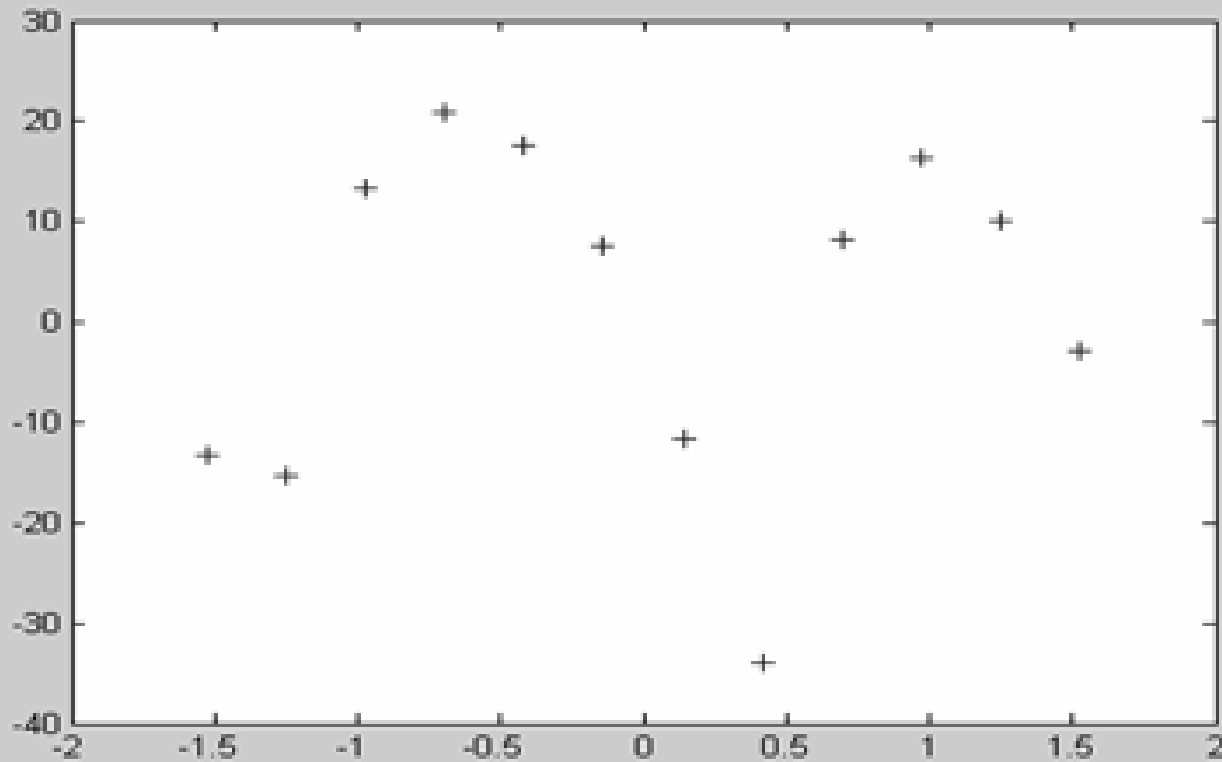
```
» logp2=polyfit(Xs,log10(Y),2);  
» logpred2=10.^polyval(logp2,Xs);  
» semilogy(X,logpred2,X,Y,'+')
```

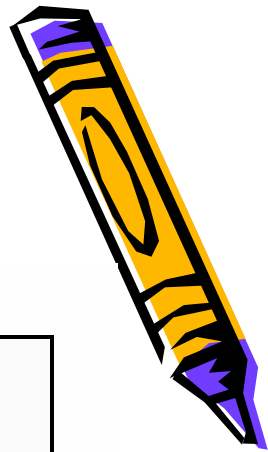




Reprezentăm grafic reziduurile precum și histograma lor.

```
»r=Y-logpred2; plot(Xs,r,'+')
```

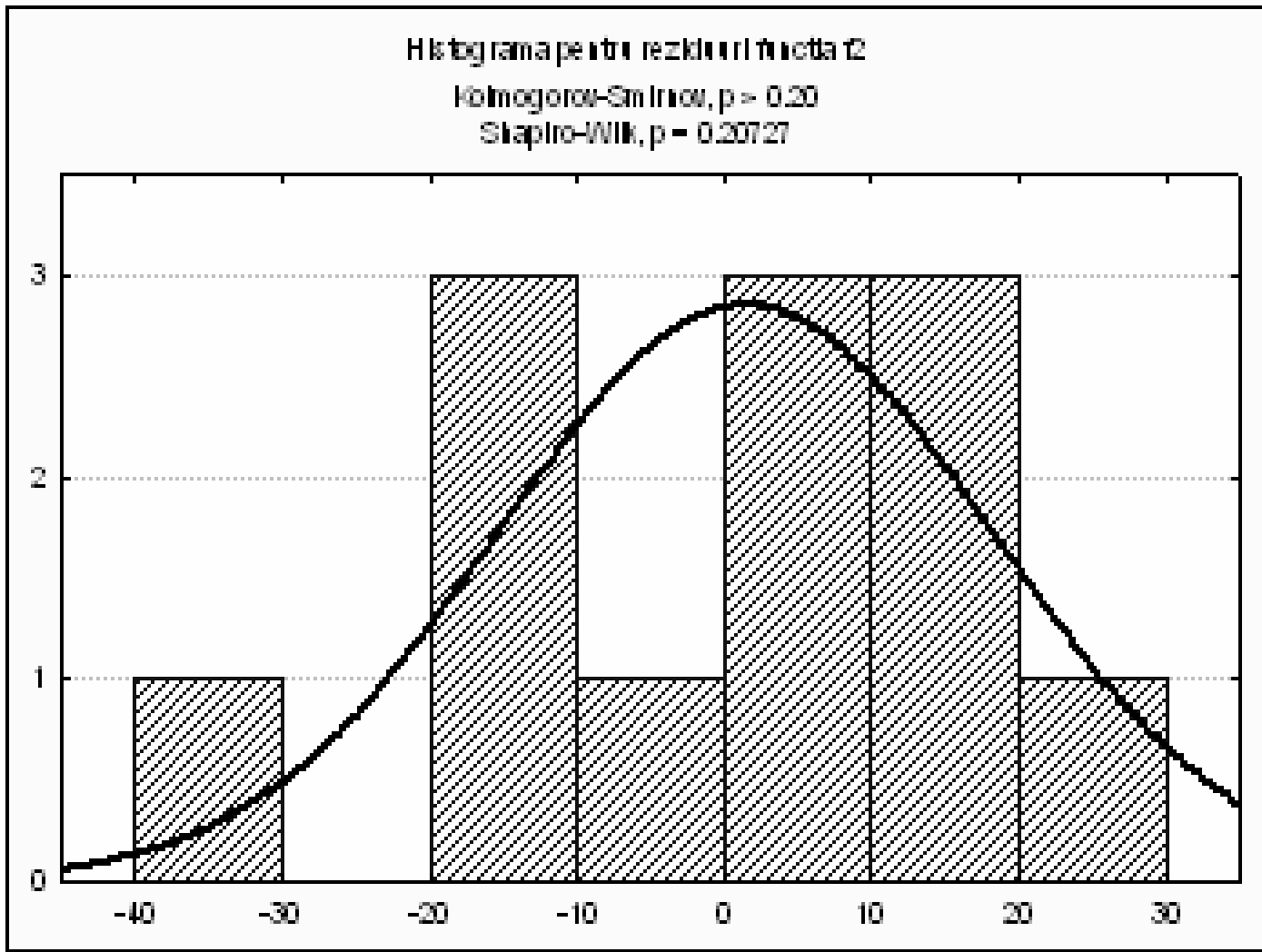


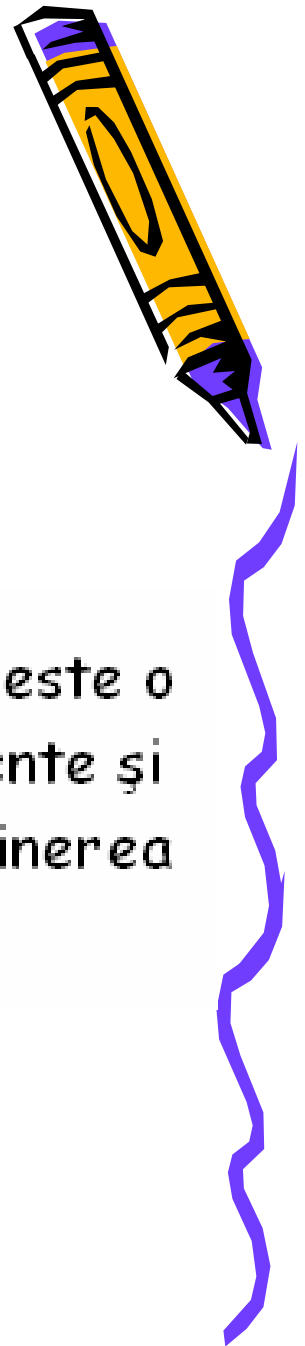


Histograma pentru reziduurile funcției

Kolmogorov-Smirnov,  $p > 0.20$

Shapiro-Wilk,  $p = 0.20727$

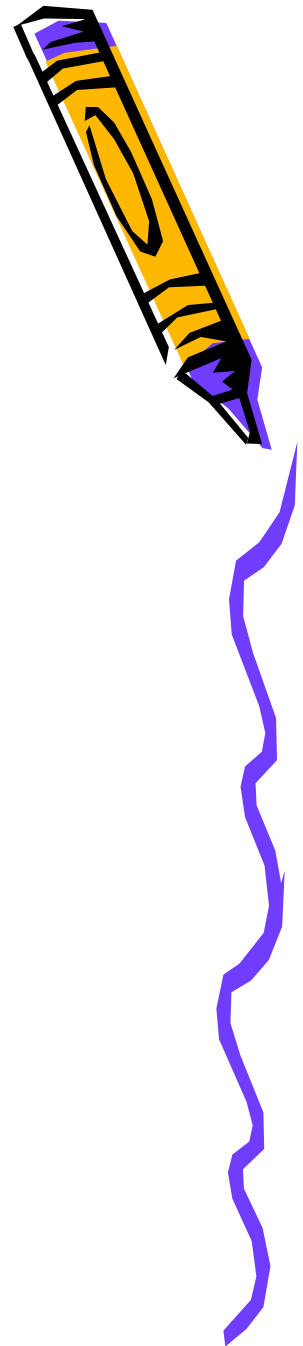




Problema utilizării regresiei neliniare la un set de date este o procedură euristică, fiind necesare mai multe experimente și folosirea cunoștințelor specifice domeniului pentru obținerea unor rezultate bune



# Regresia logistică



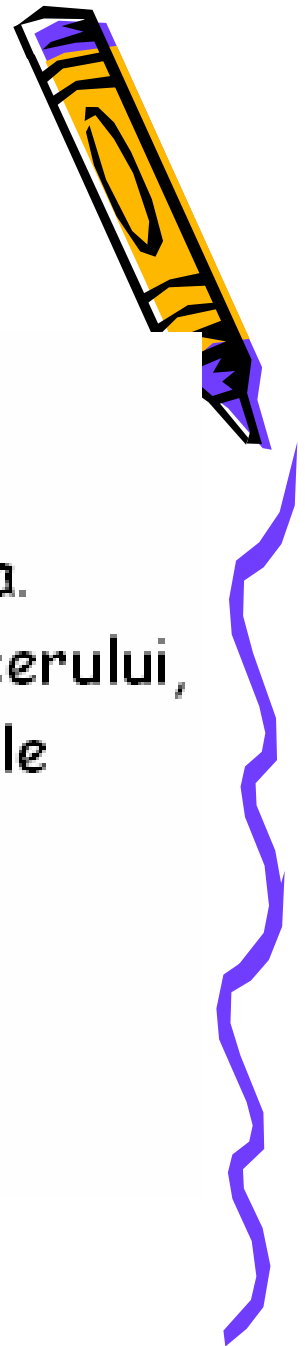


Ce se întâmplă însă în cazul în care una din variabile este categorială?

În multe studii, variabila dependentă (variabila răspuns), care trebuie dedusă din variabilele explicative, este o variabilă categorială.







De exemplu, în medicină, această variabilă poate reprezenta diagnosticarea unei boli (DA sau NU) a unui individ, pe baza unor date prelevate de la acesta. Astfel, în cursul analizei epidemiologice în studiul cancerului, se colectează date privind un anumit număr de variabile care ar putea influența riscul de îmbolnăvire. Pentru fiecare combinație a diferitelor valori ale acestor variabile trebuie estimată probabilitatea detectării apariției maladiei.



# Exemple



Dacă vom considera cazul cancerului pulmonar, ne propunem să estimăm riscul de îmbolnăvire a unui subiect de sex masculin, în vârstă de 58 ani, care a muncit timp de 40 ani ca muncitor siderurgist și care a fumat în medie 20 țigări pe zi, începând cu vârsta de 16 ani.



Altă situație se referă la încadrarea în categoria „fraudă” a unor tranzacții comerciale, pe baza anumitor caracteristici ale lor.





Spre deosebire de cazul modelelor regresive liniare multiple, în care valorile ce trebuie estimate sunt numere, în această situație în care variabila dependentă este reprezentată de o variabilă categorială (posibilitatea de a face cancer, sau de a produce o fraudă), având valorile: DA sau NU, problema se schimbă. Nu se mai poate folosi metoda regresiei multiple, pentru care variabila dependentă era numerică, continuă.



# Regresia logistică

Plecând de la aceeași filosofie ca și în cazul regresiei multiple clasice, vom construi un model ușor diferit, cunoscut sub numele de *regresie logistică liniară multiplă*, pe scurt *regresie logistică*.





Spre deosebire de regresia clasică, aici una sau mai multe variabile predictive/explicative (variabile independente) pot fi categoriale, obligatoriu variabila dependentă, deci în cazul acestui nou model este vorba de caracterul nenumeric, calitativ (categorial) al unora dintre variabilele sale.





Principiul de bază rămâne același ca și la regresia multiplă, diferența constând în faptul că dacă în primul caz estimăm valoarea variabilei dependente pe baza valorilor predictorilor, în acest caz estimăm o *transformare* a variabilei dependente.





Astfel, dacă variabila dependentă are ca valori binare afirmațiile DA și NU, spre exemplu îmbolnăvire sau nu, răspuns pozitiv sau negativ la un tratament, fraudă sau nu etc., care codate au valorile 1 și 0, atunci media acesteia reprezintă proporția indivizilor din populație cu caracteristica respectivă.

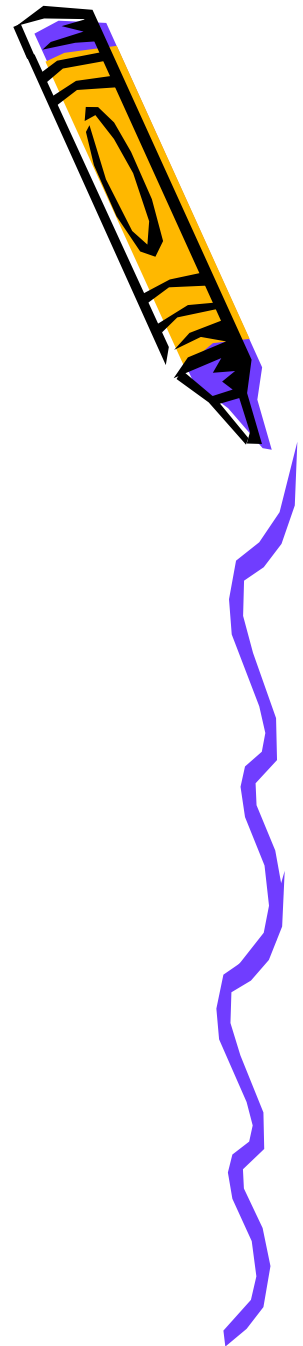




# Exemplu

să presupunem că avem pentru 100 de indivizi,  
65 răspunsuri DA și 35 răspunsuri NU, adică 65 de 1  
și 35 de 0.

Media acestor valori va fi  $(65 * 1 + 35 * 0)/100 = 65/100$   
adică 65% de DA.



În general, va rezulta că modelul regresiv logistic va estima proporția subiecților/obiectelor care vor avea aceeași caracteristică de interes pentru studiul statistic, sau echivalent, probabilitatea ca un subiect/obiect oarecare din populație să aibă o anumită caracteristică, de exemplu cancer pulmonar, escroc.





Din punct de vedere practic, în locul acestei proporții se ia în considerație o transformare a ei.

Rațiunea pentru folosirea unei transformări a proporției este aceea că, în principiu, o combinație oarecare a predictorilor din ecuația de regresie ar putea lua valori înafara intervalului  $[0, 1]$  de care poate aparține o probabilitate, obținând astfel o aberație din punct de vedere matematic.



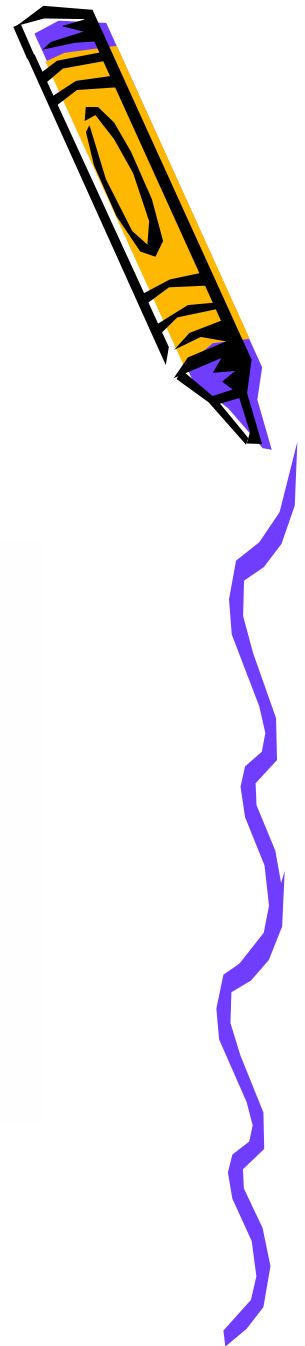
Cu alte cuvinte, se procedează la o ,normare' a intervalului valorilor produse de combinația predictorilor, adică o transformare a acestuia în intervalul  $[0, 1]$ , și astfel valorile variabilei răspuns pot fi considerate probabilități.

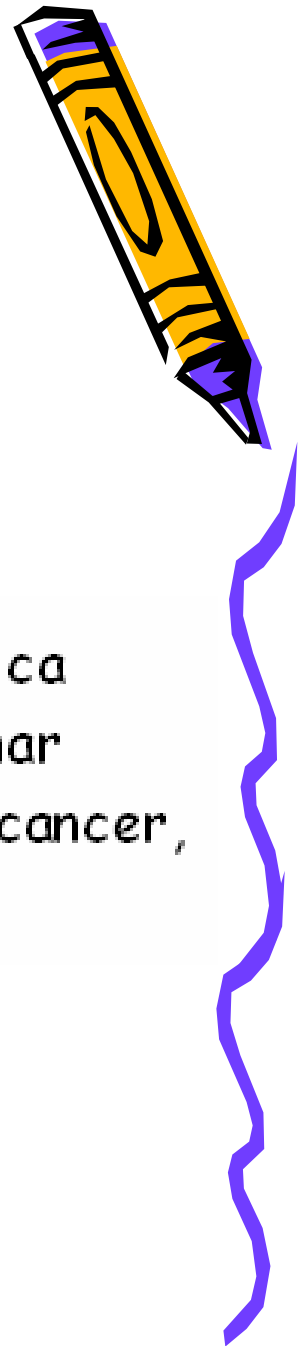


# logit

Metoda regresiei logistice constă în folosirea transformării *logit* scrisă ca  $\text{logit}(p)$ .

Concret, aici  $p$  reprezintă probabilitatea ca un individ/obiect oarecare să aibă caracteristica de interes cerută și deci  $(1 - p)$  va reprezenta probabilitatea ca acesta să nu o aibă.





În exemplul de mai sus,  $p$  este probabilitatea ca un subiect din populație să aibă cancer pulmonar sau să producă o fraudă, iar  $(1 - p)$  să nu aibă cancer, sau nu producă fraudă.





Raportul  $\frac{p}{1-p}$  se numește *șansă*,

iar transformarea  $\text{logit}(p) = \ln \frac{p}{1-p}$

este numită *logaritmul (log-ul) șansei*.





Această mărime este importantă și deoarece ajută la compararea predicțiilor între subiecții/obiectele care au sau nu o anumită caracteristică, de exemplu pentru subiecții dintr-un lot cu vârsta mai mică de 40 ani față de cei peste 40 ani.

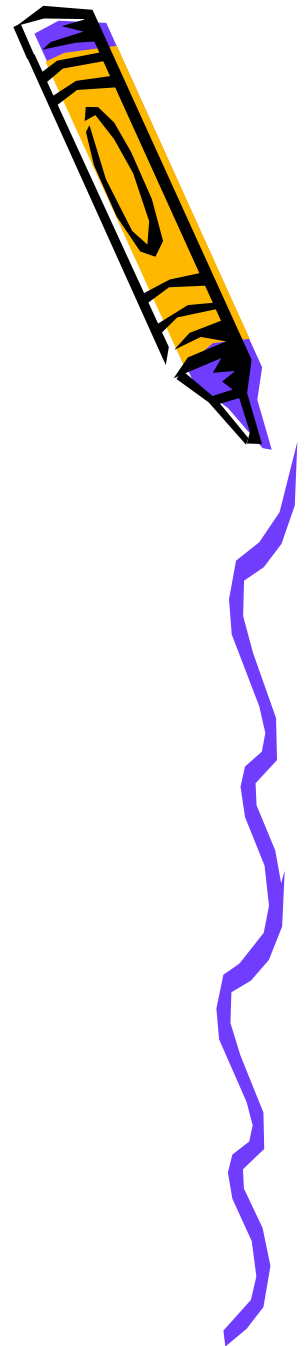






În acest fel, dacă prin această nouă caracteristică introdusă în analiza mulțimii studiate, obținem o divizare a acesteia în două grupuri, putem obține o măsură a raportului șanselor între cele două grupuri de subiecți/obiecte de a avea sau nu caracteristica de bază care îi diferențiază.





Dacă vom considera:

$$l_1 = \text{logit}(p_1)$$

log-ul șansei pentru primul grup și

$$l_2 = \text{logit}(p_2)$$

log-ul șansei pentru al doilea grup, atunci:



# Logaritmul riscului relativ



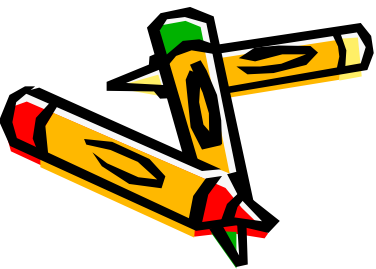
$$l_1 - l_2 = \text{logit}(p_1) - \text{logit}(p_2) = \ln \frac{p_1 \cdot (1 - p_2)}{p_2 \cdot (1 - p_1)}$$

reprezintă logaritmul *raportului șanselor*, cu alte cuvinte logaritmul *riscului relativ* pentru cele două grupuri, deoarece raportul șanselor aproximează suficient de bine riscul relativ.





Acesta este folosit uzual de exemplu în studiile epidemiologice, pentru a stabili conexiunea între maladie și expunere.





În concluzie, odată găsit  $\text{logit}(p)$ , pe baza predictorilor, dat de ecuația de regresie logistică:

$$\text{logit}(p) = \ln \frac{p}{1-p} = b_0 + b_1 X_1 + \dots + b_k X_k$$

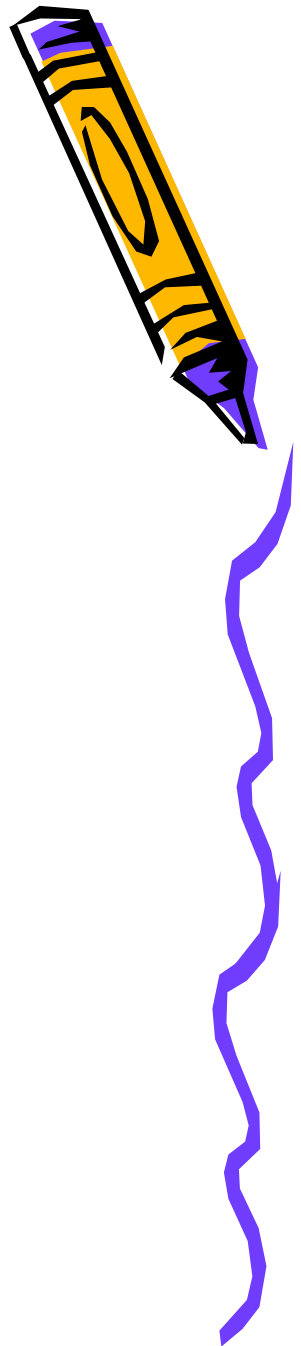
prin transformarea inversă, exponențială, se găsește valoarea probabilității  $p$  ca un anumit individ/obiect să aibă sau nu o anumită caracteristică („etichetă” de clasificare).

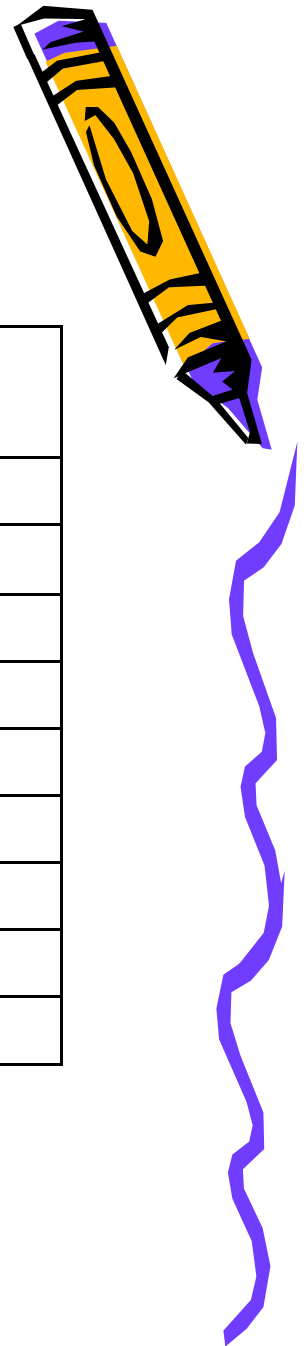


# Exemplu

Pentru stabilirea legăturilor existente între apariția hipertensiunii și următorii factori de risc: *fumat*, *obezitate*, *vârstă* (sub sau deasupra valorii alese de 40 ani -codat 0 sau 1), vom aplica o analiză regresivă logistică.

Să considerăm următoarea situație statistică, descrisă în tabelul de mai jos





E

fumat	obezitate	vârstă	număr subiecți	număr subiecți cu hipertensiune (%)
0	0	0	60	5(8%)
1	0	0	17	2(11%)
0	1	0	8	1(13%)
0	0	1	187	35(19%)
1	1	0	2	0(0%)
1	0	1	60	5(8%)
0	1	1	51	15(29%)
1	1	1	23	8(35%)
			Total=433	total=79 (18%)



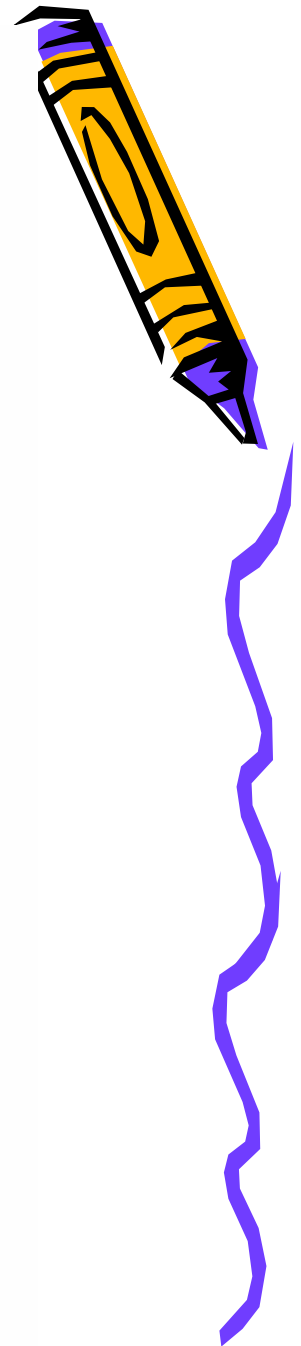
Plecând de la acest tabel general, vom construi un tabel de lucru, care rezumă, sintetic, toate posibilitățile existente.

Modul în care se construiește acest tabel este simplu: în primele două linii ale sale se iau în considerație cele două situații posibile de clasificare ale indivizilor (hipertensiune = 1, normal = 0), pentru situația în care variabilele explicative fumat, obezitate și vârstă au valorile egale cu 0, indivizi care nu fumează, nu sunt obezi și au vârsta sub 40 de ani. Mai departe se consideră toate combinațiile posibile ale predictorilor.





fumat	obezitate	vârsta	număr	hipertensiune
0	0	0	55	0
0	0	0	5	1
1	0	0	15	0
1	0	0	2	1
0	1	0	7	0
0	1	0	1	1
0	0	1	152	0
0	0	1	35	1
1	1	0	2	0
1	1	0	0	1
1	0	1	55	0
1	0	1	5	1
0	1	1	36	0
0	1	1	15	1
1	1	1	15	0
1	1	1	8	1





În cazul a 3 variabile explicative avem 16 combinații  
posibile în total, iar dacă vom considera 4 variabile  
explicative, vom avea 64 de cazuri posibile în total.  
Toate variabilele explicative au fost codate  
(0 = Nu și 1 = DA).

Calculul nu se poate face ,manual' din cauza volumului  
mare de muncă.



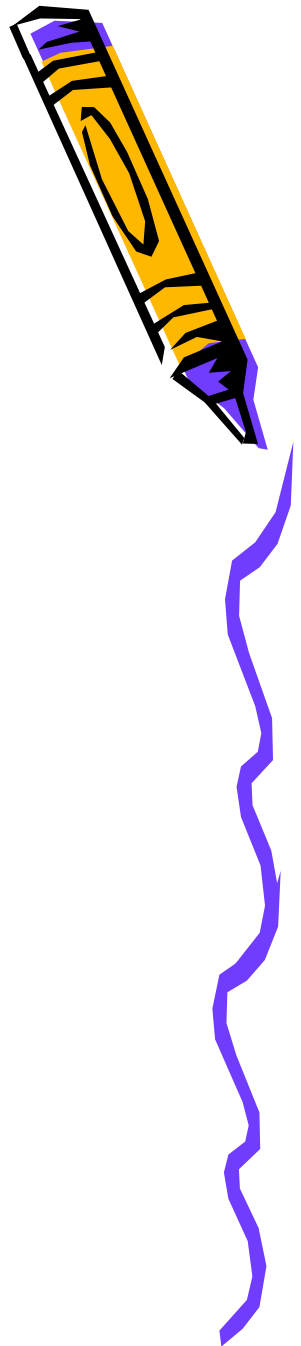
Prezentăm mai jos tabelul privind analiza regresiei logistice pentru cazul de mai sus, conform STATISTICA

	Estimare	Eroarea standard $se(b)$	Nivelul de semnificație $p$
Constantă	-2.378	0.380	
Fumat	-0.068	0.278	0.81
Obezitate	0.695	0.285	0.015
Vârstă	0.872	0.398	0.028



Se observă din acest tabel că cel mai important factor de risc pentru hipertensiune este reprezentat de obezitate ( $p = 0.015$ ), urmat de vârstă ( $p = 0.028$ ), în timp ce fumatul nu pare a fi semnificativ ( $p = 0.81$ ).

Am obținut și o clasificare a importanței variabilelor explicative asupra variabilei răspuns (ierarhizarea acestora), ierarhie obținută pe baza nivelului de semnificație  $p$ .





Ecuatia de regresie logistica corespunzatoare este dată de:

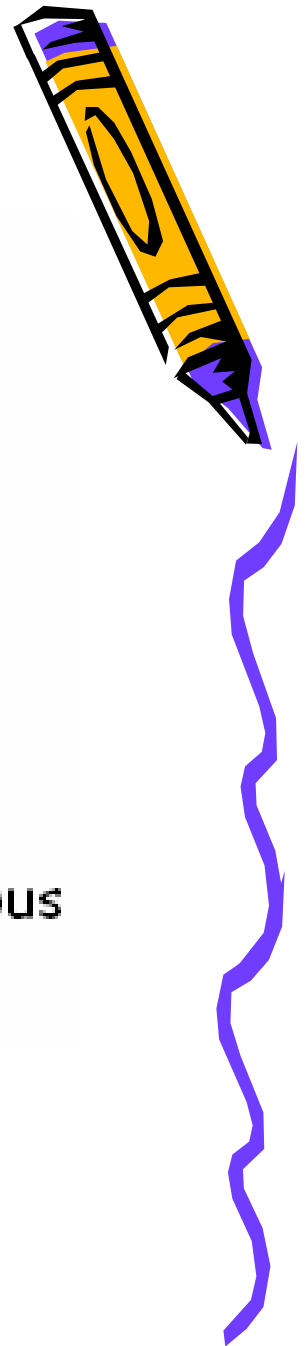
$$\text{logit}(p) = -2.378 - 0.068 \cdot \text{fumat} + 0.695 \cdot \text{obezitate} + 0.872 \cdot \text{varsta}$$

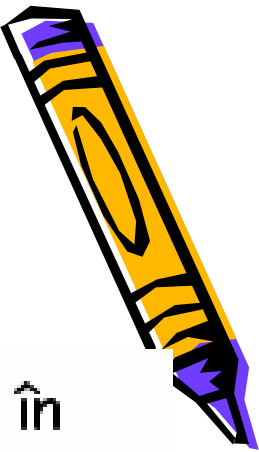


Avem de rezolvat problema clasificării unui individ oarecare ca fiind predispus sau nu la hipertensiune, pe baza atributelor (caracteristicilor) fumat, obezitate sau vârstă.

Se consideră valorile concrete ale aceluși individ privind cei trei factori de risc enumerați mai sus, care se vor introduce în ecuația de regresie.

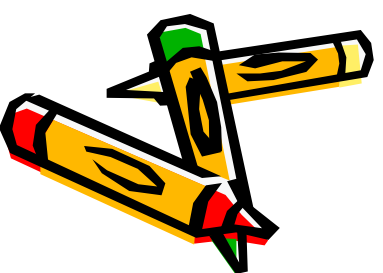
Va rezulta valoarea  $\text{logit}(p)$  corespunzătoare, care va estima riscul ca individul respectiv să fie sau nu predispus la hipertensiune.





Ecuția de regresie logistică se mai poate folosi și în scopul de a compara probabilitățile de predicție a hipertensiunii pentru diferite grupuri, de exemplu pentru cei cu vârsta sub 40 ani față de cei peste 40 ani.

Astfel, codând ca mai sus, cu 0 subiecții sub 40 ani și cu 1 pe cei peste 40 ani, și utilizând ecuația de regresie de mai sus, obținem cele două variante ale sale:





$$\text{logit}(p_{\text{varsta} > 40}) = -2.378 - 0.068 \cdot \text{fum at} + 0.695 \cdot \text{obezitate} + 0.872$$

$$\text{logit}(p_{\text{varsta} < 40}) = -2.378 - 0.068 \cdot \text{fum at} + 0.695 \cdot \text{obezitate}$$







Se obține astfel că:

$$\text{logit}(p_{\text{varsta} > 40}) - \text{logit}(p_{\text{varsta} < 40}) = 0.872.$$

Rezultă că raportul șanselor hipertensiunii asociat cu nivelul de *vârstă* considerat mai sus (de 40 ani)

este  $e^{0.872} = 2.3917$ , valoare care poate fi interpretată astfel:

„riscul de a face hipertensiune peste vârsta de 40 de ani este de 2,39 de ori mai mare decât sub această vârstă.”

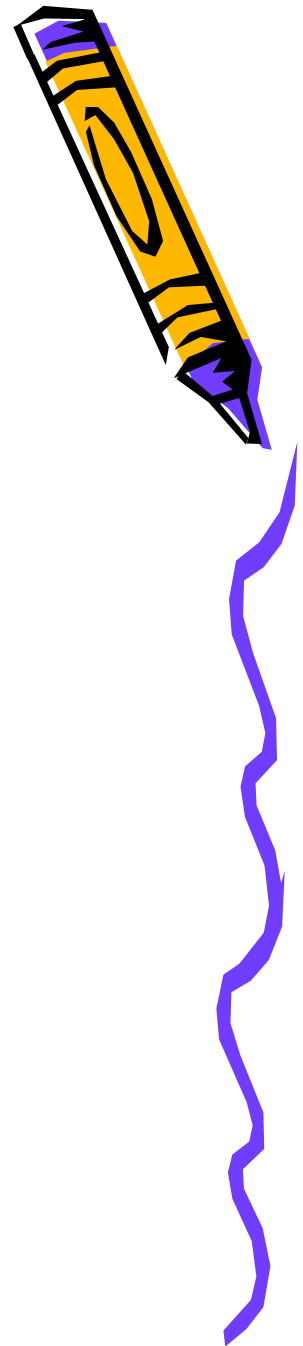


# Exemplu

Studiul lui Tuyns et al., 1977 în departamentul Ille-et-Vilaine din Bretania este un studiu caz-control cuprinzând în lotul caz 200 bărbați diagnosticați cu cancer esofagian iar în lotul control un număr de 775 adulți aleși aleator din aceeași localitate.

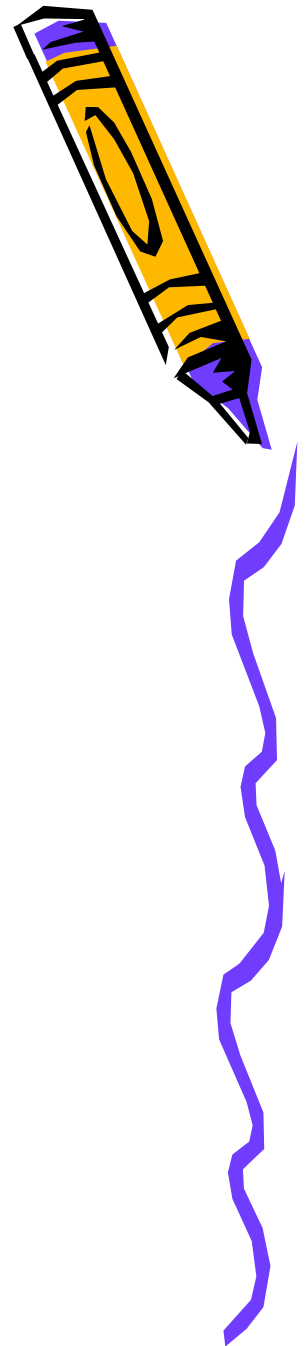
În acest studiu au fost aleși ca factori de risc consumul de alcool și de tutun măsurate în g/zi.





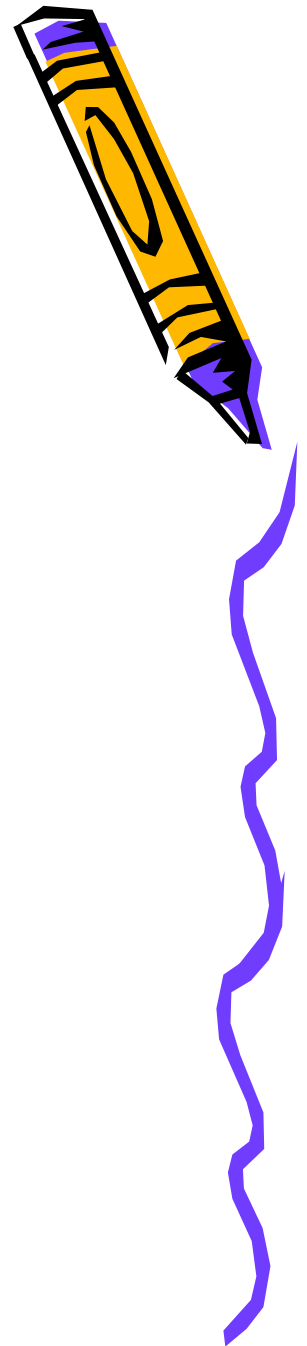
	caz	control
<i>Vârsta (ani)</i>		
25-34	1	115
35-44	9	190
45-54	46	167
55-64	76	166
65-74	55	106
75 +	13	31
medie	60,0	50,2
dev. Std	9.2	14.3





	caz	control
<i>Alcool (g / zi)</i>		
0-39	29	386
40-79	75	280
80-119	51	87
120 +	45	22
medie	84.9	44.4
dev. std.	48.4	31.9





	caz	control
<i>Tutun (g / zi)</i>		
0-9	78	447
10-19	58	178
20-29	33	99
30 +	31	51
medie	16.7	10.5
dev. std.	12.9	11.9



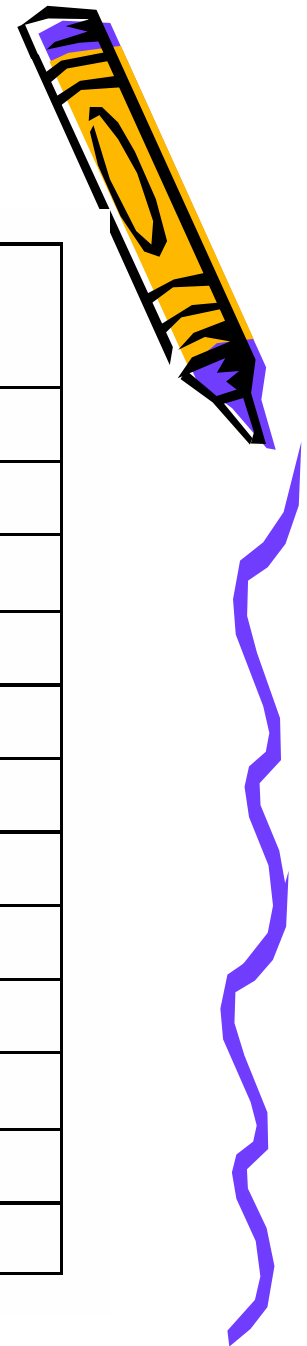


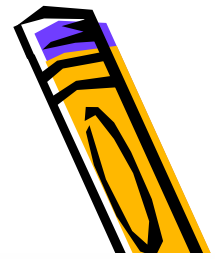
Avem șase categorii de vârstă iar factorul de risc principal este consumul de alcool.

A fost ales un prag al consumului de alcool de 80 g/zi, astfel încât în ceea ce privește expunerea, subiecții cu valori sub acest prag au fost codati cu 0 iar cei cu valori peste acest prag au fost codati cu 1.



Strat vârstă	expunere	caz	Total(caz + control)
1	1	1	10
1	0	0	106
2	1	4	30
2	0	5	169
3	1	25	54
3	0	21	159
4	1	42	69
4	0	34	173
5	1	19	37
5	0	36	124
6	1	5	5
6	0	8	39





În urma procesării computerizate (STATISTICA), s-a obținut riscul relativ (raportul șanselor) corespunzător consumului de alcool egal cu 5.31 iar intervalul corespunzător de încredere de nivel 95% este (3.66, 7.71), de unde tragem următoarea concluzie:

„există un risc semnificativ ridicat dat de depășirea limitei de 80 g/zi în declanșarea cancerului esofagian” (deoarece 1 nu aparține acestui interval).

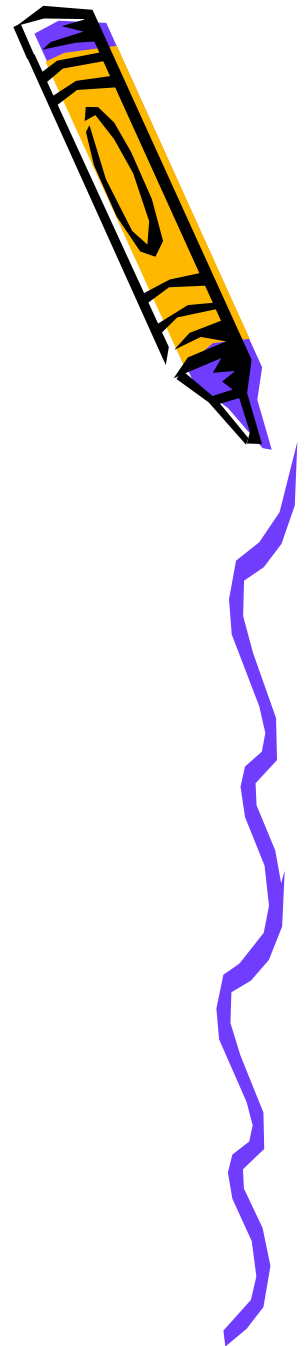






Se pot face analize regresive mai complexe, care să țină seama și de acțiunea concertată a doi sau mai mulți factori de risc.





# Analiza discriminant

Un model de regresie logistică permite prognoza unui anumit răspuns în funcție de o serie de factori predictivi, realizând în acest mod o clasificare a unor subiecți/obiecte în două clase distincte în raport cu acest răspuns.

Această împărțire a unui grup de obiecte în două categorii distincte este cunoscută în Statistică sub numele de *Analiza discriminant*.





Putem utiliza modelul regresiv logistic pentru calcularea *indicelui pronostic* (sau *index diagnostic* în medicină).

Definind  $L$  ca fiind log-ul șansei  $p$ , rezultă că:

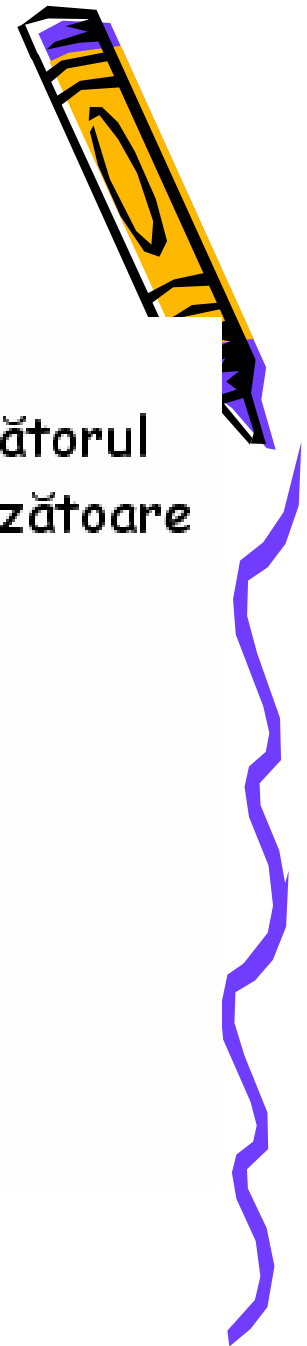
$$L = \ln \frac{p}{1-p} = b_0 + b_1 \cdot X_1 + \dots + b_k \cdot X_k$$

unde avem  $k$  variabile explicative în model.



Astfel, în cazul primului exemplu, considerând doar factorii predictivi *obezitatea* și *vârsta*, obținem următorul tabel cu valorile lui  $L$ , precum și proporțiile corespunzătoare de pacienți.

Obezitate	Vârsta	$L$	$p$	Proporție observată
Nu	Nu	-2.392	8%	0.09% (7/77)
Da	Nu	-1.697	15%	0.09% (1/11)
Nu	Da	-1.526	18%	0.18% (48/272)
Da	Da	-0.831	30%	0.31% (23/74)





Se observă din acest tabel că, de exemplu, riscul hipertensiunii este mare (30%) în cazul subiecților obezi și în vârstă de peste 40 de ani, în timp ce în cazul subiecților cu greutatea normală și sub 40 de ani, riscul scade la 8%.





Am considerat doar cazul în care variabilele explicative sunt de asemenea categoriale.

Nu există niciun impediment în a considera modelul regresiv logistic și în cazul valorilor numerice ale acestora.

Singura variabila care trebuie să fie categorială este doar variabila dependentă, adică cea care dă eticheta de clasă (categorie).

