Modele regresive

Modelele sunt reprezentări ale obiectelor sau ale situaţiilor reale. Există trei tipuri de modele:

- *modele iconice*, (iconic models) care sunt replici fizice ale obiectelor reale (de exemplu macheta unei construcţii sau un camion jucărie).
- *modele analogice,* (analog models) care nu au aceeaşi înfăţişare fizică cu obiectul ce urmează a fi modelat, dar în fond reprezintă acelaşi lucru (de exemplu, poziţia acului de la vitezometrul unui automobil reprezintă viteza vehiculului)
- *modelele matematice* reprezintă problema printr-un sistem de simboluri şi relaţii matematice.

Herbert A. Simon, deţinător al premiului Nobel pentru economie, specialist în teoria deciziilor, spunea că modelele matematice nu trebuie să fie identice cu realitatea ci cât mai apropiate de aceasta şi să conducă la rezultate mai bune decât cele obţinute prin logica bunului simţ

Un *model predictiv* creează, pe baza tuturor informaţiilor (datelor) primite, un model statistic, care face o prognoză asupra unui rezultat viitor, cu o anumită acurateţe. Modelul, construit pe baza acelor factori variabili ce influenţează rezultatul, cunoscuţi sub numele de *predictori*, va fi validat sau revizuit pe baza unor date noi. În construcţia modelului, avem de hotărât ce informaţii putem utiliza, care sunt variabilele asupra cărora putem face predicţii, cum vom determina acurateţea modelului şi, mai ales, dacă acesta poate fi folosit în lumea reală, deoarece acest tip de modelare este conceput ca să rezolve probleme reale, cu date reale.

Prognoza are drept scop producerea unei *ieşiri* (output) numerice corecte pentru o nouă *intrare* (input).

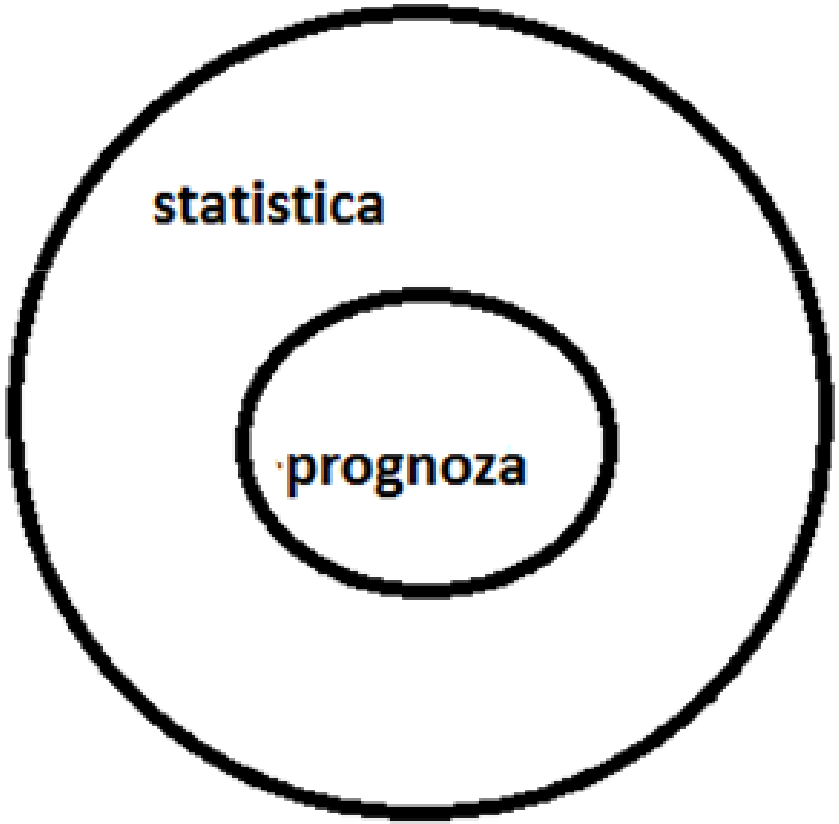
Să nu uităm însă vorbele lui Nils Bohr: „*Prognoza este foarte dificilă, mai ales dacă este vorba despre viitor*”, pentru a sublinia importanţa validării unui model.

Nu este greu să construim un model care să se potrivească perfect cu datele existente, dar important este să construim acel model care identifică acele caracteristici din datele existente, ce se vor regăsi şi în viitor.

Informaticianul şi statisticianul au păreri diferite asupra prognozei

Punctul de vedere al informaticianului:

Punctul de vedere al statisticianului:



Informaticianul:

- Ce algoritm să utilizez sau să inventez?
- Care sunt proprietățile algoritmului(complexitate, timp de rulare)?

Statisticianul:

- Care ipoteze asupra datelor sunt convenabile?
- Care este cel mai bun rezultat pe care îl putem obține in aceste condiții?
- Cum să construiesc un predictor/un algoritm care să rezolve problema?

Informaticianul:

- Mulţimea de antrenament $\Rightarrow$ Algoritm $\Rightarrow$ Prognoză
- Proprietăţile algoritmului

Statisticianul:

- Model: $(X_1, Y_1),...(X_n, Y_n) \sim P \in \mathsf{P}$
- Notând cu $\hat{m}(X)$ prognoza, trebuie să găsesc $\hat{m}$ optimal, adică pe cel care satiface condiţia:

$$\sup_{P \in P} E(Y - \hat{m}(X))^2 = \inf_{\hat{m}} \sup_{P \in P} E(Y - \hat{m}(X))^2$$

- Sunt interesat de intervalul de încredere pentru $\hat{m}(X)$

Modele regresive

Modelele regresive sunt modele statistice care stabilesc relaţia între valorile a două sau mai multe variabile aleatoare, 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.

Intuitiv, având o mulţime de valori experimentale din $\mathbf{R^2}$, considerate a fi valorile a două variabile aleatoare, încercăm să determinăm curba care este „cea mai apropiată” de punctele respective.

Metoda cea mai des utilizată în acest scop este *metoda celor mai mici pătrate*. Dacă relaţia care stabileşte legătura între variabila dependentă şi variabilele independente este una liniară, se vorbeşte despre *regresia liniară*, altfel fiind vorba de *regresia neliniară* (polinomială, exponenţială, logaritmică etc.).

Diagrama de împrăştiere

Prezentăm mecanismele cu care se pot evidenţia legăturile între secvenţe de date, provenind de la două sau mai multe variabile aleatoare.

În cazul a două serii statistice $\{x_i\}_{1\le i\le n}$ şi $\{y_i\}_{1\le i\le n}$ definite pe acelaşi lot de obiecte putem considera seria cuplurilor de observaţii $\{(x_i, y_i)\}_{1\le i\le n}$ definite de cele două variabile statistice pe acelaşi obiect i.

Acest cuplu de observaţii este reprezentat grafic folosind *norul* de puncte definit de reprezentarea bidimensională a punctelor (x_i, y_i) – aşa numita *diagramă de împrăştiere*.

1. Exemplu

Într-o fabrică s-a constatat că viteza (m/min) benzii de lucru afectează numărul de defecţiuni descoperite în timpul verificării. Următorul tabel prezintă diferite viteze ale benzii şi numărul defectelor găsite.

v	6	6	7	7	8	8	9	9	10	10	11	11	12	12	13	13
n	21	18	19	20	22	19	21	24	23	25	24	26	28	27	27	28

Prezentăm diagrama „norului” de împrăştiere a cuplurilor (viteza, număr defecţiuni) pentru lotul de 16 de observaţii:

```
>> X=[6 6 7 7 8 8 9 9 10 10 11 11 12 12 13 13];
>> Y=[21 18 19 20 22 19 21 24 23 25 24 26 25 28 27 28];
>> plot(X,Y,'ok')
```

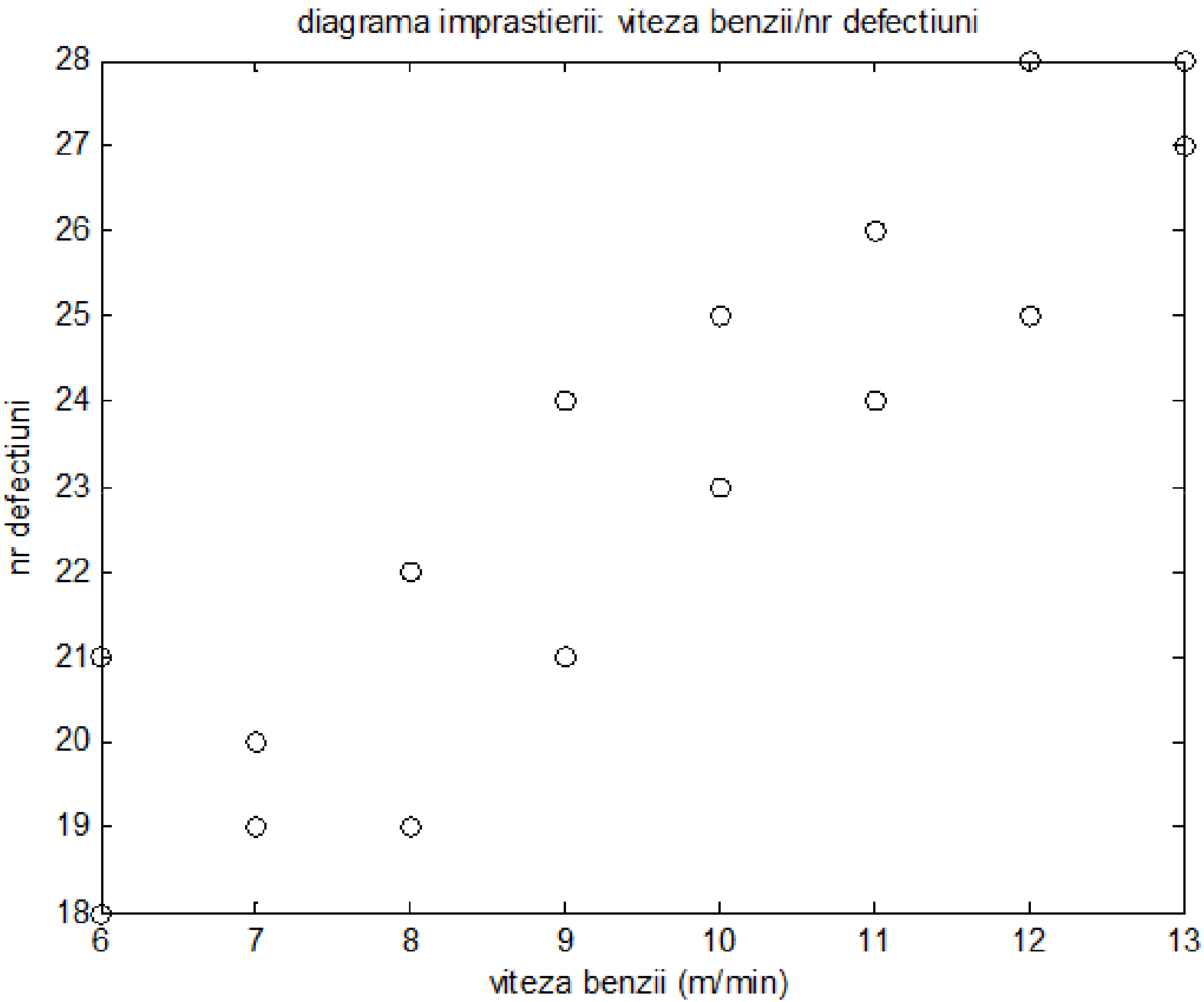


diagrama imprastierii: viteza benzii/nr defectiuni
nr defectiuni
viteza benzii (m/min)
28
27
26
25
24
23
22
21
20
19
18
6
7
8
9
10
11
12
13

Diagrama împrăştierii dă informaţii importante privind legătura între cele două serii statistice fiind astfel un instrument util într-o analiză statistică.

Dacă norul are forma unei figuri geometrice alungite, simetrică faţă de o axă, există o legătură liniară între variabile. Forma de cerc sau pătrat a norului implică independenţa variabilelor.

Este importantă alegerea unităţilor de măsură a celor două variabile pentru a nu fi modificată forma norului.

O mare parte a studiilor statistice uzuale se ocupă cu analiza relaţiei între două variabile statistice ce corespund aceluiaşi grup de subiecţi/obiecte.

Cel mai cunoscut exemplu se referă la relaţia ce există între înălţimea şi greutatea unui individ ce corespunde unor anumite standarde geografice, rasiale etc. Pentru a o identifica, se studiază relaţia dintre cele două caracteristici măsurate pe indivizii dintr-un anumit lot.

Suntem interesaţi în a descrie relaţia care ar putea exista între cele două variabile, analizând legătura între cele două serii de observaţii.

Concret, se analizează dacă tendinţa ascendentă a uneia implică o tendinţă ascendentă, descendentă sau nici o tendinţă a celeilalte. Scopul final, în ipoteza existenţei unei legături reale între cele două variabile, este *prognoza* valorilor uneia în raport cu valorile celeilalte pe baza ecuaţiei de regresie.

Coeficientul de corelaţie

Posibilele asociaţii între valorile a două variabile statistice continue, prelevate de la acelaşi grup de subiecţi, sunt date de *coeficientul de corelaţie*.

Acest coeficient poate fi calculat pentru orice set de date.

Coeficientul de corelaţie are relevanţă statistică, dacă sunt îndeplinite următoarele două condiţii:

1. cele două variabile sunt definite de acelaşi lot de obiecte, cuplurile de date corespunzând aceluiaşi obiect;
2. cel puţin una din variabile să aibă o repartiţie aproximativ normală.

Considerăm, două serii statistice $\{x_i\}_{1\le i\le n}$ şi $\{y_i\}_{1\le i\le n}$ corespunzătoare variabilelor statistice X şi Y, generate de un grup de obiecte.

Prin *coeficientul de corelaţie* r al celor două variabile, numit şi *Pearson's* r, vom înţelege numărul real r, cuprins între -1 şi 1, definit de formula:

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

unde $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Formula poate fi 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

Calculăm coeficientul de corelaţie a celor două variabile din exemplul nr 1

```
>> X=[6 6 7 7 8 8 9 9 10 10 11 11 12 12 13 13];
>> Y=[21 18 19 20 22 19 21 24 23 25 24 26 25 28 27 28];
>> r =(X*Y'-16*mean(X)*mean(Y))/sqrt((X*X'-16*mean(X)^2)*
(Y*Y'-16*mean(Y)^2))
r =
    0.9151
```

Construcţia intervalului de încredere 95% pentru *r*

Intervalul de încredere 95% pentru *r* este intervalul care-l conţine pe *r* cu o probabilitate de 0.95.

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

Aplicând transformarea inversă, obţinem intervalul de încredere 95% pentru *r*, dat de: $\left(\frac{e^{2z_1}-1}{e^{2z_1}+1}, \frac{e^{2z_2}-1}{e^{2z_2}+1}\right)$

Exemplu

Construim intervalul de încredere 95% pentru r din exemplul nr 1

```
>> r = 0.9274
>>z=1/2*log((1+r)/(1-r))
z =
        1.6396
>> z1=z-1.96/sqrt(17)
z1 =
        1.1642
>> z2=z+1.96/sqrt(17)
z2 =
        2.1149
>> a=(exp(2*z1)-1)/(exp(2*z1)+1)
a =
        0.8224
>> b=(exp(2*z2)-1)/(exp(2*z2)+1)
b =
        0.9713
```

Aşadar, în acest caz intervalul de încredere 95% pentru r este:
(0.8224, 0.9713).

Coeficientul de corelaţie *r* (Pearson) ia valori cuprinse între –1 şi +1.

- O valoare a lui *r* apropiată de +1 indică o corelaţie pozitivă puternică, (tendinţa de creştere puternică a unei variabile atunci când cealaltă variabilă creşte)
- O valoare a lui *r* apropiată de -1 indică o corelaţie negativă puternică, (tendinţa de descreştere puternică a unei variabile atunci când cealaltă variabilă creşte)
- Valoarea 0 a coeficientului de corelaţie indică independenţa liniară între cele două variabile.

Coeficientul de corelaţie *r* (Pearson) măsoară cât de ***puternică*** este legătura dintre două variabile.

Este foarte important cât de ***semnificativă*** este această legătură dintre două variabile.

În statistică un rezultat este *statistic semnificativ*, dacă este improbabil să fie obţinut din întâmplare.

De obicei *nivelul de semnificaţie* (prag de probabilitate) mai mic de 0.05 înseamnă că rezultatul observat are cel mult 5% şanse de a fi obţinut printr-o întâmplare, ceea ce ne permite să-l considerăm semnificativ.

Dacă inegalitatea:

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

este adevărată, putem considera că cele două variabile sunt într-adevăr corelate.

În exemplul nr 1 avem:

```
>>r*sqrt(15)
ans =
        3.5442
```

ceea ce înseamnă că legătura între viteza benzii de lucru şi numărul de defecţiuni este semnificativă.

Pentru a stabili cât de semnificativă este legătura dintre cele două variabile, se utilizează nivelul de semnificaţie *p* asociat calculării coeficientului *r* :

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

Putem calcula în Matlab matricea coeficienţilor de corelaţie şi matricea nivelului de semnificaţie *p* folosind funcţia **corrcoef**:

În exemplul nr 1 avem:

```
>> [r,p]=corrcoef(X,Y)
r =
    1.0000    0.9274
    0.9274    1.0000
p =
    1.0000    0.0000
    0.0000    1.0000
```

rezultând r = 0.9274 şi p=0.

Cu cât este mai mic nivelul de semnificaţie, cu atât este mai *semnificativă* legătura.

Pe de altă parte, cu cât valoarea lui $|r|$ este mai apropiată de 1, cu atât mai *puternică* este legătura.

Factorul esenţial este dimensiunea eşantionului.

- În cazul eşantioanelor mici se poate obţine din întâmplare o corelaţie puternică; înainte de a concluziona este absolut necesară consultarea nivelului de semnificaţie *p*;
- În cazul eşantioanelor mari este uşor de obţinut un nivel de semnificaţie mic şi astfel este necesar să se determine cât de puternică este legătura.

Prezentarea corelaţiei

Prezentarea corelaţiei dintre două variabile statistice trebuie să urmeze următorul model:

1. Se prezintă mai întâi diagrama de împrăştiere a norului de puncte.
2. Valoarea coeficientului de corelaţie r trebuie să aibă cel puţin două zecimale şi să fie însoţită de nivelul de semnificaţie p şi de intervalul de încredere corespunzător, dacă este posibil.
3. Numărul de observaţii analizate trebuie menţionat.

Analiza matricei corelaţiilor

Luând în considerare un set de variabile (atribute), $X_1, X_2, ..., X_k$, $k > 2$, matricea de tip $k \times k$ al cărui element (i,j) este coeficientul de corelaţie al variabilelor (X_i, X_j) se numeşte *matricea corelaţiilor*.

La fel ca şi în cazul cuplurilor de variabile, putem analiza corelaţiile între fiecare două perechi de variabile, precum şi ‚norul' împrăştierii lor .

Avantajul prezentării legăturii între atributele obiectelor cu ajutorul matricei corelaţiilor şi nu cu corelaţiile fiecărei perechi de atribute în parte constă în faptul că astfel avem o privire de ansamblu asupra tuturor conexiunilor între atributele analizate.

De asemenea, ‚norul' punctelor reprezentate, ca si alte reprezentări grafice sunt mult mai sugestive în cazul reprezentării colective a datelor decât luate separat.

2. Exemplu

Într-o firmă de IT, se presupune că Y, rezultatul vânzărilor, depinde de X1, suma cheltuită pentru reclama tradiţională la TV şi în ziare, şi X2, suma cheltuită pentru reclama prin Internet, (sumele sunt în zeci de mii de euro).

X1	5	7	7	5	9	11	12	13	17	17	18	19
X2	2.9	2.5	3	1.5	4	4.5	5	5.5	6	7.5	8	7
Y	1	2	2.5	1.5	3.5	4	5	6	6.5	8	8	7.5

Prezentăm o asemenea analiză multiplă (multivariată) a acestor date.

Studiem cât de puternică este legătura între variabila răspuns Y şi variabilele predictive X1, X2, scriind matricea corelaţiilor şi calculând nivelul de semnificaţie *p*:

```
>> Y=[5 7 7 5 9 11 12 13 15 17 18 19];
>> X1=[2 2.5 3 1.5 4 4.5 5 5.5 6 7.5 8 7];
>>X2=[1 2 2.5 1.5 3.5 4 5 6 6.5 8 8 7.5];
>> A=[X1' X2' Y'];[r,p]=corrcoef(A)
r =
    1.0000    0.9899    0.9825
    0.9899    1.0000    0.9858
    0.9825    0.9858    1.0000
p =
    1.0000    0.0000    0.0000
    0.0000    1.0000    0.0000
    0.0000    0.0000    1.0000
```

Diagrama împrăştierii

```
>> plot3(X1,X2,Y,'o')
>> grid on
>> axis square
```

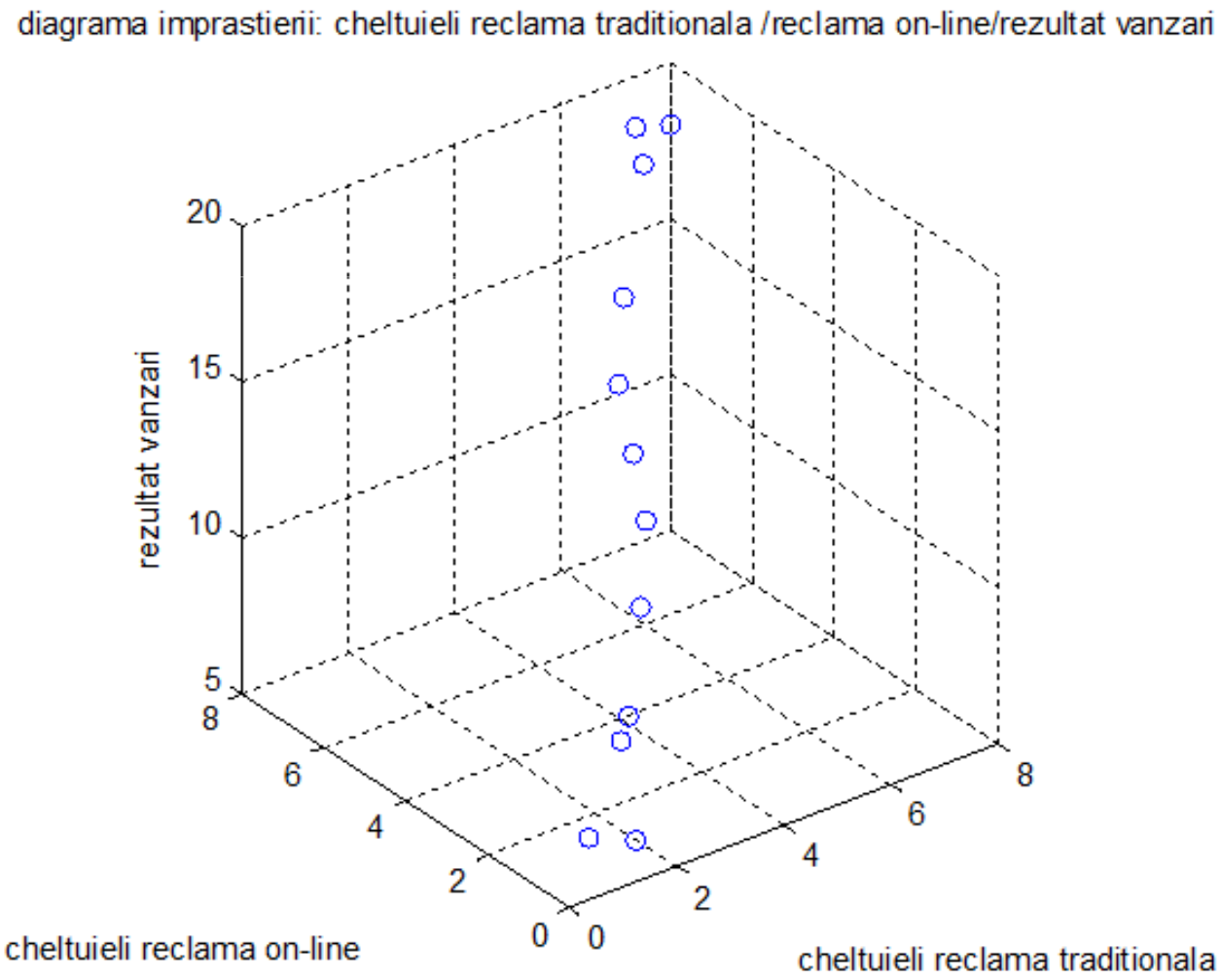
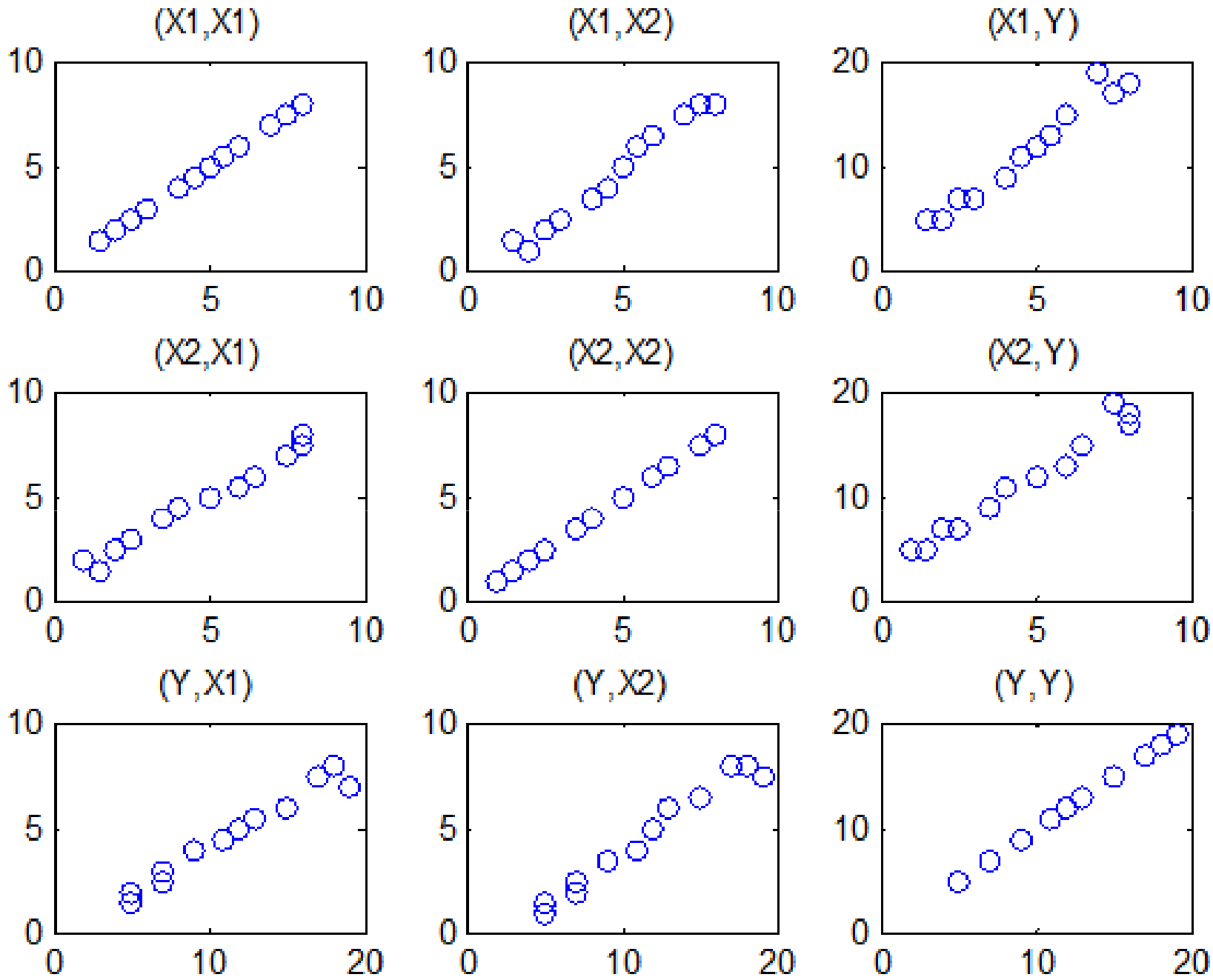


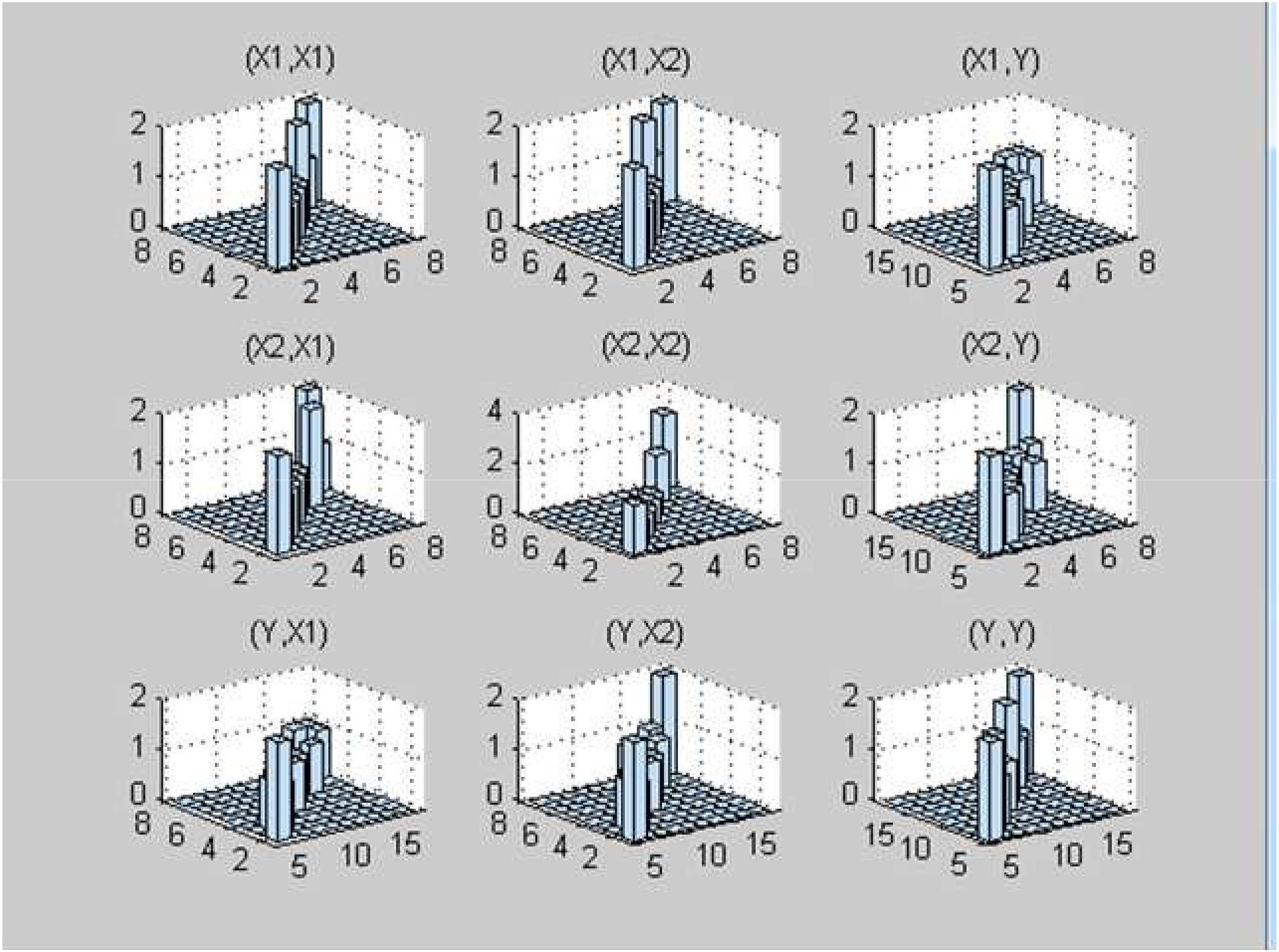
Diagrama corelaţiilor multiple

În ceea ce priveşte ilustrarea grafică a corelaţiilor de mai sus, vizualizarea este foarte utilă, rezumând sugestiv toate informaţiile numerice prezentate arid în tabel

```
>> subplot(331);plot(X1,X1,'o');
>> subplot(332);plot(X1,X2,'o');
>> subplot(333);plot(X1,Y,'o');
>> subplot(334);plot(X2,X1,'o');
>> subplot(335);plot(X2,X2,'o');
>> subplot(336);plot(X2,Y,'o');
>> subplot(337);plot(Y,X1,'o');
>> subplot(338);plot(Y,X2,'o');
>> subplot(339);plot(Y,Y,'o');
```


(X1,X1)
(X1,X2)
(X1,Y)
(X2,X1)
(X2,X2)
(X2,Y)
(Y,X1)
(Y,X2)
(Y,Y)
0
5
10
20

```
>> subplot(331);A1=[X1' X1'];hist3(A1)
>> subplot(332);A2=[X1' X2'];hist3(A2)
>> subplot(333);A3=[X1' Y'];hist3(A3)
>> subplot(334);A4=[X2' X1'];hist3(A4)
>> subplot(335);A5=[X2' X2'];hist3(A5)
>> subplot(336);A6=[X2' Y'];hist3(A6)
>> subplot(337);A7=[Y' X1'];hist3(A7)
>> subplot(338);A8=[Y' X2'];hist3(A8)
>> subplot(339);A9=[Y' Y'];hist3(A9)
```


(X1,X1)
(X1,X2)
(X1,Y)
(X2,X1)
(X2,X2)
(X2,Y)
(Y,X1)
(Y,X2)
(Y,Y)

Regresia liniară

Regresia liniară este tehnica statistică ce permite modelarea relaţiei liniare între o variabilă explicativă (notată cu X) şi variabila ce urmează a fi explicată (notată cu Y).

Teoremă

Coeficientul de corelaţie r a două variabile aleatoare X şi Y ia valori în intervalul [-1, 1].
Dacă variabilele sunt independente coeficientul de corelaţie este nul.
Coeficientul de corelaţie r este egal cu ± 1 dacă şi numai dacă variabilele X şi Y verifică ecuaţia:

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

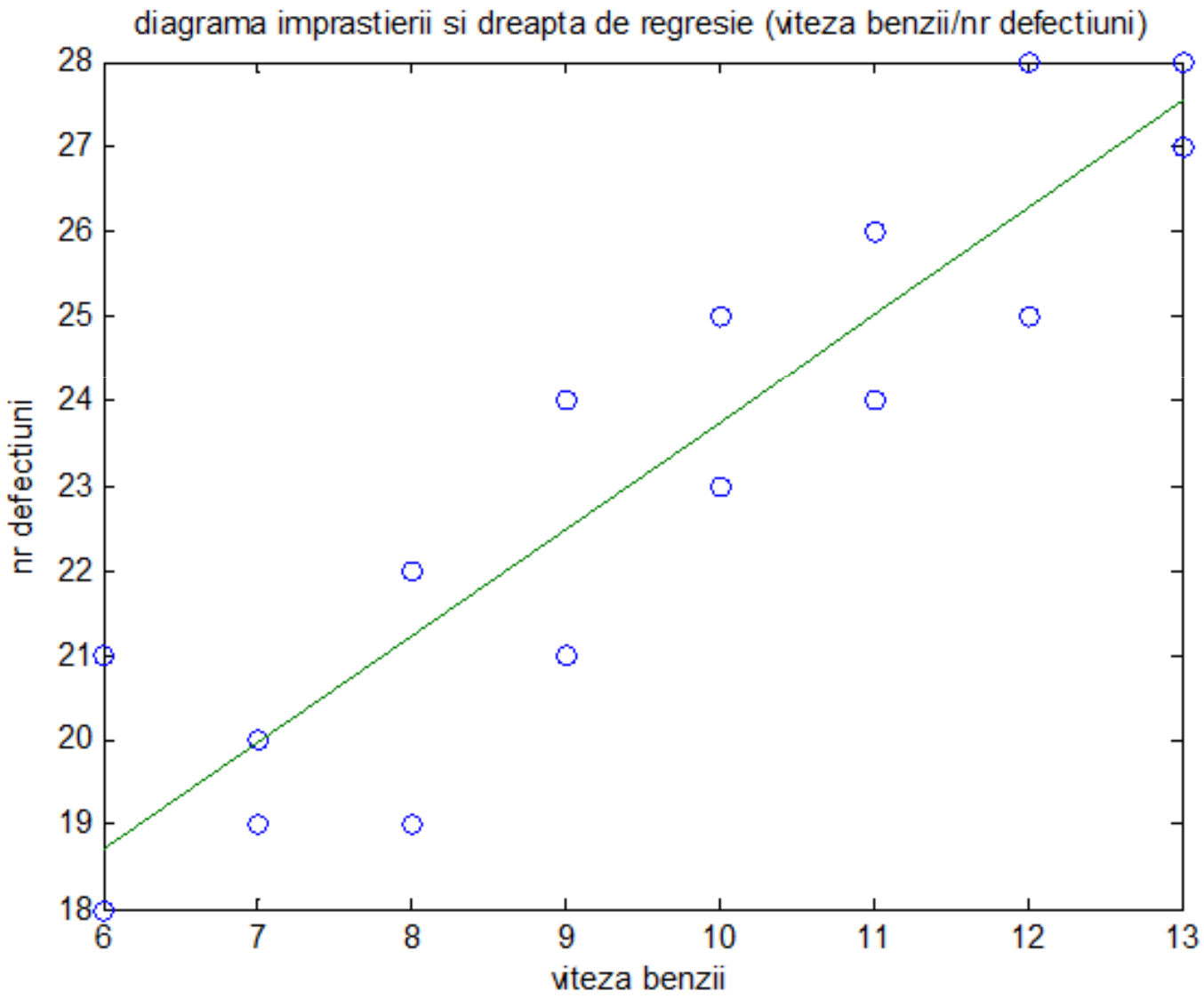
Reciproca nu este totdeauna adevărată, adică există variabile dependente având coeficientul de corelaţie zero.

Ecuaţia de regresie

În cazul în care există o legătură liniară între două variabile (numerice), se consideră una dintre variabile ca *variabilă independentă* (*predictor)*, iar cealaltă variabilă ca *variabilă dependentă* (*răspuns*).

Legătura liniară dintre cele două variabile este descrisă de o ecuaţie liniară, așa-numita *ecuaţia de regresie*, căreia îi corespunde geometric *dreapta de regresie*.

În exemplul nr 1 avem:



Dreapta de regresie

Dreapta de regresie este acea dreaptă (ideală) ce trece prin norul de puncte format de perechile de date ale celor două variabile şi care minimizează distanţa între date şi ea, prin minimizarea sumei pătratelor distanţelor. Ecuaţia de dreptei regresie este de forma:

$$y = a + bx$$

unde a se numeşte *interceptor*, iar b se numeşte *coeficientul de regresie* .

Ecuaţia dreptei de regresie se stabileşte pe baza metodei "*celor mai mici pătrate*", care în cazul nostru constă în calcularea distanţelor dintre punctele observate (reale) corespunzătoare cuplului de serii statistice şi punctele (imaginare) de pe o anumită dreaptă (de regresie), ce trece prin mijlocul norului de puncte generate de cuplurile de date, distanţe cunoscute sub numele de *reziduuri* .

Datele existente satisfac ecuaţiile:

$$y_i = a + bx_i + \varepsilon_i \quad 1 \le i \le p$$

unde:

- x_i şi y_i, $1 \le i \le p$ sunt numere reale cunoscute;
- ε_i sunt erorile.

Metoda celor mai mici pătrate constă în aflarea parametrilor a şi b pentru care expresia $S(a,b) = \sqrt{\sum_{i=1}^{p} (y_i - a - b \cdot x_i)^2}$ este minimă.

Regresia liniară este utilizată, dacă sunt îndeplinite următoarele trei ipoteze de lucru :

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

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.

Dreapta de regresie este dată de ecuaţia:

$$y = a + b \cdot x \text{, unde:}$$

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

cu notaţiile $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

3. Exemplu

Următoarea baza de date conţine informaţii despre preţul de vânzare a 15 apartamente dintr-un cartier Ω al unui oraş Z, în trimestrul al III-lea din 2011. Prezentăm datele referitoare la preţ (în mii de euro) şi suprafaţa locuibilă (în metri pătraţi)

supr	30	32	35	48	49	50	55	60	65	70	75	80	82	85	90
preţ	30	29	31	36	38	40	48	47	53	49	58	60	59	71	65

```
>> X1=[ 30 32 35 48 49 50 55 60 65 70 75 80 82 85 90];
>> Y=[30 29 31 36 38 40 48 47 53 49 58 60 59 71 65];
>> A=[X1' Y'];[r p]=corrcoef(A)
r =
   1.0000   0.9729
   0.9729   1.0000
p =
   1.0000   0.0000
   0.0000   1.0000
```

Se observă că variabilele sunt puternic (pozitiv) corelate, coeficientul de corelaţie fiind 0.9729 şi nivelul de semnificaţie este $p = 0.0000$.

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

```
>> b=(X1*Y'-15*mean(X1)*mean(Y))/((norm(X1))^2-15*(mean(X1))^2)
b =
   0.6572
>> a=mean(Y)-b*mean(X1)
a =
   7.9053

>> plot(X1,a+b*X1);hold on
>> plot(X1,Y,'ko');hold off
```

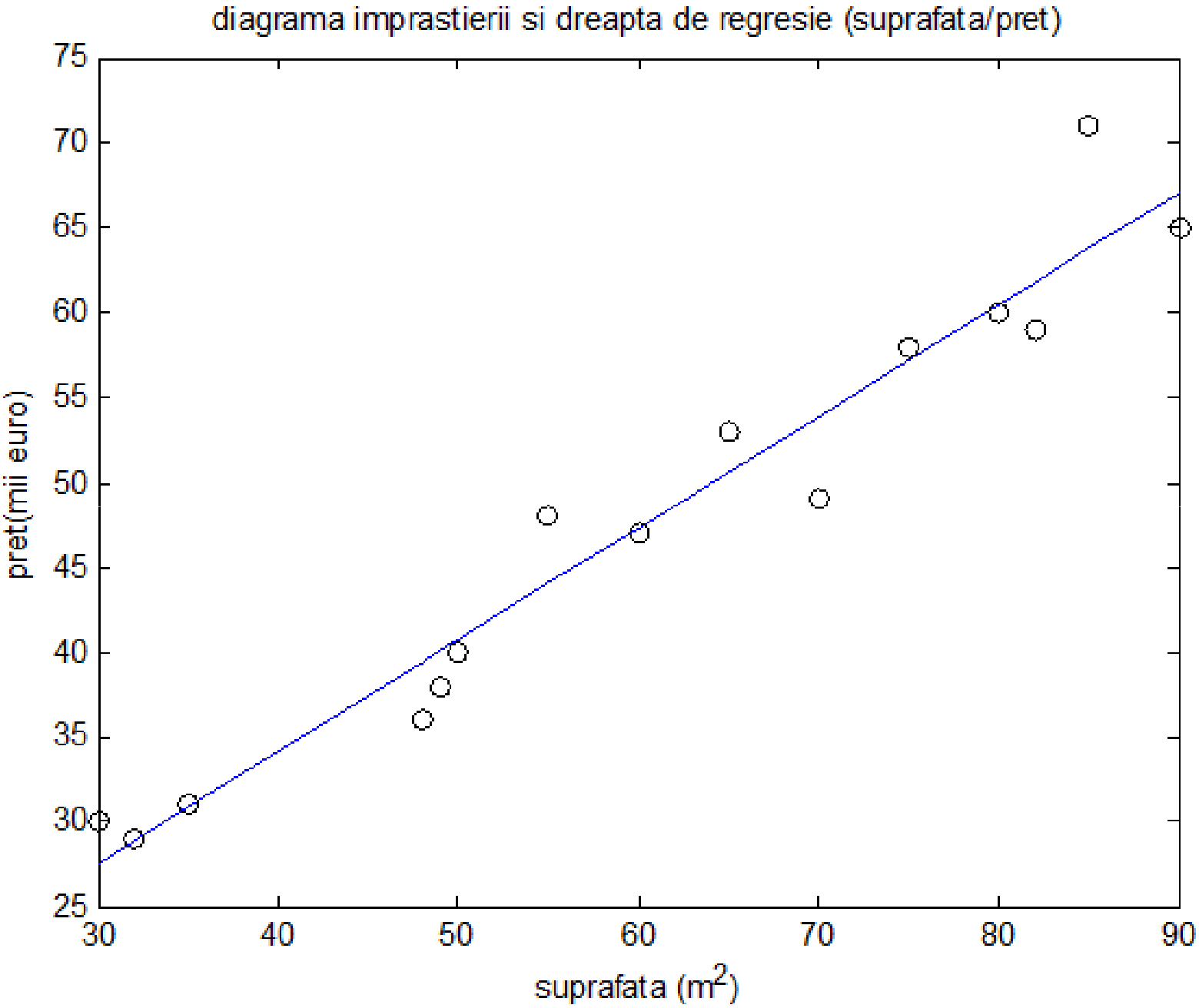


diagrama imprastierii si dreapta de regresie (suprafata/pret)
pret(mii euro)
suprafata (m2)
75
70
65
60
55
50
45
40
35
30
25
30
40
50
60
70
80
90

Dreapta de regresie obţinută poate fi utilizată în prognoză.

De exemplu, pe baza datelor avute la dispoziţie, preţul de vânzare al unui apartament de $72m^2$, reaspectiv de 87 ar fi:

```
>> syms X1
>> f=a+b*X1;\

>> p1=subs(f,X1,72)
p1 =
  55.2235

>> p2=subs(f,X1,87)
p2 =
  65.0814
```

Pentru calculul coeficienților a și b putem apela la scrierea matricială a ecuațiilor

$$y_i = a + bx_i + \varepsilon_i \quad 1 \le i \le p,$$

și anume:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_p \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix}.$$

Presupunând că matricea $\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_p \end{pmatrix}$ este de rang 2, utilizând metoda celor mai mici pătrate, putem calcula parametrii a și b:

$$\begin{pmatrix} a \\ b \end{pmatrix} = \left(\left(\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_p \end{pmatrix}' \cdot \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_p \end{pmatrix} \right)^{-1} \cdot \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_p \end{pmatrix}' \cdot \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} \right)$$

Exemplu

Reluăm exemplul nr.3

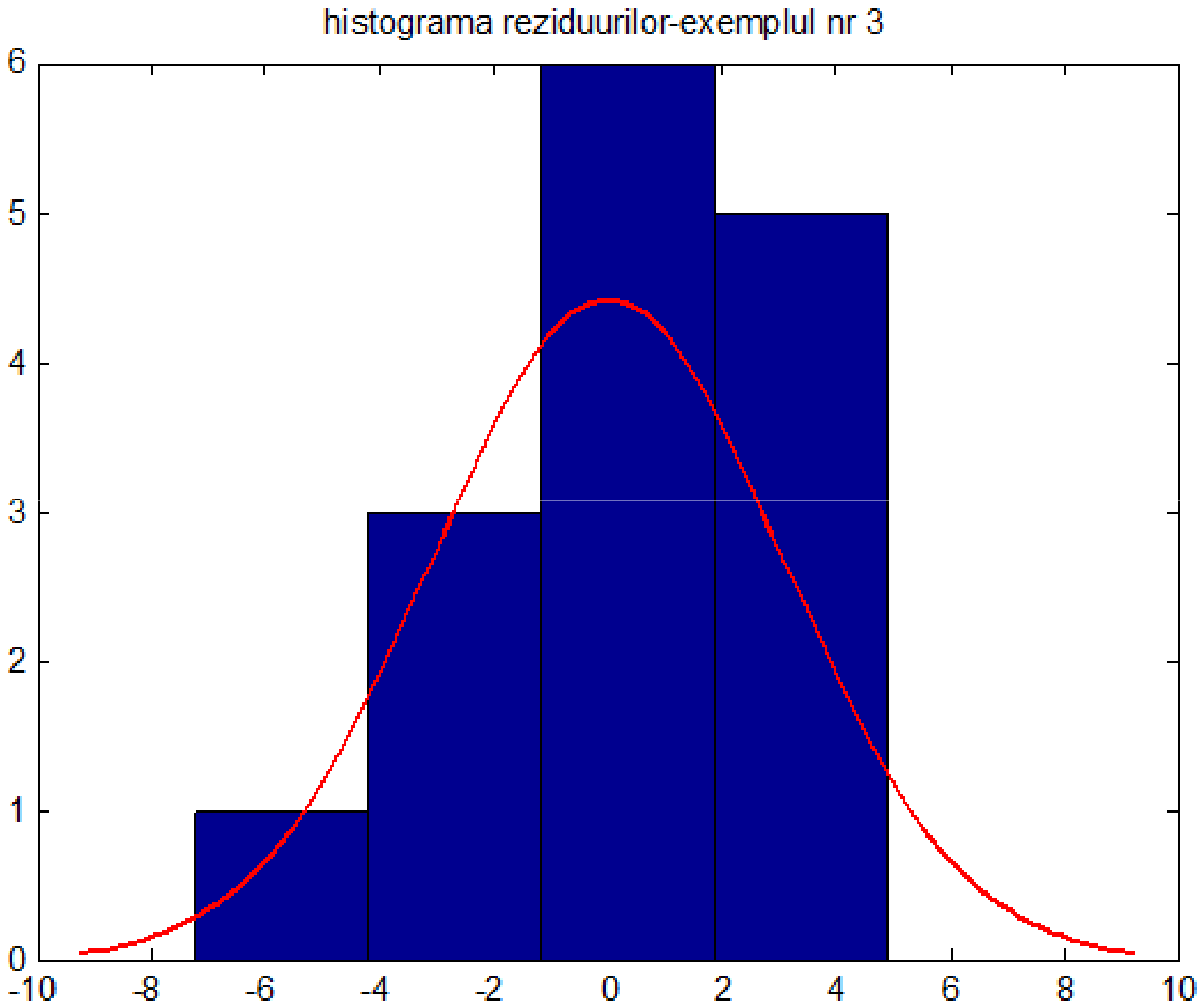
```
>> X1=[ 30 32 35 48 49 50 55 60 65 70 75 80 82 85 90];
>> i=1; B=[1]; for i=2:15 B=[1 [B]]; end
>> X=[B' X1'];
>> a=inv(X'*X)*X'*Y'
a =
   7.9053
   0.6572
```

În Matlab, problema se poate rezolva mai simplu:

```
>> a=X\Y'
a =
  7.9053
  0.6572
```

Vom calcula reziduurile, media acestora şi vom desena histograma lor, suprapunând graficul densităţii de repartiţie normală

```
>> Rez=a(1)+a(2)*X1-Y;
>> mean(Rez)
ans =
 -2.2974e-01
>>histfit(Rez)
```


histograma reziduurilor-exemplul nr 3
6
5
4
3
2
1
0
-10
-8
-6
-4
-2
0
2
4
6
8
10

Regresia neliniară

În cazul când legătura dintre cele două variabile statistice nu este liniară şi totuşi există, avem de a face cu o *regresie neliniară*.

Variabila dependentă este o combinaţie neliniară a variabilelor independente.

Ne vom limita la modelarea relaţiei dintre o variabilă explicativă şi variabila ce urmează a fi explicată printr-o ecuaţie a unei curbe de regresie:

- ecuaţie polinomială,
- ecuaţie mixt exponenţială.

Regresia polinomială

Regresia polinomială va furniza o ecuaţie de forma:

$$Y = b_0 + b_1 X + b_2 X^2 + ... + b_n X^n ,$$

unde, ca de obicei, X este variabila predictoare, iar Y este variabila prognozată, adică variabila răspuns. Pentru a determina coeficienţii de regresie şi interceptorul vom aplica *metoda celor mai mici pătrate*:

Presupunem că datele existente satisfac ecuaţiile:

$$Y_i = b_0 + b_1 \cdot X_i + b_2 \cdot X_i^2 + ... + b_n \cdot X_i^n + \varepsilon_i , \quad 1 \le i \le p ,$$

unde:

- $X_i, Y_i, \quad 1 \le i \le p$ sunt numere reale cunoscute (reţinem că $p > n$),
- ε_i sunt variabile aleatoare necunoscute.

Sub formă matriceală, modelul devine:

$$Y = X \cdot b + \varepsilon ,$$

unde:

- $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix}$ este un vector aleator p-dimensional,
- $X = \begin{pmatrix} 1 & X_1 & \dots & X_1^n \\ 1 & X_2 & \dots & X_2^n \\ \dots & \dots & \dots & \dots \\ 1 & X_p & \dots & X_p^n \end{pmatrix}$,
- b este vectorul $(n+1)$-dimensional, ale cărui componente sunt parametrii necunoscuți ai modelului,
- ε este vectorul p-dimensional al erorilor.

Metoda celor mai mici pătrate constă în aflarea vectorului b, pentru care expresia:

$$S(b) = \|Y - Xb\|^2 = \sqrt{\sum_{i=1}^{p} (Y_i - b_0 - b_1 \cdot X_i - b_2 \cdot X_i^2 - ... - b_n \cdot X_i^n)^2} \ ,$$

este minimă, presupunând că matricea X are rangul n. Astfel $X' \cdot X$ este inversabilă şi avem:

$$b = (X' \cdot X)^{-1} \cdot X' \cdot Y \ .$$

4. Exemplu

Considerăm situaţia numărului de pacienţi internaţi într-o secţie a unui spital pe parcursul a 11 săptămâni, situaţie reflectată în următorul tabel:

săptămâna	1	2	3	4	5	6	7	8	9	10	11
nr pacienţi	50	92	116	125	135	150	145	130	120	95	85

Determinăm curba de regresie polinomială, considerând pentru început că datele existente satisfac relaţiile:

$$Y_i = b_0 + b_1 \cdot t_i + b_2 \cdot t_i^2 + \varepsilon_i, \quad 1 \le i \le 6 .$$

Utilizând metoda celor mai mici pătrate, vom rezolva ecuaţia matriceală:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \\ Y_{10} \\ Y_{11} \end{pmatrix} = \begin{pmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \\ 1 & t_4 & t_4^2 \\ 1 & t_5 & t_5^2 \\ 1 & t_6 & t_6^2 \\ 1 & t_7 & t_7^2 \\ 1 & t_8 & t_8^2 \\ 1 & t_9 & t_9^2 \\ 1 & t_{10} & t_{10}^2 \\ 1 & t_{11} & t_{11}^2 \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}; \text{ notam } X = \begin{pmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \\ 1 & t_4 & t_4^2 \\ 1 & t_5 & t_5^2 \\ 1 & t_6 & t_6^2 \\ 1 & t_7 & t_7^2 \\ 1 & t_8 & t_8^2 \\ 1 & t_9 & t_9^2 \\ 1 & t_{10} & t_{10}^2 \\ 1 & t_{11} & t_{11}^2 \end{pmatrix}$$

și astfel obținem coeficienții polinomului de gradul al doilea:

$$b = (X' \cdot X)^{-1} \cdot X' \cdot Y \ ,$$

```
>> t=1:11; P=[50 92 116 125 135 150 145 130 120 95 85];
>> i=1;B=1;for i=2:11 B=[1 B];end
>> for i=1:11 t2(i)=t(i)^2;end
>> X=[B' t' t2'];b2=X\P';
>> X=[B' t' t2'];b2=X\P'
b2 =
   21.0848
   38.9000
   -3.0758
```

Desenăm diagrama împrăștierii și polinomul de regresie obținut:

```
>>P2=b2(1)+b2(2)*t+b2(3)*t.^2;plot(t,P,'O',t,P2)
```

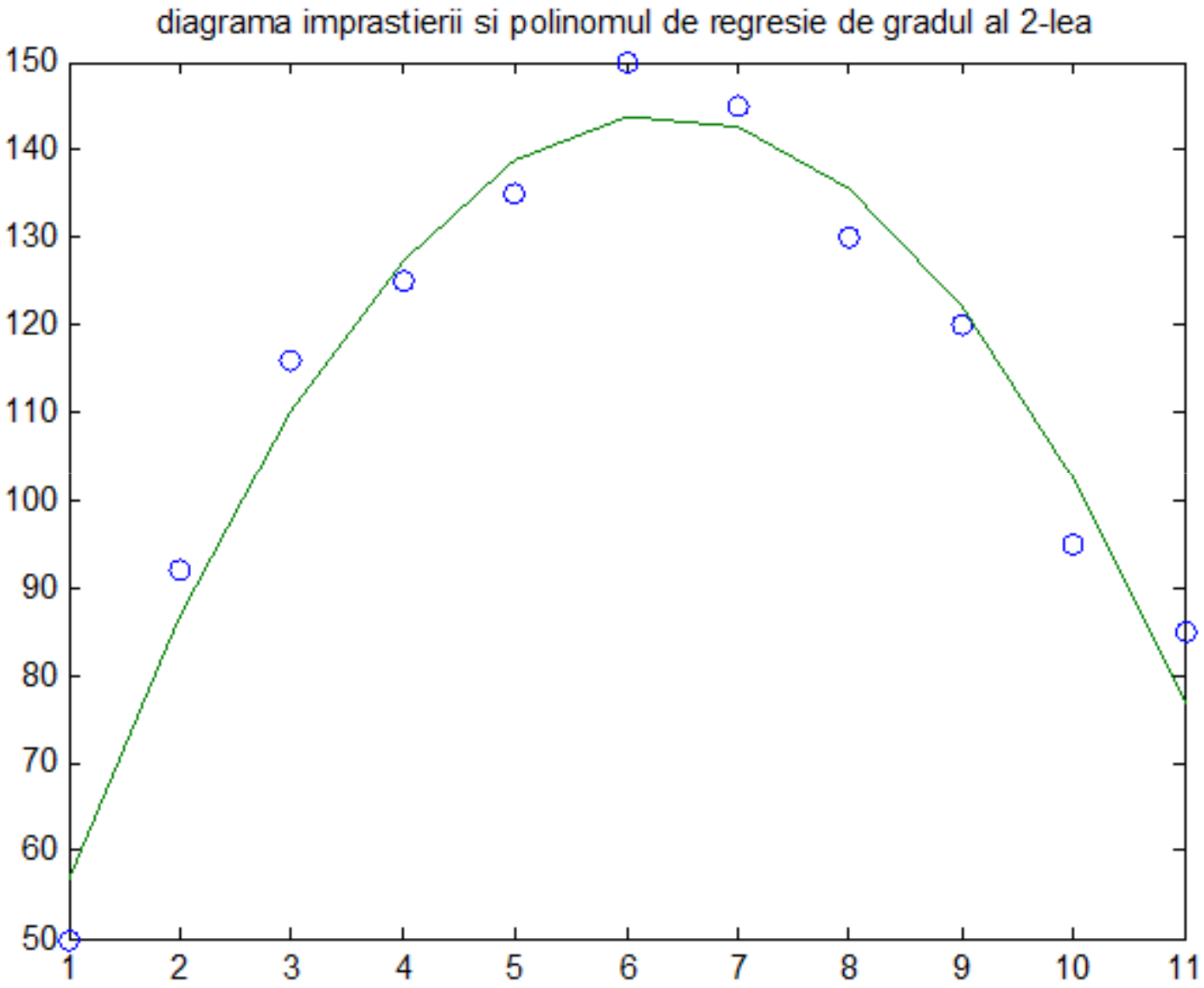
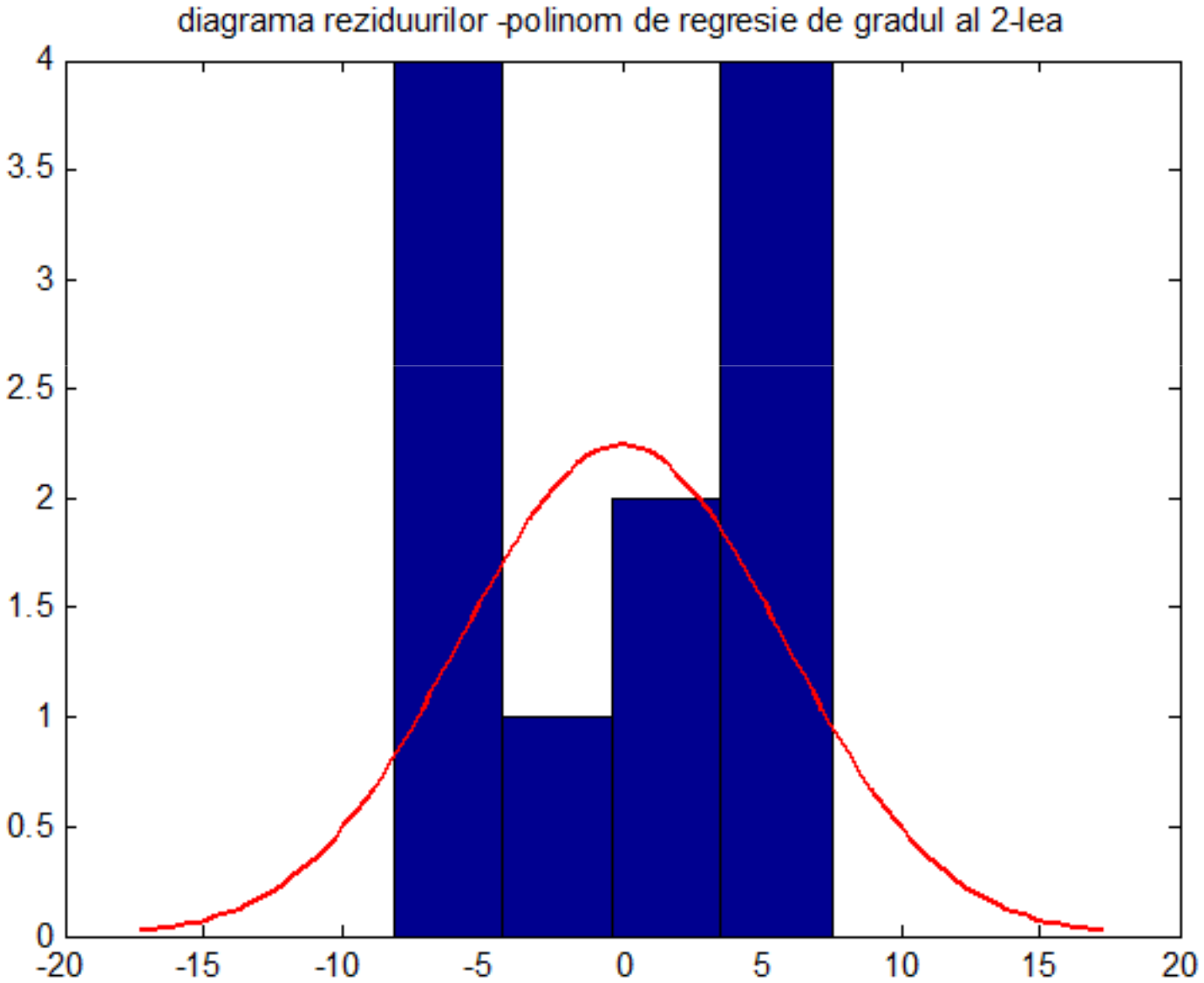


diagrama imprastierii si polinomul de regresie de gradul al 2-lea
150
140
130
120
110
100
90
80
70
60
50
1
2
3
4
5
6
7
8
9
10
11

```
>> Rez=P2-P;
>> mean(Rez)
ans =
  3.6173e-014
>> histfit(Rez)
```



Încercăm să determinăm un polinom de regresie de grad 4, ceea ce înseamnă să presupunem că:

$$Y_i = b_0 + b_1 \cdot t_i + b_2 \cdot t_i^2 + b_3 \cdot t_i^3 + b_4 \cdot t_i^4 + \varepsilon_i, \quad 1 \le i \le 6,$$

```
>> t=1:11; P=[50 92 116 125 135 150 145 130 120 95 85];
>> i=1;B=1;for i=2:11 B=[1 B];end
>> for i=1:11 t2(i)=t(i)^2;end
>> for i=1:11 t3(i)=t(i)^3;end
>> for i=1:11 t4(i)=t(i)^4;end
>> X=[B' t' t2' t3' t4'];b4=X\P'
b4 =
   11.5152
   44.8918
   -3.4257
   -0.1249
    0.0108
>> P4=b4(1)+b4(2)*t+b4(3)*t.^2+ b4(4)*t.^3+b4(5)*t.^4;plot(t,P,'O',t,P4)
```

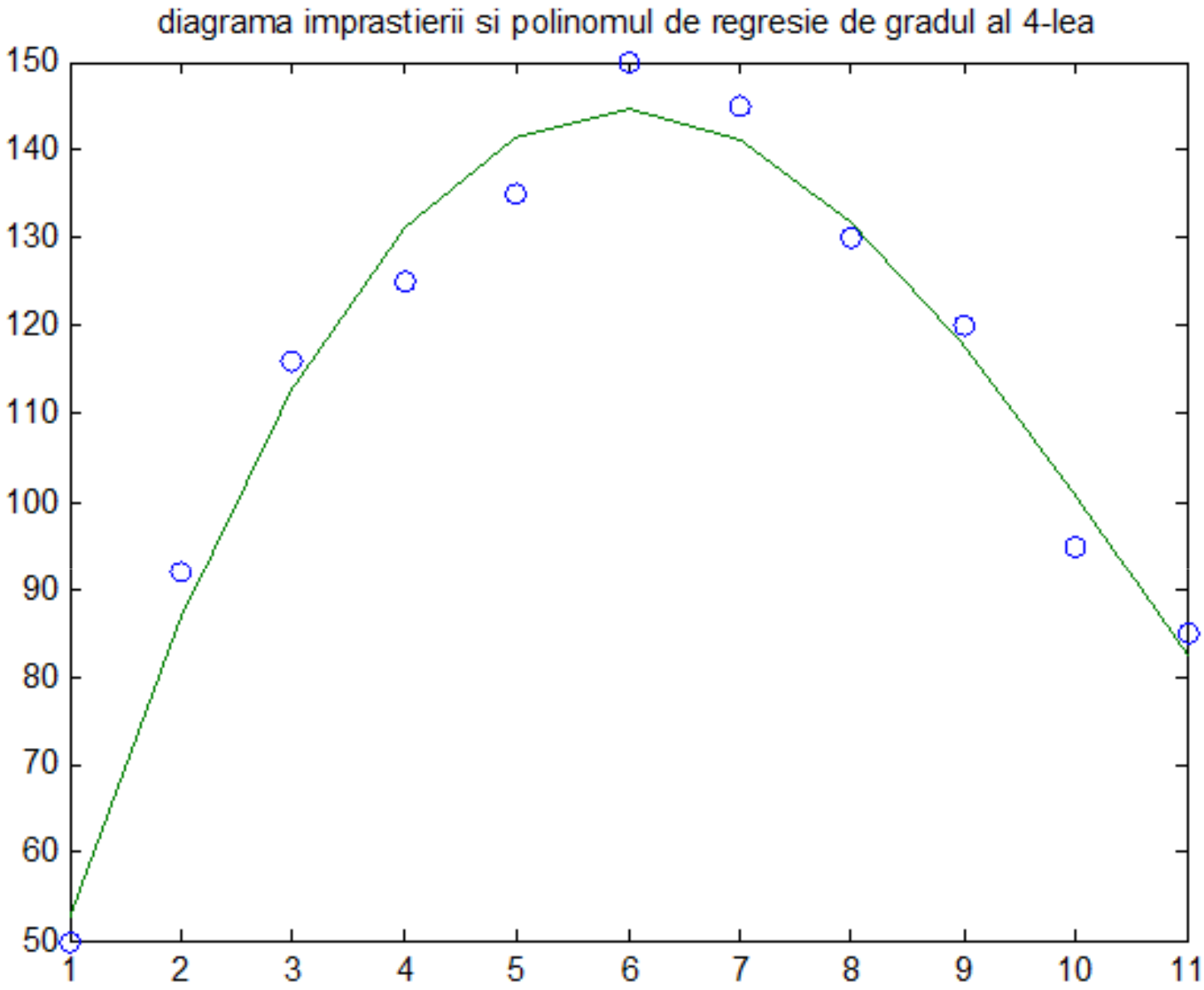
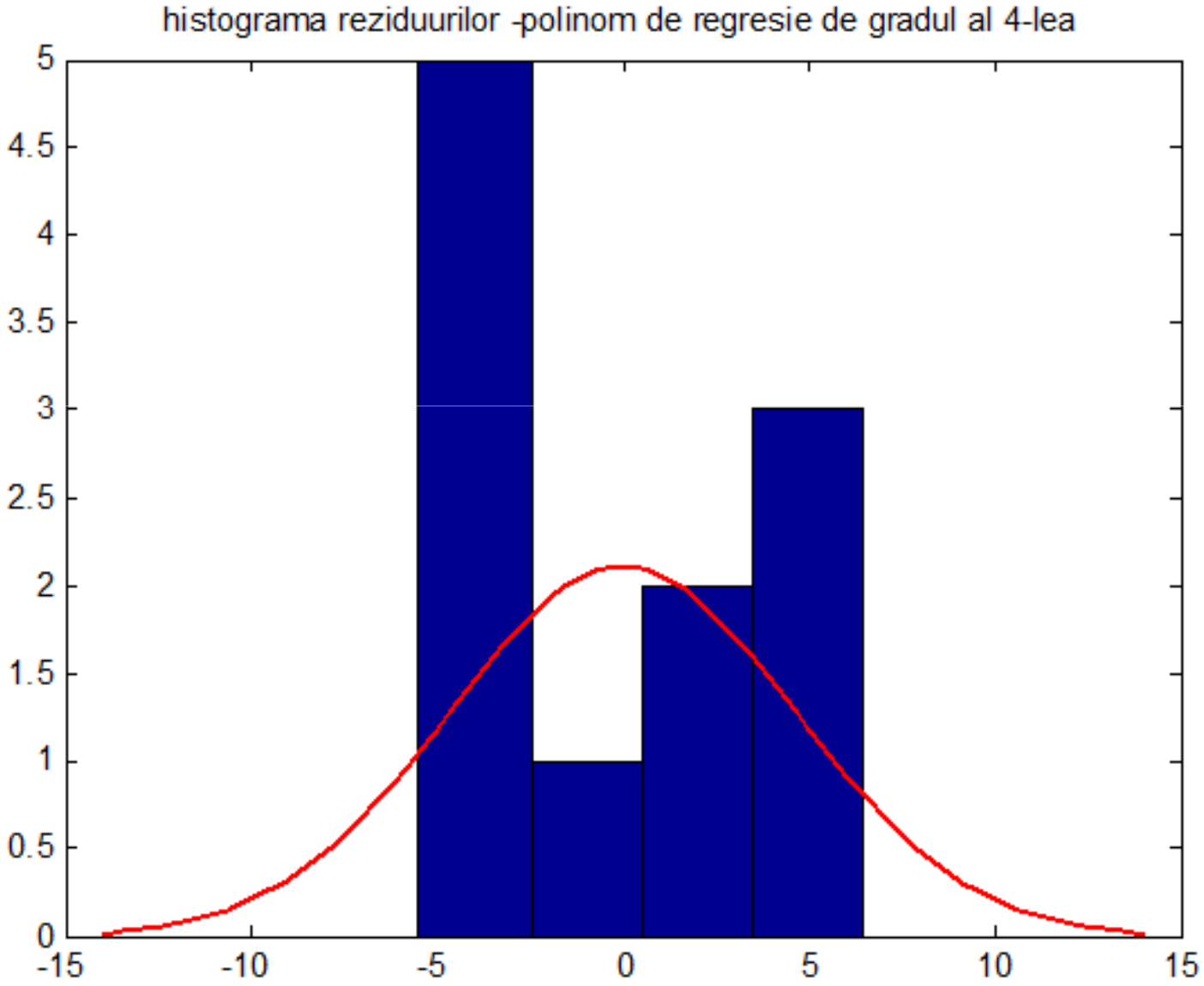


diagrama imprastierii si polinomul de regresie de gradul al 4-lea
150
140
130
120
110
100
90
80
70
60
50
1
2
3
4
5
6
7
8
9
10
11

```
>> Rez=P4-P;mean(Rez)
ans =
    2.0670e-014
>> histfit(Rez)
```



Polinoamele de regresie pot fi utilizate pentru a prognoza numărul de pacienţi din următoarele săptămâni.
Pentru a decide care polinom merită utilizat, vom calcula mediile pătratelor reziduurilor (mediile pătratelor erorii) pentru a stabili care curbă de regresie dă o aproximare mai bună.

```
>> MSE2=mean((P-P2).^2)
MSE2 =
   30.2788

>> MSE4= mean(( P-P4).^2)
MSE4 =
   19.7839
```

Vom prognoza totuşi numărul pacienţilor ce urmează a fi internaţi în săptămâna a 12-a folosind ambele polinoame:

```
>>  Yp2=b2(1)+b2(2)*12+b2(3)*12^2
Yp2 =
   44.9758

>>  Yp4=b4(1)+b4(2)*12+b4(3)*12^2+b4(4)*12^3+b4(5)*12^4
Yp4 =
   64.6364
```

În general, pentru a obține un rezultat bun este necesar un număr mare de date, să zicem n.

Cu ajutorul a $\frac{3}{4}n$ din aceste date, aleator alese scriem ecuația polinomului de regresie de gradul al 2-lea respectiv al 4-lea. Cu ajutorul acestor ecuații obținem prognoza pentru celelalte $\frac{n}{4}$ date care nu au fost utilizate și calculăm media pătratelor reziduurilor obținute pentru cele două situații considerate.

Reluăm procedeul încă de 4 ori, alegând de fiecare dată aleator alte $\frac{3}{4}n$ date. Vom alege aceea curbă polinomială de regresie pentru care media pătratelor reziduurilor obținute este cea mai mică.

Polinoame de regresie in Matlab

Prezentăm un studiu de caz în Matlab, privind aproximarea unui set de date $\{(x_i, y_i), 1 \le i \le n\}$ printr-o funcţie polinomială. Funcţiile folosite sunt:

- polyfit - funcţie ce returnează coeficienţii polinomului în ordinea $b_n, b_{n-1}, \ldots, b_1, b_0$.
- polyval - funcţie ce returnează valorile polinomului în punctele $\{x_i, 1 \le i \le n\}$.

Reluăm exemplul nr 4 calculând pentru început coeficienţii polinoamelor de regresie de gradul al 2-lea, respectiv al 4-lea, şi apoi reziduurile în aceste cazuri:

```
>> t=1:11; P=[50 92 116 125 135 150 145 130 120 95 85];
>> b2=polyfit(t,P,2)
b2 =
  -3.0758   38.9000   21.0848

>> b4=polyfit(t,P,4)
b4 =
   0.0108   -0.1249   -3.4257   44.8918   11.5152

>> pol2 = polyval(b2,t,P);
>> pol4 = polyval(b4,t,P);
   8.1818    2.5874

>> MSE2=mean((P-pol2).^2);MSE4=mean((P-pol4).^2);[MSE2 MSE4]
ans =
  30.2788   19.7839
```

Regresie mixt exponenţială

Considerând un set de date $\{(x_i, y_i), 1 \le i \le n\}$, vom construi funcţia de regresie *mixt exponenţială*:

$$y = b_0 + b_1 \cdot e^{-x} + b_2 \cdot x \cdot e^{-x}.$$

Presupunem că datele existente satisfac relaţiile:

$$Y_i = b_0 + b_1 \cdot e^{-X_i} + b_2 \cdot X_i e^{-X_i} + \varepsilon_i, \quad 1 \le i \le p$$

unde:

- $X_i, Y_i, \quad 1 \le i \le p$ sunt numere reale cunoscute,
- ε_i sunt variabile aleatoare necunoscute.

Sub formă matriceală modelul devine:

$$Y = X \cdot b + \varepsilon ,$$

unde:

- $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix}$ este un vector aleator p-dimensional,
- $X = \begin{pmatrix} 1 & e^{-X_1} & X_1 e^{-X_1} \\ 1 & e^{-X_2} & X_2 e^{-X_2} \\ \vdots & \vdots & \vdots \\ 1 & e^{-X_p} & X_p e^{-X_p} \end{pmatrix}$,
- b este vectorul 3-dimensional ale cărui componente sunt parametrii necunoscuți ai modelului, anume b_0, b_1, b_2 ,
- ε este vectorul p-dimensional al erorilor.

Metoda celor mai mici pătrate constă în aflarea vectorului b pentru care expresia:

$$S(b) = \|Y - Xb\|^2 = \sqrt{\sum_{i=1}^{p} (Y_i - b_0 - b_1 \cdot e^{-X_i} - b_2 \cdot X_i e^{-X_i})^2} \ ,$$

este minimă, presupunând că matricea X are rangul n.

Astfel $X' \cdot X$ este inversabilă şi avem:

$$b = (X' \cdot X)^{-1} \cdot X' \cdot Y \ .$$

Exemplu

Reluăm exemplul anterior construind cu o funcţie de regresie mixt exponenţială de forma:

$$f(X) = b_0 + b_1 \cdot e^{-X} + b_2 \cdot Xe^{-X} .$$

Pentru a determina parametrii b_0, b_1, b_2 vom utiliza metoda celor mai mici pătrate:

```
>> t=1:11; P=[50 92 116 125 135 150 145 130 120 95 85];
>> i=1;B=1;for i=2:11 B=[1 B];end
>> for i=1:11 t1(i)=exp(-t(i));end
>> for i=1:11 t2(i)=t(i)*exp(-t(i));end
>> X=[B' t1' t2'];b2=X\P'
b2 =
  122.7532
 -236.5601
   33.0078
>> C=b2(1)+b2(2)*exp(-t)+b2(3)*t.*exp(-t);plot(t,P,'O',t,C)
```

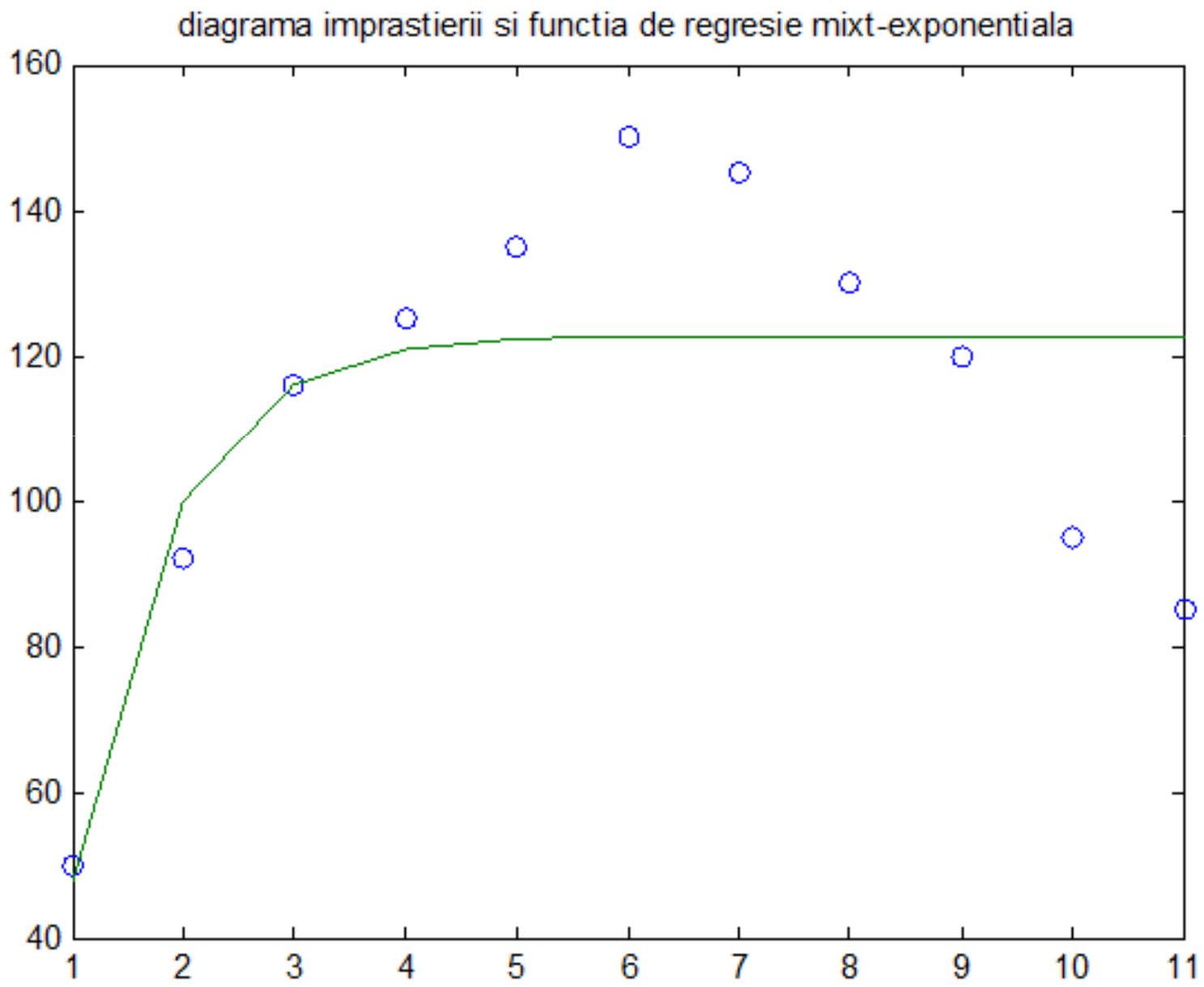
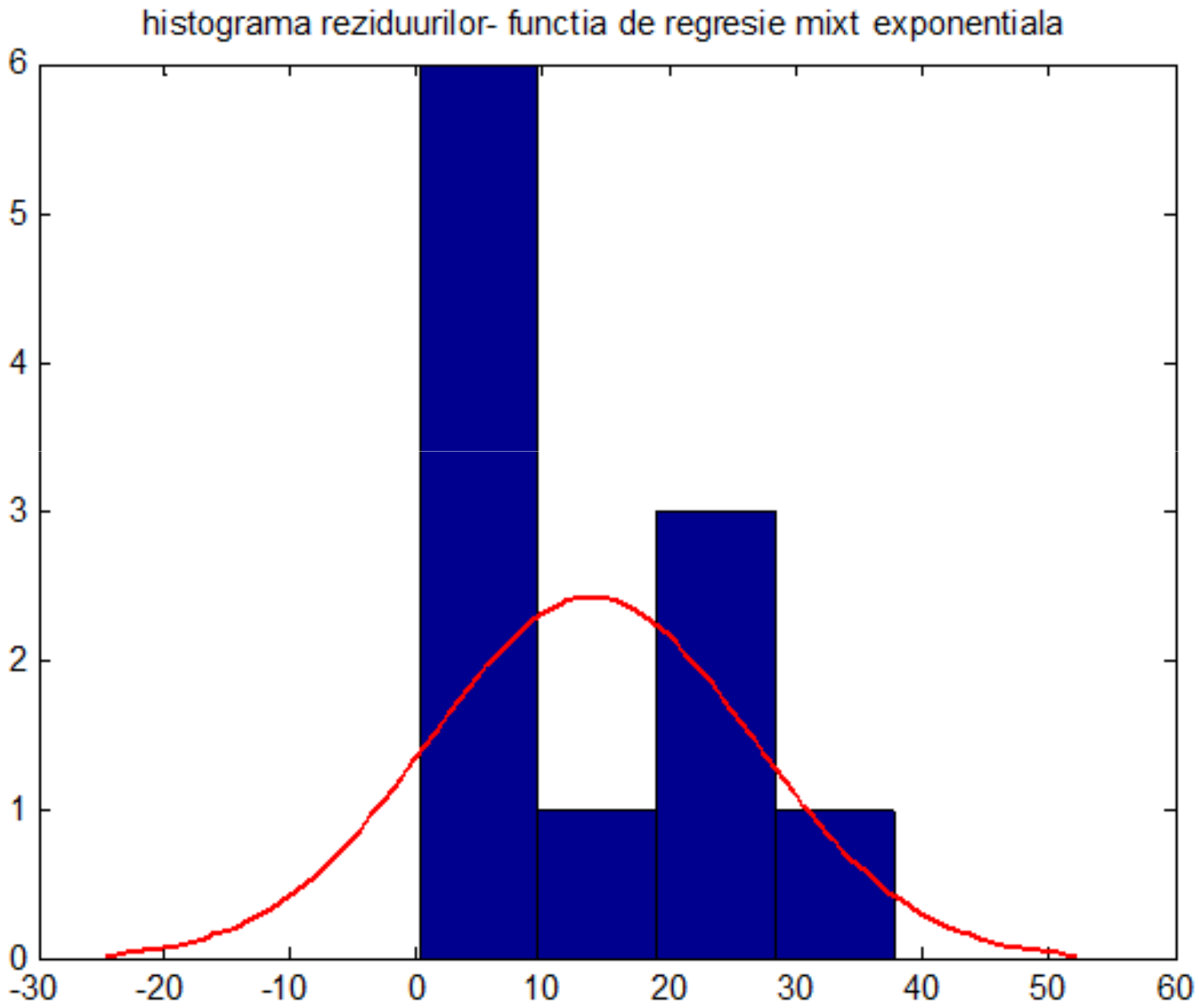


diagrama imprastierii si functia de regresie mixt-exponentiala
160
140
120
100
80
60
40
1
2
3
4
5
6
7
8
9
10
11

```
>> rez=abs(P-C);  MSE=mean((P-C).^2)
 MSE =
     340.1303
```

Comparând media pătratelor reziduurilor cu cele obţinute în cazul folosirii regresiei polinomiale, se observă că rezultatele obţinute sunt mai puţin bune, lucru evident şi din figură (diagrama împrăștierii vs. curba de regresie mixt exponenţială).

```
>>histfit(rez)
```


histograma reziduurilor- functia de regresie mixt exponentiala
6
5
4
3
2
1
0
-30
-20
-10
0
10
20
30
40
50
60

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 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 şi mai multe variabile predictive.

Să considerăm un set de n variabile aleatoare $X_1, X_2, \ldots, X_n$, considerate independente în sensul că, din punct de vedere probabilistic, reprezintă variabilele ce guvernează n factori predictivi care acţionează independent unul faţă de altul.
Variabilă dependentă (prognozată), notată cu Y, este variabilă aleatoare ale cărei valori vor fi estimate pe baza analizei regresive liniare multiple, plecând de la valorile factorilor predictivi.

5. Exemplu

Prezentăm baza de date completă ce conţine informaţii despre preţul de vânzare a 15 apartamente dintr-un cartier Ω al unui oraş Z, în trimestrul al III-lea din 2011 (parţial, această bază a fost prezentată în exemplul nr 3).

supr	30	32	35	48	49	50	55	60	65	70	75	80	82	85	90
preţ	30	29	31	36	38	40	48	47	53	49	58	60	59	71	65
nr cam	1	1	1	2	2	2	2	3	3	3	3	3	4	4	4
nr bai	1	1	1	1	1	1	1	1	2	2	2	2	1	2	2

Variabilele predictive sunt caracteristicile apartamentelor (suprafaţă, număr camere, număr băi) şi variabila dependentă este reprezentată de preţul apartamentului.

Variabilele predictive sunt caracteristicile apartamentelor (suprafață, număr camere, număr băi) și variabila dependentă este reprezentată de prețul apartamentului.

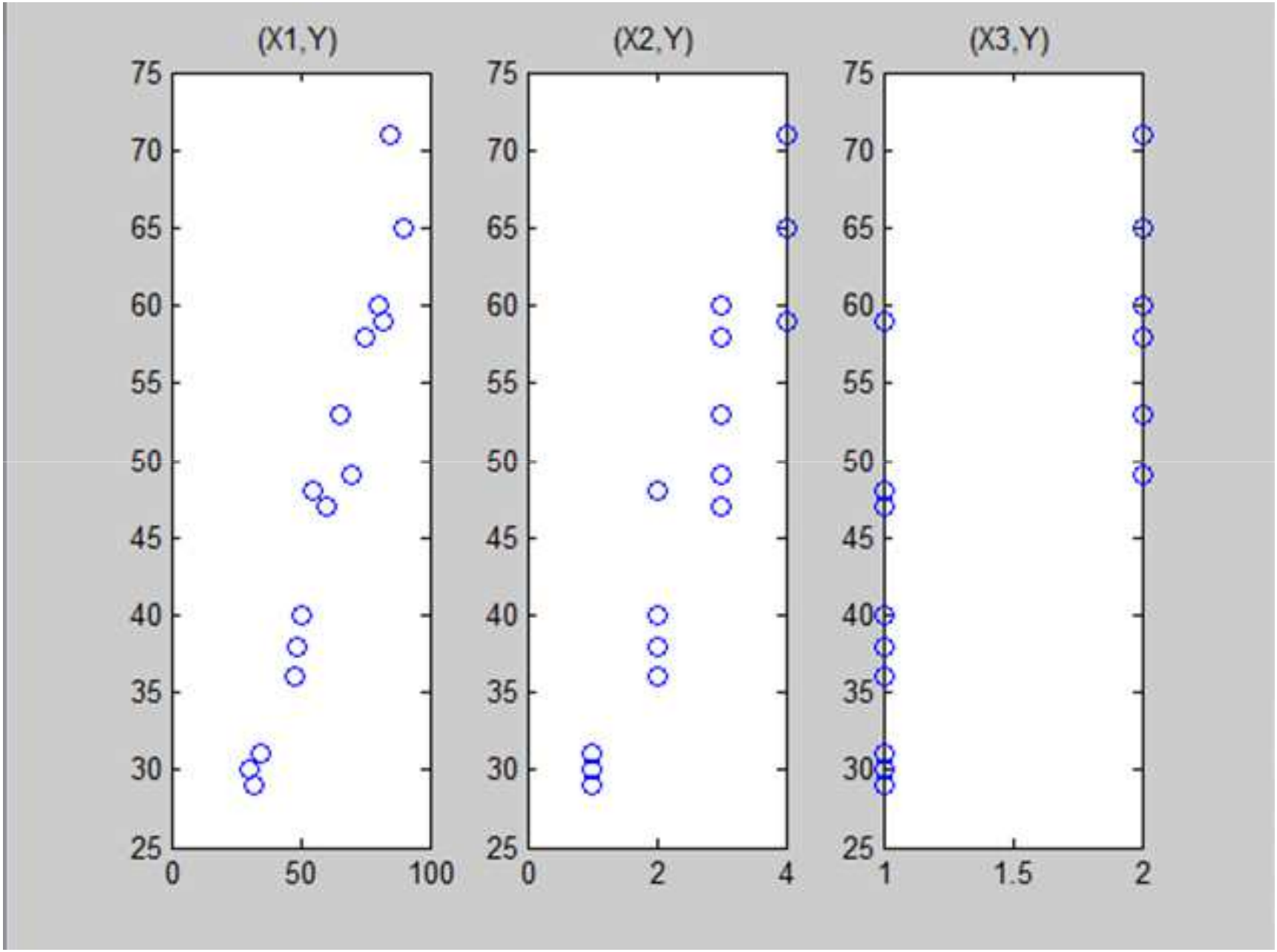
```
>> Y=[30 29 31 36 38 40 48 47 53 49 58 60 59 71 65];
>> X1=[ 30 32 35 48 49 50 55 60 65 70 75 80 82 85 90];
>> X2=[1 1 1 2 2 2 2 3 3 3 3 3 4 4 4];
>> X3=[ 1 1 1 1 1 1 1 1 2 2 2 2 1 2 2 ];
>> A=[X1' X2' X3' Y'];[r,p]=corrcoef(A)
r =
    1.0000    0.9687    0.7321    0.9729
    0.9687    1.0000    0.6378    0.9408
    0.7321    0.6378    1.0000    0.7436
    0.9729    0.9408    0.7436    1.0000
p =
    1.0000    0.0000    0.0019    0.0000
    0.0000    1.0000    0.0105    0.0000
    0.0019    0.0105    1.0000    0.0015
    0.0000    0.0000    0.0015    1.0000
```

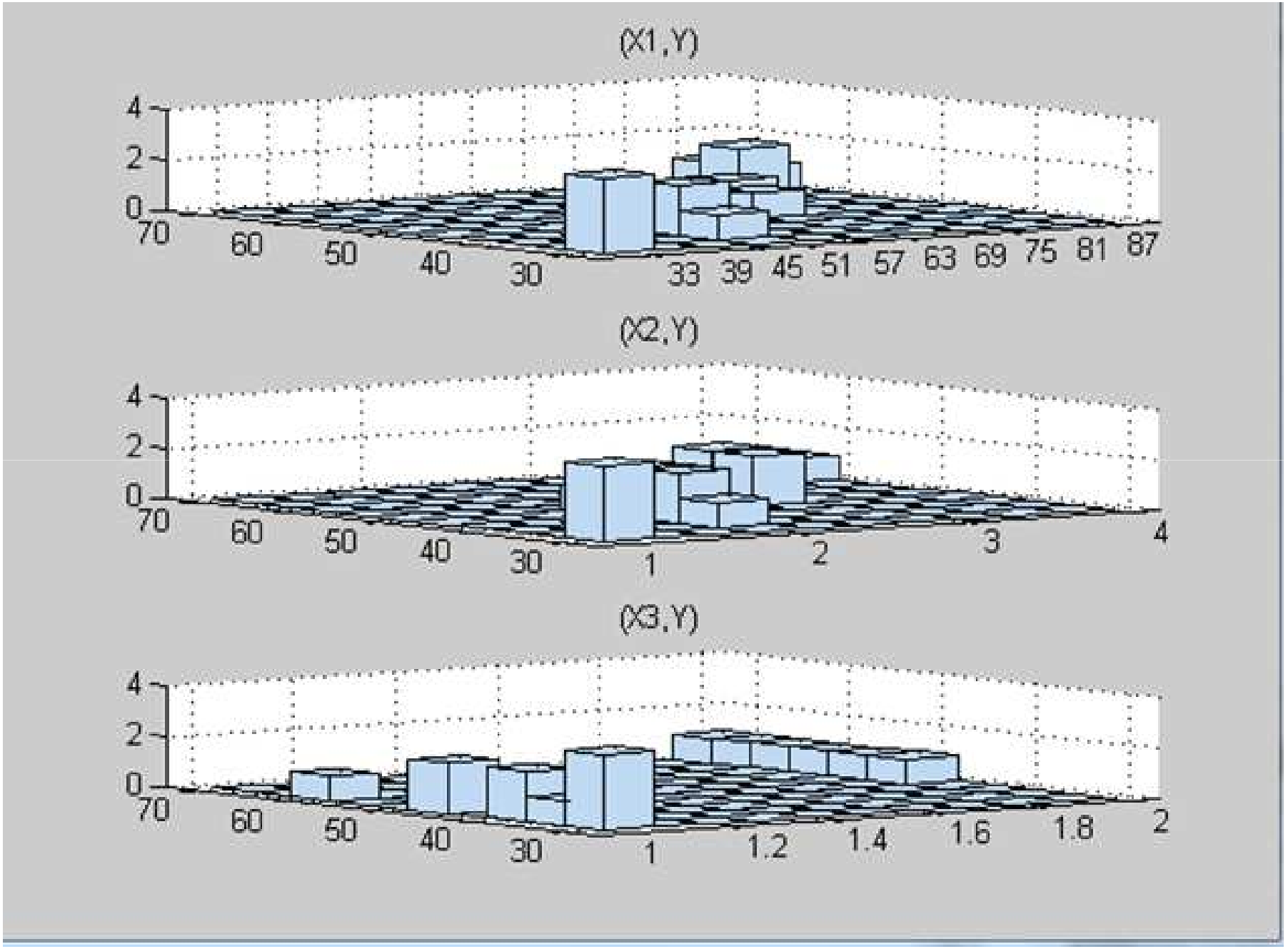
Legătura între variabila răspuns și variabilele predictive este puternică (în special X1 și Y, respectiv X2 și Y, și semnificativă.

Nefiind interesaţi de legăturile existente între variabilele predictictive, vom desena diagramele împrăştierii şi histogramele corespunzătoare doar în cazurile (variabilă predictivă, predictor):

```
>> subplot(131);plot(X1,Y,'o');
>> subplot(132);plot(X2,Y,'o');
>> subplot(133);plot(X3,Y,'o');

>> subplot(311);A1=[X1',Y']; hist3(A1);
>> subplot(312);A2=[X2',Y']; hist3(A2);
>> subplot(313);A3=[X3',Y']; hist3(A3);
```


(X1,Y)
(X2,Y)
(X3,Y)
75
70
65
60
55
50
45
40
35
30
25
0
50
100
2
4
1
1.5
2


(X1,Y)
4
2
0
70
60
50
40
30
33
39
45
51
57
63
69
75
81
87
(X2,Y)
4
2
0
70
60
50
40
30
1
2
3
4
(X3,Y)
4
2
0
70
60
50
40
30
1
1.2
1.4
1.6
1.8
2

Practic, deducând ecuaţia de regresie pe baza datelor culese de la acest lot de apartamente, se poate prognoza preţul de vânzare al unui nou apartament, introducând în ecuaţie valorile particulare ale acestuia pentru fiecare variabilă predictivă în parte.

Aşadar, pentru regresia multi-liniară, dispunem de n variabile predictive sau explicative, a căror efecte sunt independente, în principiu, unele de altele, şi de o variabilă prognozată, numită *răspuns*, pe care vrem să o deducem, cunoscând valorile celor n variabile predictive.
Legătura între variabila răspuns Y şi variabilele predictive X_i trebuie să fie una liniară.

Odată stabilită validitatea condiţiilor de aplicare a metodei, se trece la obţinerea ecuaţiei de regresie liniară multiplă, care este de forma:

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

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

Metoda regresiei multiple standard reprezintă o generalizare a regresiei simple, obţinându-se hiperplanul care trece prin „norul" multidimensional al tuturor variabilelor.

Metoda regresiei multiple standard reprezintă o generalizare a regresiei simple, obţinându-se hiperplanul care trece prin „norul” multidimensional al tuturor variabilelor.

Pentru a determina coeficienţii de regresie şi interceptorul, vom aplica *metoda celor mai mici pătrate*. Presupunem că datele existente satisfac ecuaţiile:

$$Y_i = b_0 X_0^i + b_1 \cdot X_1^i + b_2 \cdot X_2^i + ... + b_n \cdot X_n^i + \varepsilon_i, \quad 1 \le i \le p,$$

unde:

- $X_j^i, 1 \le j \le n, 1 \le i \le p$ şi $Y_i, 1 \le i \le p$ sunt numere reale cunoscute (reţinem că $p > n$).
- $X_0^i = 1, 1 \le i \le p$.
- ε_i sunt variabile aleatoare necunoscute.

Sub formă matriceală modelul devine:

$$Y = X \cdot b + \varepsilon ,$$

unde:

- $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix}$ este un vector aleator p-dimensional,
- $X = \begin{pmatrix} 1 & X_1^1 & \dots & X_n^1 \\ 1 & X_1^2 & \dots & X_n^2 \\ \dots & \dots & \dots & \dots \\ 1 & X_1^p & \dots & X_n^p \end{pmatrix}$ este matricea obţinută prin concatenarea celor p variabile X_i ,
- b este vectorul $(n+1)$-dimensional ale cărui componente sunt parametrii necunoscuţi ai modelului,
- ε este vectorul p-dimensional al erorilor.

Deoarece X este o matrice de rang n, $X' \cdot X$ este inversabilă şi astfel:

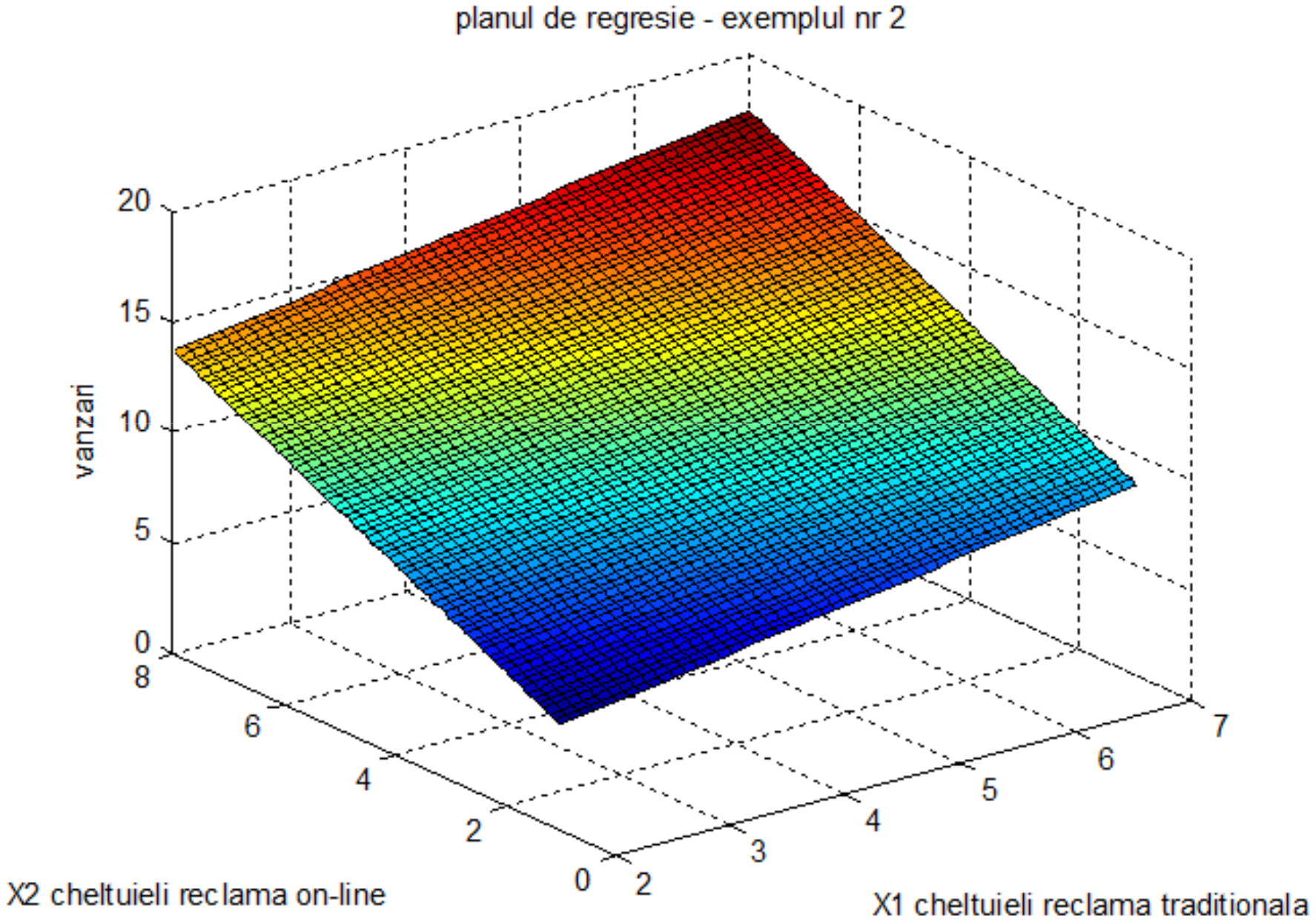
$$b = (X' \cdot X)^{-1} \cdot X' \cdot Y .$$

Exemplu

Reluăm exemplul nr 2 cu firma de IT, unde se presupune că Y, rezultatul vânzărilor, depinde de X1, suma cheltuită pentru reclama tradiţională la TV şi în ziare, şi X2, suma cheltuită pentru reclama prin Internet.
Vom scrie ecuaţia planului de regresie şi îl vom desena.

```
>> Y=[5 7 7 5 9 11 12 13 15 17 18 19];
>> X1=[2 2.5 3 1.5 4 4.5 5 5.5 6 7.5 8 7];
>> X2=[1 2 2.5 1.5 3.5 4 5 6 6.5 8 8 7.5];
>> i=1; B=[1]; for i=2:12 B=[B [1]];end
>> X=[ B' X1' X2' ];
>> b=X\Y'
b =
   1.9921
   0.7672
   1.2748
>> [X1,X2]=meshgrid(2:.1:7,1:.1:8);
>>surf(X1,X2, b(1)+b(2)*X1+b(3)*X2)
```

Reamintim că în MATLAB nu este nevoie să utilizăm formula $b = \left(X' \cdot X\right)^{-1} \cdot X' \cdot Y$, obţinând acelaşi rezultat cu b=X\Y':


planul de regresie - exemplul nr 2
vanzari
20
15
10
5
0
8
6
4
2
0
2
3
4
5
6
7
X2 cheltuieli reclama on-line
X1 cheltuieli reclama traditionala

Dacă se utilizează 85000 EUR pentru publicitate clasică şi 60.000 EUR pentru publicitate prin Internet, suntem interesaţi la cât se prognozează cifra vânzărilor.

Vom apela la Symbolic Math:

```
>> syms X1 X2
>> f=b(1)+b(2)*X1+b(3)*X2;
>> f1=subs(f,[X1,X2],[8.5,6.0])
f1 =
            16.1616
```

Exemplu

Reluăm exemplul nr 5, determinând coeficienţii hiperplanului de regresie:

```
>> Y=[30 29 31 36 38 40 48 47 53 49 58 60 59 71 65];
>> X1=[ 30 32 35 48 49 50 55 60 65 70 75 80 82 85 90];
>> X2=[1 1 1 2 2 2 2 3 3 3 3 3 4 4 4];
>> X3=[ 1 1 1 1 1 1 1 1 2 2 2 2 1 2 2 ];
>> i=1; B=[1]; for i=2:15 B=[B [1]];end
>> X=[ B' X1' X2' X3']
>> b=inv(X'*X)*X'*Y'
b =
   7.8582
   0.5769
   0.8070
   2.0385
```

Ecuaţia hiperplanului de regresie este:

$$Y = 7.8582 + 0.5769 \cdot X_1 + 0.8070 \cdot X_2 + 2.0385 \cdot X_3$$

Folosind ecuaţia hiperplanului de regresie, putem estima valoarea de vânzare a unui apartament din cartierul Ω al oraşului Z: să presupunem un apartament cu 3 camere, două băi, cu suprafaţa totală de 74 m^2 :

```
>>  syms X1 X2 X3
>> a=subs(b(1)+b(2)*X1+b(3)*X2+b(4)*X3, [X1, X2, X3], [74,3,2])
a =
   57.0453
```

În cazul regresiei liniare multiple apare problema ierarhizării predictorilor în scopul păstrării numai a acelora care, într-adevăr, au o influenţă semnificativă asupra variabilei efect, renunţând astfel la cei nesemnificativi. Astfel se iau cei care sunt puternic şi mai ales semnificativ corelaţi cu variabila răspuns.
Reducerea numărului atributelor unei baze de date este o problemă serioasă, care implică tehnici variate.

6. Exemplu

În exemplul următor vom reduce numărul atributelor, vom determina hiperplanele de regresie în cazul numărului total se atribute şi în cazul numărului redus de atribute şi vom calcula sumă pătratelor reziduurilor în cele două cazuri

Să considerăm analiza regresivă multiplă utilizată în exprimarea (predicţia) *indexului* de rezistenţă a muşchiului respirator PEmax (exprimat în cm H_2O) în funcţie de variabilele predictoare reprezentate de înălţime (H -cm), greutate (G-kg), vârstă (ani), sex, procentul masei corporale (%, BMP), volumul respirator forţat per secundă (FEV_1), volumul rezidual (RV), capacitatea funcţională reziduală (FRC) şi capacitatea totală a plămânului (TLC), pentru un lot de 25 bolnavii cu fibroză cistică . Variabila dependentă a acestui model este reprezentată de indexul de rezistenţă a muşchiului respirator (PEmax). Tabelul următor sintetizează toate aceste caracteristici.

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

Suntem interesaţi doar de ultima coloană din matricea corelaţiilor şi matricea valorilor nivelului de semnificaţie, în această coloană fiind coeficienţii de corelaţie ai predictorilor cu variabila răspuns.

r =

Column 10
0.6135
-0.2886
0.5992
0.6329
0.2295
0.3061
-0.3156
-0.4172
-0.1816
1.0000

p=

Column 10
0.0011
0.1618
0.0015
0.0007
0.2698
0.1367
0.1244
0.0380
0.3849
1.0000

Calculăm hiperplanul de regresie şi suma pătratelor reziduurilor utilizând toţi predictorii:

```
>> i=1; B=[1]; for i=2:25 B=[1 B];end
>> B2=[B' V' s' H' G' BMP' FEV' RV' FRC' TLC']; b2=B2\PE'
b2 =
  254.9404
   -3.9563
  -11.8768
   -0.6649
    3.6155
   -1.7278
    0.4066
    0.2586
   -0.5103
    0.1220
>>r2=PE-b2(1)-b2(2)*V-b2(3)*s-b2(4)*H-b2(5)*G-b2(6)*BMP-b2(7)*FEV-
b2(8)*RV-b2(9)*FRC-b2(10)*TLC;
>> MSE2=mean(r2.^2)
MSE2 =
  419.2110
```

Calculăm hiperplanul de regresie şi suma pătratelor reziduurilor utilizând predictorii puternic şi semnificativ corelaţi cu variabila răspuns:
V, H,G, FRC'

```
>> B1=[B' V' H' G' FRC']; b1=B1\PE'
b1 =
   65.7249
    1.5651
   -0.0609
    0.8295
   -0.0119
>> r1=PE-b1(1)-b1(2)*V-b1(3)*H-b1(4)*G-b1(5)*FRC;
>> MSE1=mean(r1.^2)
MSE1 =
  633.5354
```

Utilizăm ecuaţia de regresie multiplă pentru a obţine 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 exemplului nr 6, pentru un anumit pacient căruia i se cunosc valorile celor nouă parametri medicali predictivi, i se poate prognoza, cu o acurateţe suficientă, valoarea PEmax, prin introducerea în ecuaţia de regresie a valorilor sale individuale.
Spunem că, astfel, se obţine o *valoare prognostic* (*index prognostic*), pe baza datelor cunoscute.

În analiza regresivă se va evita un model cu multe variabile predictive provenind dintr-un eşantion mic.

Ca regulă generală, numărul de predictori nu trebuie să depăşească valoarea $n/10$, unde n este volumul eşantionului. În plus dacă există variabile explicative puternic corelate între ele, acestea nu vor fi incluse toate în model, ci doar un reprezentant al lor.

Pentru mai multă siguranţă, se verifică capacitatea modelului pe alt eşantion, dacă acest lucru este posibil.

Regresia multiliniara in Matlab

b = regress(y,X) returnează un vector coloană de *p* componente ce reprezintă coeficienţii hiperplanului de regresie definit de variabila răspuns *y* şi de cei *p* predictori.
X este matricea cu *n* linii şi *p* coloane a celor *p* predictori, caracterizaţi de *n* attribute, în timp ce y este un vector coloană cu *n* componente, care reprezintă cele *n* răspunsuri observate.

Dacă în X sau y există date lipsă, notate NaN, regress le ignoră.

În calcule, X trebuie să conţină o coloană ale cărei elemente sunt 1.

7.Exemplu

Considerăm baza de date din Matlab, numită carsmall, ce conţine 92 de automobile ce au ca predictori greutatea şi puterea motorului şi ca răspuns kilometrajul

```
>> load carsmall
>> x1 = Weight;x2 = Horsepower;y = MPG;
>> A=[x1 x2 y];
>> X = [ones(size(x1)) x1 x2];
>> b = regress(y,X)
b =
   47.7694
   -0.0066
   -0.0420
```

Regresia logistică

În tehnicile regresive prezentate, variabila răspuns care urma a fi dedusă din variabilele explicative, este o variabilă numerică.
În anumite cazuri această variabilă este categorială, având valorile: DA sau NU, caz în care este vorba de un tip special de clasificare.

De exemplu,

- în medicină, această variabilă poate reprezenta diagnosticarea unei boli (DA sau NU) a unui individ, pe baza unor analize clinice şi biochimice ale acestuia. Pentru fiecare combinaţie a diferitelor valori ale unor variabile care ar putea influenţa riscul de îmbolnăvire, trebuie estimată probabilitatea apariţiei maladiei.
- altă situaţie se referă la încadrarea în categoria „fraudă" a unor tranzacţii comerciale, pe baza anumitor caracteristici ale lor.

Principiul de bază rămâne acelaşi ca şi la regresia multiplă, cu deosebirea că în acest caz estimăm o *transformare* a variabilei dependente.

Astfel, dacă variabila dependentă are ca valori binare afirmaţiile DA şi NU, codate cu valorile 1 şi 0, atunci media acesteia reprezintă proporţia indivizilor din populaţie cu caracteristica respectivă.

De exemplu presupunem că avem pentru 100 de indivizi, 73 răspunsuri DA şi 27 răspunsuri NU, adică 73 de 1 şi 27 de 0. Media acestor valori va fi:

$$(73 * 1 + 27 * 0)/100 = 73/100$$

adică 73% de DA.

Modelul regresiv logistic estimează probabilitatea ca un obiect oarecare din populaţie să aibă o anumită caracteristică, de exemplu o anumită boală, tentativă de fraudă etc.

O combinaţie oarecare a predictorilor din ecuaţia de regresie poate lua valori în afara intervalului [0, 1] şi din acest motiv se ia în considerare o transformare a intervalului obţinut, în intervalul [0,1], situaţie în care valorile variabilei răspuns pot fi considerate probabilităţi.

Metoda regresiei logistice constă în folosirea transformării *logit* scrisă ca *logit*(*p*).

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

În exemplele prezentate, p este probabilitatea ca un subiect din populaţie să aibă cancer, sau să producă o fraudă, iar $(1-p)$ să nu aibă cancer sau nu producă fraudă.

Raportul $\frac{p}{1-p}$ se numeşte *şansă*, iar transformarea:

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

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

Pe baza predictorilor $X_1, X_2, ..., X_k$, $logit(p)$ este dat de ecuaţia de regresie logistică:

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

Prin transformarea inversă, se găseşte valoarea probabilităţii p ca un anumit obiect să aibă sau nu o anumită *etichetă* de clasificare:

$$p = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + ... + b_k X_k)}} .$$

Pentru a calcula coeficienţii $b_0, b_1, ..., b_k$ utilizăm metoda celor mai mici pătrate.

Presupunem că datele existente satisfac relaţiile:

$$\text{logit}(p_i) = \ln \frac{p_i}{1-p_i} = b_0 + b_1 X_1^i + \ldots + b_k X_k^i + \varepsilon_i, 1 \le i \le m ,$$

unde:

- $X_j^i, \quad 1 \le j \le k, \quad 1 \le i \le m$ sunt numere reale cunoscute;

reţinem că $m > k$,

- $p_i, \quad 1 \le i \le m$ reprezintă probabilitatea ca obiectul indice i să aibă o anumită etichetă de clasificare,
- ε_i sunt variabile aleatoare necunoscute.

Sub formă matriceală, modelul devine: $\text{logit}(p) = X \cdot b + \varepsilon$, unde

- $\text{logit}(p) = \begin{pmatrix} \ln \frac{p_1}{1-p_1} \\ \ln \frac{p_2}{1-p_2} \\ \vdots \\ \ln \frac{p_m}{1-p_m} \end{pmatrix}$ este un vector aleator m-dimensional,

- $X = \begin{pmatrix} 1 & X_1^1 & \dots & X_k^1 \\ 1 & X_1^2 & \dots & X_k^2 \\ \dots & \dots & \dots & \dots \\ 1 & X_1^m & \dots & X_k^m \end{pmatrix}$ este matricea obţinută prin concatenarea celor m variabile X_i,

- b este vectorul $(k+1)$-dimensional ale cărui componente sunt parametrii necunoscuţi ai modelului,

- ε este vectorul m-dimensional al erorilor.

Metoda celor mai mici pătrate constă în aflarea vectorului b pentru care expresia:

$$S(b) = \left\| \text{logit}(p) - Xb \right\|^2 = \sqrt{\sum_{i=1}^{m} (\ln \frac{p_i}{1-p_i} - b_0 - b_1 X_1^i + \ldots - b_k X_k^i)^2} \ ,$$

este minimă, presupunând că matricea X are rangul k.

Astfel $X' \cdot X$ este inversabilă şi avem:

$$b = (X' \cdot X)^{-1} \cdot X' \cdot \text{logit}(p) \, .$$

8.Exemplu

Pentru stabilirea legăturilor existente între apariţia hipertensiunii şi 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. În vederea utilizării sale, să considerăm următoarea situaţie statistică, descrisă în tabelul de mai jos

fumat	obezitate	vârstă	Nr. subiecţi	Nr subiecţi hipertensivi
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 toate posibilităţile existente în felul următor:

- în primele două linii ale sale se iau în consideraţie cele două situaţii posibile de clasificare ale indivizilor (hipertensiune = 1, normal = 0), pentru situaţia în care variabilele explicative fumat, obezitate şi vârstă au valorile egale cu 0, indivizi care nu fumează, nu sunt obezi şi au vârsta sub 40 de ani.

Mai departe se consideră toate combinaţiile posibile ale predictorilor.

De reţinut: *trebuie luate în consideraţie toate cazurile posibile privind combinaţiile variabilelor explicative.*

fumat	obezitate	vârstă	Nr subiecți	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

```
>> A=[ 0 0 0;  1 0 0;  0 1 0 ;  0 0 1; 1 1 0 ; 1 0 1 ; 0 1 1 ;  1 1 1 ];
>> i=1; B=[1]; for i=2:8 B=[1 B];end
>> X=[B' A];
>> logitp =[log(5/55) log(2/15) log(1/7) log(35/152) 0 log(5/55) log(15/36)
log(8/15)];
>> b=inv(X'*X)*X'*logitp'
b =
   -2.3991
   -0.4116
    1.2073
    0.2471
```

Ecuaţia de regresie logistică corespunzătoare este dată de:

$$\text{logit}(p) = - \quad 2.3991 - 0.4116 \cdot \text{fumat} + 1.2073 \cdot \text{obezitate} + 0.2471 \cdot \text{varsta} .$$

Se observă utilitatea acestei ecuaţii atunci când trebuie rezolvată problema clasificării unui individ oarecare ca fiind predispus sau nu la hipertensiune, pe baza atributelor (caracteristicilor): fumat, obezitate şi vârstă.

Astfel, se consideră valorile concrete ale acelui individ privind cei trei factori de risc enumeraţi mai sus, care se vor introduce în ecuaţia de regresie.

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

De exemplu:

- dacă respectivul este nefumător, nu este obez şi are sub 40 de ani, rezultă:

$\text{logit}(p) = -2.3991$,

deci, probabilitatea să aibă hipertensiune este egală cu 8.32%.

- dacă pacientul este nefumător, obez şi are peste 40 de ani, rezultă

$\text{logit}(p) = -0.9447$,

şi astfel probabilitatea ca să aibă hipertensiune este egală cu 28%.

Ecuaţia de regresie logistică se mai poate folosi şi în scopul de a compara probabilităţile de predicţie a hipertensiunii pentru diferite grupuri, de exemplu pentru cei cu vârsta sub 40 ani faţă de cei peste 40 ani.
Astfel, codând ca mai sus, cu 0 subiecţii sub 40 ani şi cu 1 pe cei peste 40 ani, şi utilizând ecuaţia de regresie de mai sus, obţinem cele două variante ale sale:

$$\text{logit}(p_{\text{var sta}>40}) = - \quad 2.3991 - 0.4116 \cdot \text{fumat} + 1.2073 \cdot \text{obezitate} + 0.2471$$

şi:

$$\text{logit}(p_{\text{var sta}<40}) = - \quad 2.3991 - 0.4116 \cdot \text{fumat} + 1.2073 \cdot \text{obezitate} .$$

Se obţine astfel că:

$$\text{logit}(p_{\text{var sta}>40}) - \text{logit}(p_{\text{var sta}<40}) = 0.2471 .$$

Rezultă că raportul şanselor hipertensiunii asociat cu nivelul de *vârstă* considerat mai sus (de 40 ani) este $e^{0.2471} = 1.2803$, valoare care poate fi interpretată astfel: *riscul de a face hipertensiune peste vârsta de 40 de ani este de 1.28 de ori mai mare decât sub această vârstă.*

În exemplul nr 8 am considerat cazul în care variabilele explicative sunt, de asemenea, categoriale.

Nu există niciun impediment în a considera modelul regresiv logistic şi 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).

9. Exemplu

Fibroza ficatului este un indicator important al bolilor hepatice.
Tehnic fiecare dintre cele 6 stadii al fibrozei corespunde unui anumit scor MetavirF, începând cu MetavirF = 0 nu există fibroză şi încheind cu MetavirF = 4 ciroza.
În ultimii ani se caută o manieră de evaluare neinvazivă a fibrozei, evitând biopsia prin combinarea testelor biochimice cu metode de imagistică medicală.

Fibroscanul este o tehnică de ultimă generaţie, ce permite cuantificarea fibrozei hepatice pe baza analizei deplasarii unei unde elastice de soc care se se propaga in ţesutul hepatic.

In contractul nr 2076/2007 s-a realizat o baza de date formată din 722 de pacienţi de la Clinica medicală nr 3, UMF Cluj-Napoca, pacienţi cu diferite stadii ale fibrozei ficatului. Fiecare pacient este caracterizat de 26 de parametrii clinici şi biochimici, de valoarea rigidităţii ficatului (stiffness_dată de fibroscan) şi de stadiul fibrozei, obţinut prin biopsie.
Această bază de darte va constitui *mulţimea de antrenament.*

Prezentăm 5 pacienţi, în diferite stadii ale fibrozei, cu parametrii clinici şi biochimici şi valorile rigidităţii ficatului.

	P_1	P_2	P_3	P_4	P_5
Metavir F	**0**	**1**	**2**	**3**	**4**
Stiffness	4.9	5,3	6	10.6	27
Sex	2	2	1	1	1
Age	45	56	31	37	46
BMI (body mass index)	23.1	24.3	34	25.8	30.1
Glycemia	108	110	96	99	84
Triglycerides	349	54	154	134	93
Cholesterol	220	133	197	162	152
Aspartate aminotransferase	27	62	46	48	105
Alanin aminotransferase	42	61	117	60	167
Gamma glutamyl transpeptidase	117	19	41	46	187
Total bilirubin	0.4	1.2	0.9	0.5	1
Alkaline phosphatase	156	283	248	184	246
Prothrombin index	92.2	102	107	79.1	42.7
Tqs (tocopheryl quinones)	15.3	15.2	15	18.4	25.4
INR(prothrombin time ratio)	1.06	0.98	0.95	1.19	1.85
Prolonged activated partial thromboplastin time	27.9	30.9	28.4	29	30.9
Haematids (erytthrocytes)	4.64	4.91	4.81	4.1	5.11
Hemoglobin	13.5	14.5	14.9	14.1	15.2
Hematocrit	40.9	40.9	42.6	40.5	45.2
Medium eritrocity volume	88.2	83.2	88.6	89.9	88.5
Average eritrocitary hemoglobin	29.2	29.5	31	33.4	29.7
Av. concentr. of hemoglobinin a red blood cell	33.1	35.6	35	37.2	33.5
Leukocytes	6.3	83.2	88	5.98	88.6
Thrombocytes	236	29.5	31	130	29.7
Sideraemia	70	35.5	35	132	33.5

Am aplicat un model de regresie logistică, în care am găsit probabilitatea ca un pacient să fie

- în stradiul MetavirF = 0 vs MetavirF $\in \{1,2,3,4\}$
- în stradiul MetavirF $\in \{0,1\}$vs MetavirF $\in \{2,3,4\}$
- în stradiul MetavirF $\in \{0,1,2\}$vs MetavirF $\in \{3,4\}$
- în stradiul MetavirF $\in \{0,1,2,3\}$vs MetavirF = 4

studiu care se poate face şi cu ajutorul curbelor ROC, despre care vom vorbi ulterior.

Pentru diagnosticul automat rezultatul nu este deosebit, fiind interesaţi de clasificarea pacienţilor într-un anume stadiu, lucru ce s-a realizat cu alte metode de clasificare.

Ceea ce ne interesează este că putem face o reducere a atributelor, lucru important atât pentru rularea programului, cât şi pentru serioase economii în domeniul meduical.

Acele atribute care nu au o legătură puternică şi semnificativă cu stadiul fibrozei vor fi eliminate.

Vom stabili probabilitatea P_1 ca un pacient să fie în în stradiul

MetavirF = 0 vs MetavirF $\in \{1,2,3,4\}$.

$$\text{logit}(P_1) = -29.27 - 1.21*\text{stiffness} - 24.1*\text{trigliceride} - 0.24*\text{hematocr} + 0.47*\text{vem} -- 1.65*\text{hem} + 1.17*\text{chem} + 362.28*\text{tg-uln}$$

Nivelul de semnificaţie al parametrilor predictivi ce apar in această ecuaţie:

	p-level
stiffness	0.0005
trigliceride	0.0174
hematocrit	0.0424
vem	0.0014
hem	0.0001
chem	0.0001
tg_uln	0.0175

Vom stabili probabilitatea P_2, ca un pacient să fie în stradiul MetavirF $\in \{0,1\}$ vs MetavirF $\in \{2,3,4\}$

$\text{logit}(P_2) = -7.70 - 0.42*\text{stiffness} - 0.32*\text{hem} + 0.34*\text{chem} + 0.004*\text{trombocite}$

Nivelul de semnificație al parametrilor ce apar in ecuație:

	p-level
stiffness	0.0000
hem	0.0022
chem	0.0001
trombocite	0.0445

Stabilim probabilitatea P_3, ca un pacient să fie în stradiul MetavirF $\in \{0,1,2\}$vs MetavirF $\in \{3,4\}$

$$\text{logit}(P_3) = 4.48 - 0.35*\text{stiffness} + 0.60*\text{trigliceride} - 2.87*\text{alat} + 0.010*\text{sideremia} + 120.99*\text{alat\_uln} - 132.84*\text{cst\_uln}$$

Nivelul de semnificație al parametrilor predictivi ce apar in ecuație

	p-level
stiffness	0.0000
trigliceride	0.0165
alat	0.0206
sideremie	0.0155
alat_uln	0.0203
cst_uln	0.0162

Stabilim probabilitatea P_4 ca un pacient să fie în stradiul

MetavirF $\in \{0,1,2,3\}$ vs MetavirF = 4

$$\text{logit}(P_4) = 3.03 - 0.23*\text{stiffness} + 0.047*\text{ind.d} + 0.42*\text{hemoglobin} - 0.39*\text{hem} + \\ + 0.009*\text{trombocite} - 0.472*\text{sex}$$

Nivelul de semnificaţie al parametrilor predictivi ce apar in ecuaţie:

	p-level
stiffness	0.0000
ind.d	0.0095
hemoglobin	0.0469
hem	0.0102
trombocite	0.0052
sex	0.0323

În concluzie un model de regresie logistică permite prognozarea unui anumit răspuns în funcţie de o serie de factori predictivi, realizând în acest mod o clasificare a unor obiecte în două clase distincte în raport cu acest răspuns.

Pe de altă parte, prin utilizarea unui model de regresie logistică putem elimina o serie de atribute cu nivel de semnificaţie scăzut, în problema noastră, deci regresia logistică este şi o tehnică de *Features selection*.

Analiza supravieţuirii (Survival Analysis)

Una dintre ramurile importante ale Statisticii, cu aplicaţii deosebit de interesante mai ales în domeniul medical şi al sistemelor mecanice, este *analiza supravieţuirii* (*Survival Analysis*).

În medicină este cunoscută chiar sub acest nume, în timp ce în ştiinţele inginereşti se numeşte *teoria siguranţei în funcţionare* (*Reliability Theory*), iar în economie este cunoscută ca *analiza duratei* (*Duration Analysis*).

Indiferent de context, elementul de bază al acestei teorii este termenul de „moarte", „eşec", „cădere", „absenţă", „ieşire din funcţiune" etc., care este privit ca un *eveniment* în analiza supravieţuirii.

Pentru a înţelege mai bine despre ce este vorba vom considera cazul medical, de la care de fapt s-a şi pornit în dezvoltarea acestei ramuri statistice.

Astfel, problema analizei timpilor de supravieţuire, aşa cum o indică şi numele, se referă, în principiu, la supravieţuirea unui pacient în urma unei operaţii serioase, tratamentului unei anumite boli cu sfârşit letal, e.g. cancer, SIDA etc. Practic, se înregistrează timpul din momentul începutului procesului medical (operaţie, tratament etc.) până în momentul decesului, timpul respectiv fiind numit *timp de supravieţuire*, iar analiza sa făcând obiectul analizei supravietuirii.

Analiza supraviețuirii utilizează așa-numitele *date cenzurate*, adică unele observații sunt incomplete.

Spre exemplu, să ne imaginăm că un grup de pacienți cu cancer sunt urmăriți în cadrul unui experiment o anumită perioadă de timp (*follow-up*). După trecerea acestei perioade, pacienții care au supraviețuit nu mai sunt supravegheați și, atunci când se analizează timpul de supraviețuire, nu se mai știe cu exactitate dacă mai sunt încă în viață. Pe de altă parte, unii pacienți pot părăsi grupul în timpul perioadei de supraveghere, fără a se mai cunoaște situația lor ulterioară.

Datele privind aceste tipuri de pacienți sunt *date cenzurate*.

Să menţionăm faptul că, în multe cercetări clinice, prin timp de supravieţuire se înţelege şi timpul până la apariţia unui simptom, până la revenirea bolii, până la remisiune etc.

În acest context, prin *probabilitate de supravieţuire* se înţelege proporţia indivizilor unui grup, supus unui anumit experiment medical comun, care ar putea supravieţui o anumită perioadă de timp (e.g. operaţie de transplant de inimă, chimioterapie etc.), în anumite circumstanţe date.

Tehnica clasică de calculare a timpului de supraviețuire este următoarea:

Se notează variabila aleatoare care reprezintă timpul de supraviețuire cu X.

Probabilitatea de supraviețuire se calculează prin împărțirea timpului în mici intervale $(0, t_1),\ldots, (t_{k-1}, t_k),\ldots,$ şi estimarea apoi a probabilității:

$$P\{X \leq t_n\} = P\{X \leq t_1\}P\{X \leq t_2 | X = t_1\}\ldots P\{X \leq t_n | X = t_{n\text{-}1}\}.$$

De exemplu probabilitatea ca un pacient să supraviețuiască două zile în urma unui transplant de ficat este produsul dintre probabilitatea ca pacientul să supraviețuiască o zi şi probabilitatea ca paciientul să supraviețuiască a doua zi, condiționată de faptul că a supraviețuit prima zi.

O problemă comună atât domeniului cercetării medicale - timpul de supraviețuire, cât și cercetării inginerești - timpul de funcționare sigură a unui mecanism este determinarea efectului unor variabile continue (independente) asupra timpului de supraviețuire, în particular identificarea existenței corelației între predictori și timpul de supraviețuire.

Metoda clasică de analiza supraviețuirii este bazată pe tehnici ca, de exemplu, *curba de supraviețuire Kaplan-Meier* (*Kaplan-Meier survival curve*), *analiza tabelelor de viață* (*life table analysis*), *testul logrank* (*logrank test*), *raportul hazardului* (*hazard ratio*).

Curba de supravieţiure *Kaplan-Meier*

Am stabilit anterior modul de calcul al timpului de supraviețuire

Notând cu p_{100} probabilitatea de a supraviețui a 100-a zi condiționată de faptul că a supraviețuit primele 99 de zile, atunci probabilitatea de a supravietui 100 de zile dupa transplant este:

$$p_1 \cdot p_2 \cdot \ldots \cdot p_{99} \cdot p_{100}$$

Probabilitatea de a supraviețui a 100-a zi este estimata a fi raportul dintre cei ce au supraviețuit a 100-a şi cei ce supraviețuiseră primele 99 de zile.

Probabilitatea p este 1 în zilele când nimeni nu moare, ceea ce simplifică mult calculele, fiind necesar calcularea probabilităţilor în zilele în care moare cel puțin o persoană.

10.Exemplu

Vom prezenta un experiment pentru prognoza apariţiei răului de mare:
21 de subiecţii au fost plasaţi iîntr-un simulator, timp de 2 ore.
Timpul de supravieţuire al unui subiect este calculat din minutul 0 până când i se făcea rău.
Unii subiecţi au renunţat la experiment înainte de a li se face rău, în timp ce alţii au rezistat doua ore fără a avea rău de mare, ambele cazuri ducând la observaţii cenzurate.

Tabelul următor, care prezintă situaţia acestui experiment, observaţiile cenzurate sunt notate cu * este un aşa numit *tabel de viaţă* (life table):

Am notat: T=timp de supravieţuire (min); p = proporţia de supravieţuire

nr	1	2	3	4	5	6	7	8	9	.....	21
T	30	50	50*	51	66*	82	92	120*	120*	.....	120*
p	0.95	0.9		0.85		0.79	0.74				

unde $p_1 = \frac{20}{21} = 0.9524$, $p_2 = \frac{19}{21} = 0.9048$, $p_4 = \frac{17}{20} = 0.8500$,

$p_6 = \frac{15}{19} = 0.7895$, $p_7 = \frac{14}{19} = 0.7368$

Proporţiile de supravieţuire se calculează doar pentru cele 5 date necenzurate
Vom avea doar 5 estimări ale probabilităţii de supravieţuire.

La minutul 30 probabilitatea de supravieţuire (fracţia de subiecţi care au supravieţuit) este $P_{30} = \frac{20}{21} = 0.9524$

La începutul intervalului (30',50') procentul de subiecţi rămaşi în simulator este 0.9524. Pentru a calcula probabilitatea de supravieţuire la minutul 50, conform formulei prezentate vom înmulţi P_{30} cu procentul de pacienţi rămaşi în simulator la sfârşitul intervalului (30',50'), adică $\frac{19}{21}$.

$$P_{50} = 0.9524 \cdot \frac{19}{21} = 0.8615 .$$

Se ţine seama de subiecţii cenzuraţi, atât numărătorul cât şi numitorul sunt micşoraţi, în ziua când apare un astfel de subiect.

Probabilitatea de supravieţuire după 51’ este $P_{51} = 0.8615 \cdot \frac{17}{20} = 0.7323$

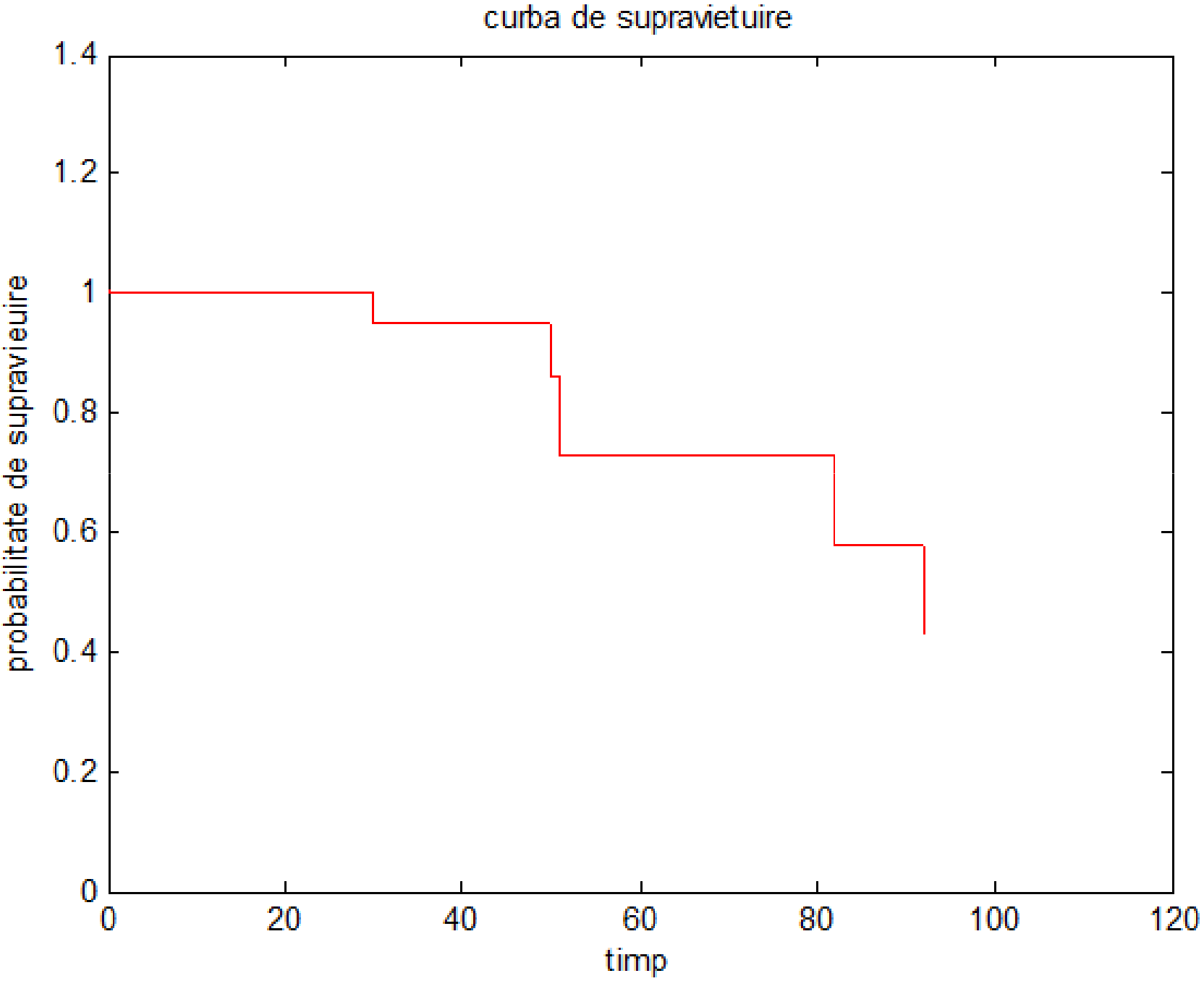
Probabilitatea de supravieţuire după 82’ este $P_{82} = 0.7323 \cdot \frac{15}{19} = 0.5781$

Probabilitatea de supravieţuire după 92’ este $P_{92} = 0.5781 \cdot \frac{14}{19} = 0.4260$

Reţinem că în primele 30 min, probabilitatea este 1, şi nu putem estima supravietuirea dincolo de ultima observaţie necenzurată din minutul 92.

Vom desena curba de supravieţuire corespunzătoare:

```
>>x=[0 0.01 30 50 51 82 92];y=[ 1.01 1 0.95 0.86 0.73 0.58 0.43];
>>x1=[ 92 120];y1=[0 0];
>> [xb, yb]=stairs(x, y); [x1b y1b]=stairs(x1,y1);
>>plot(xb,yb,'r',x1b,y1b,'k')
```


curba de supravietuire
probabilitate de supravieuire
timp
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
120

Estimator

Un estimator de parametru necunoscut θ a unui model este o funcţie ce face să corespundă unui set de observaţii $x_1, x_2, \ldots, x_n$ provenite din model, valoarea $\hat{\theta}$, numită estimare.

Estimatorul este o variabilă aleatoare deoarece valoarea lui depinde de eşantionul obţinut în mod aleatoriu.

11. Exemple

Dacă dorim să calculăm înălţimea medie a copiilor de 10 ani, se poate alege un eşantion de şcoli situate în medii diferite, şi media înălţimii copiilor de 10 ani din acest eşantion, numită şi medie empirică, este un estimator al înălţimii medii a copiilor de 10 ani.

Dacă dorim să determinăm procentajul electorilor hotărâţi să voteze pentru candidatul A vom face un sondaj pe un eşantion reprezentativ. Procentajul obţinut pe acest eşantion este un estimator al procentajului evotanţilor candidatului A.

De cele mai multe ori un estimator este o medie, o proporţie sau o dispersie.

ecdf - funcţia cumulativă de distribuţie empirică (Matlab)

ecdf - funcţia cumulativă de distribuţie empirică

Funcţia cumulativă de distribuţie empirică a unui set de valori numerice este definită pentru orice valoare reală *x* ca proporţia de observaţii mai mici sau egale cu *x*.
Este o funcţie în scară. Intuitiv, lăţimea unei trepte depinde de distanţa dintre date consecutive şi înălţimea depinde de numărul de valori
egale cu *x*.
Funcţia este monotonă. Limita sa la $-\infty$ este 0, iar limita la $+\infty$ este 1.

[f,x] = ecdf(y) calculează estimatorul Kaplan-Meier a funcţiei cumulative de distribuţie empirice (cdf = cumulative distribution function), unde

- y este un vector ce are drept componente valorile datelor
- f este vectorul ce are drept componente valorile funcţiei cumulative de distribuţie empirice în x.

12. Exemplu

Vom genera aleator timpi de supravieţuire şi timpi cenzuraţi, date ce vor avea o repartţie exponenţială.

>> y = exprnd(10,50,1); % generează un vector coloană cu 50 de componente, ce au o repartiţie exponenţială de medie $\mu = 10$, componente ce reprezintă timpii de supravieţuire

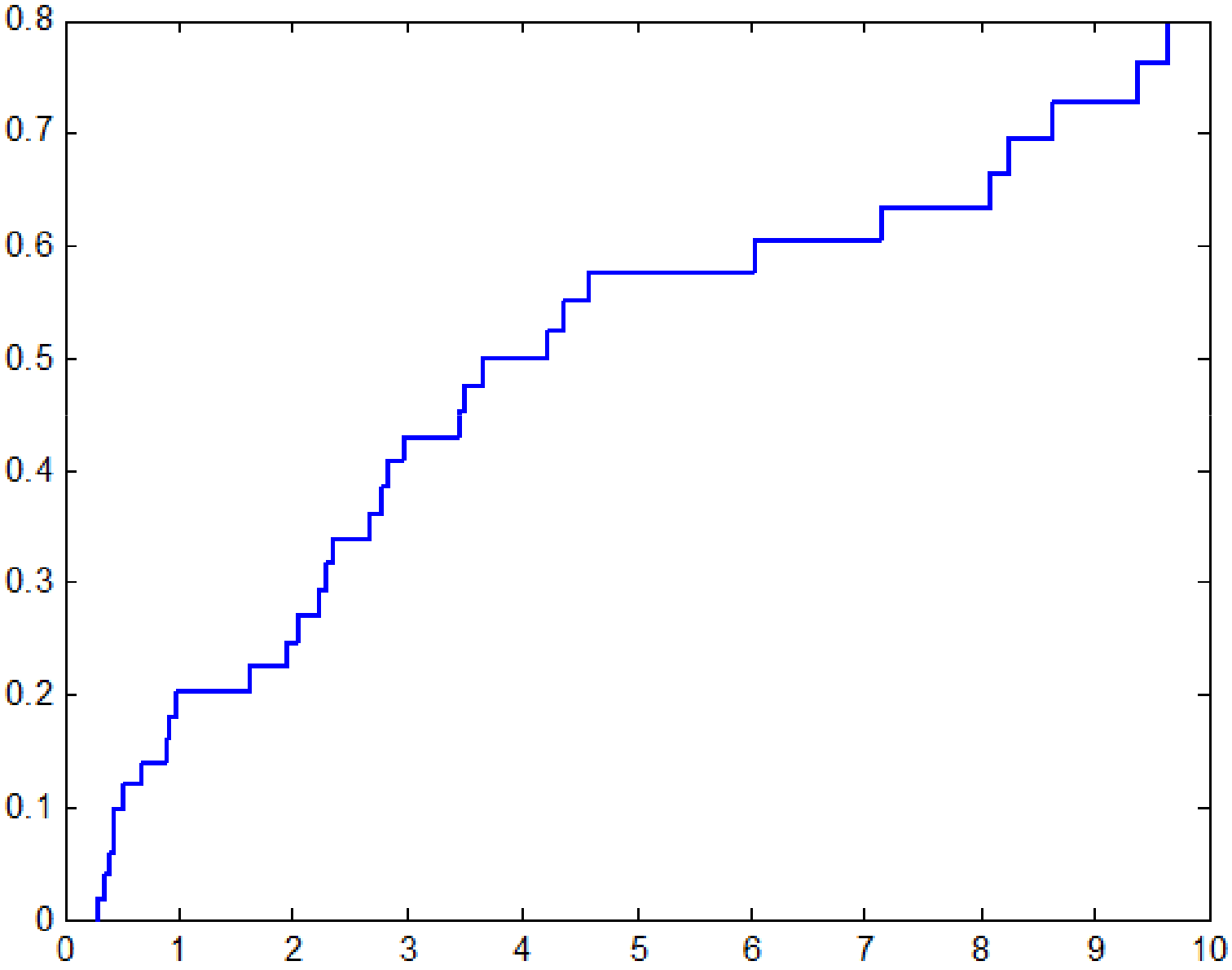
>>d = exprnd(20,50,1); % timpii de renunţare (drop-out) care au o repartiţie exponenţială de medie $\mu = 20$.

>>t = min(y,d);

>>censored = (y>d); % se obţine un vector boolean de aceasi dimensiune cu y; avem 1 pentru observaţiile cenzurate.

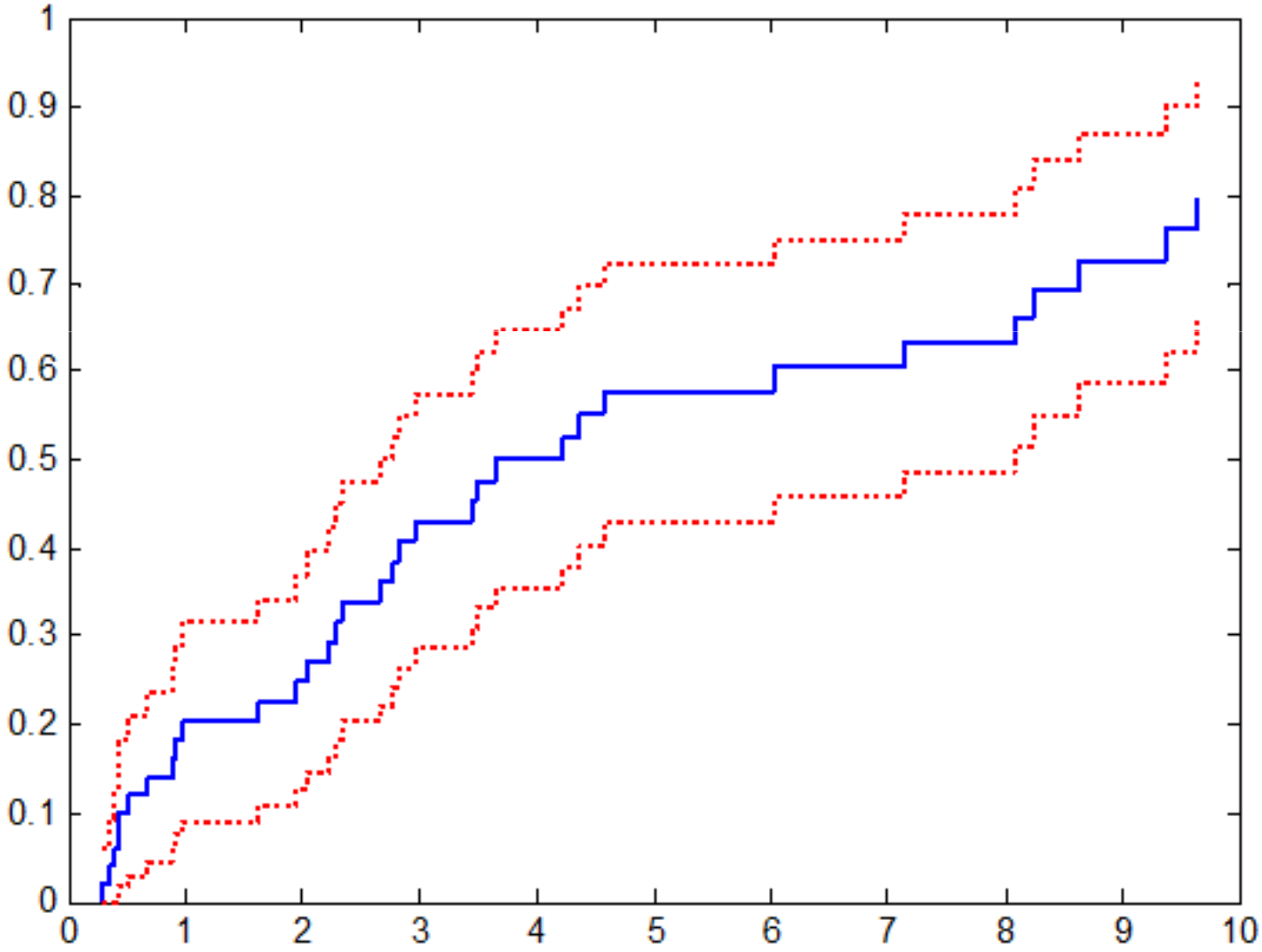
% Vom calcula şi desena funcţia cumulativă de distribuţie empirică;
vom folosi parametrul 'censoring', a cărui valoare este vectorul boolean
de aceeaşi dimensiune cu x.
Pentru datele cenzurate se atribuie valoarea 1.

```
>> [f,x,flo,fup] = ecdf(t,'censoring',censored);
>>stairs(x,f,'LineWidth',2)
>>hold on
```


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
9
10

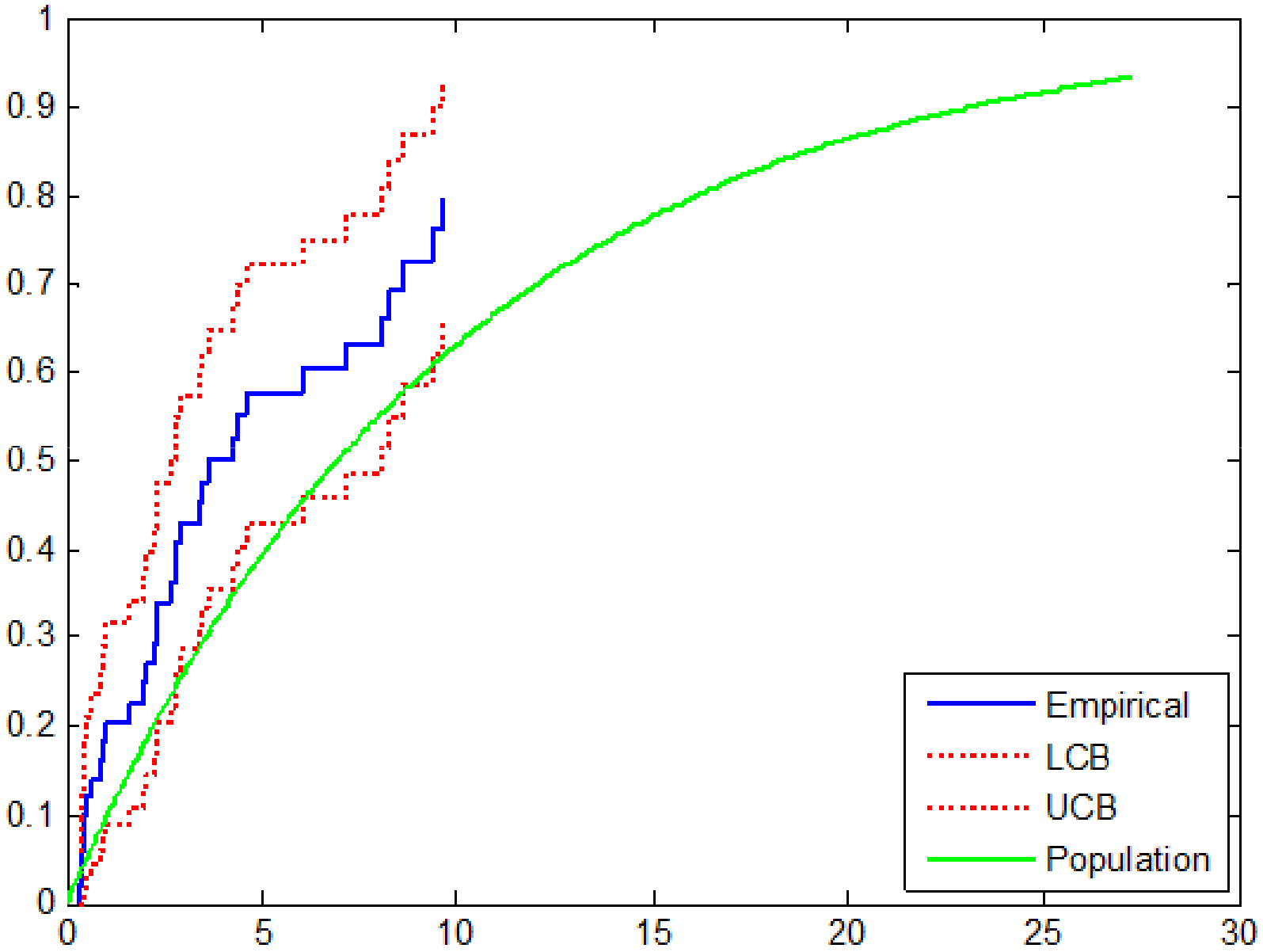
Vom calcula şi desena limitele intervalelor de încredere

```
>>stairs(x,flo,'r:','LineWidth',2)
>>stairs(x,fup,'r:','LineWidth',2)
```



Vom desena în acelaşi cadran şi funcţia de distribuţie cumulativă a populaţiei:

```
>>xx = 0:.1:max(t);
>>yy = 1-exp(-xx/10);
>>plot(xx,yy,'g-','LineWidth',2)
>>legend('Empirical','LCB','UCB','Population',...
       'Location','SE')
>>hold off
```


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
25
30
Empirical
LCB
UCB
Population

Dacă suntem interesaţi în a compara experienţa de supravieţuire la două grupuri de subiecţi, putem calcula curba Kaplan-Meier pentru fiecare grup. O asemenea abordare dă o comparaţie în anumite valori de timp, arbitrare, iar dacă cele două curbe sunt semnificativ diferite, numai în anumite valori ale timpului.

10. Exemplu

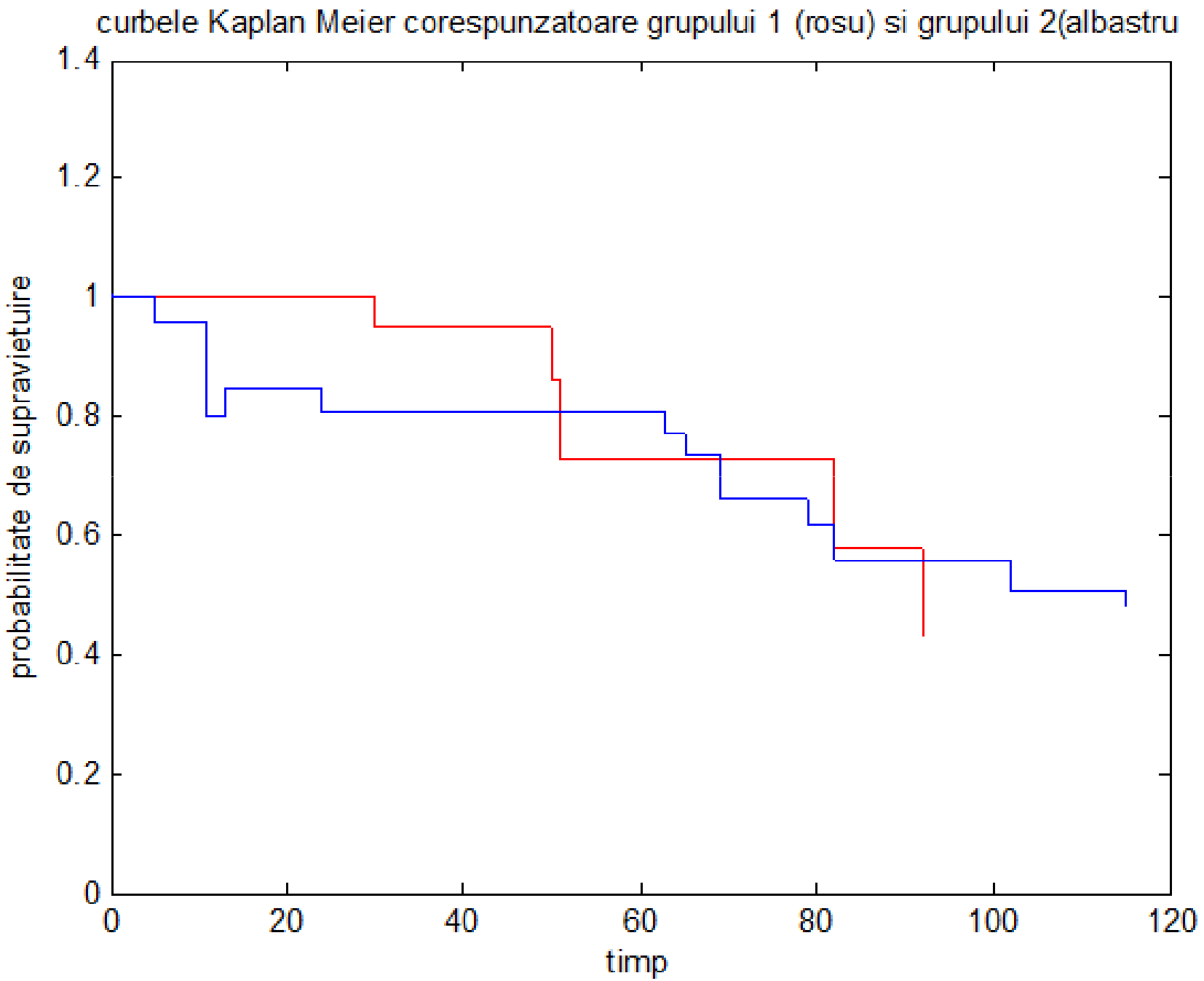
Vom considera un al doilea grup de subiecţi testaţi pentru răul de mare, în acelaşi simulator, doar că frecvenţa şi acceleraţia mişcării s-a dublat fată de primul experiment. Tabelul următor datele obţinute:

nr	1	2	3	4	5	6	7	8	9	10	11	12
T	5	6*	11	11	13	24	63	65	69	69	79	82
p	0.96			0.89	0.85	0.81	0.77	0.74		0,66	0,62	

nr	13	14	15	16	17	.....	28
T	82	102	115	120*	120*	.....	120*
p	0.56	0.51	0,48				

Vom desena curbele Kaplan-Meier în acelaşi sistem de axe:

```
>>x=[0 0.01 30 50 51 82 92];y=[ 1.01 1 0.95 0.86 0.73 0.58 0.43];
>>x1=[ 92 120];y1=[0 0];
>> [xb, yb]=stairs(x, y); [x1b y1b]=stairs(x1,y1);
>>x2=[0 0.01 5 11 13 24 63 65 69 79 82 102 115];
>> y2=[1.01 1 0.96 0.80 0.85 0.81 0.77 0.74 0.66 0.62 0.56 0.51 0.48];
>> [x2b, y2b]=stairs(x2, y2); [x3b y3b]=stairs(x3,y3);
>> plot(xb,yb,'r',x1b,y1b,'k', x2b,y2b,'b',x3b,y3b,'k')
```


curbele Kaplan Meier corespunzatoare grupului 1 (rosu) si grupului 2(albastru
probabilitate de supravietuire
timp
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
120

Testul log-rank

Metoda cea mai cunoscută de comparare a timpurilor de supraviețuire pentru două grupuri independente de subiecți este *testul log-rank*.

Testul log-rank este o metodă non-parametrică de testarea ipotezei nule. H_0, în sensul că verificăm dacă nu există diferențe semnificative între curbele de supraviețuire.

Curbele Kaplan-Meier şi testul logrank se folosesc când variabila predictor este categorială (de exemplu medicament vs placebo) sau când ia un număr mic de valori, care pot fi considerate categoriale.

Testul logrank se utilizează pentru a detecta diferenţele între curbele de supravieţuire în condiţiile în care rata moratalităţii (ieşirii din funcţiune) într-un grup este semnificativ mai mare decât rata corespunzătoare în al doilea grup şi raportul acestor rate este constant în timp.

Acest test dă ponderi egale tuturor evenimentelor (decese, ieşiri din funcţiune)

$$T = \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Principiul testului log-rank constă în divizarea scalei timpilor de supravieţuire, ignorând timpii cenzuraţi.

13. Exemplu

Considerăm două grupuri de 10 și repectiv 8 pacienți ce au suferit un transplant de ficat, cărora li se administrează medicație diferită. Prezentăm timpul de supraviețuire pentru cele două grupuri, în luni.

G1	1* 3 4* 5 5 6* 7 7 7* 8
G2	2 2 3* 4 6* 6* 7 10

Reamintim că * înseamnă dată cenzurată, în cazul nostru pacient retras în luna t_i din studiu.

În tabelele următoare vom folosi notaţiile:

- t – timpul în care are loc evenimentul (în luni);
- n - numărul de subiecţi care sunt încă sub observaţie în luna t
- n_1 - numărul de subiecţi din grupul 1 care sunt încă sub observaţie în luna t
- n_2 - numărul de subiecţi din grupul 2 care sunt încă sub observaţie în luna t
- r – numărul de decese raportat în luna t.
- c- numărul devalori cenzurate raportat în luna t.
- σ_1 – numărul de decese din grupul 1, raportat în luna t.
- σ_2 – numărul de decese din grupul 2, raportat în luna t.
- e_1 – numărul de decese din grupul 1, aşteptat în luna t.
- e_2 – numărul de decese din grupul 2, aşteptat în luna t.

Luna nr 1

t	n	n_1	n_2	r	c	σ_1	σ_2	e_1	e_2
1	18	10	8	0	1	0	0	0	0

Luna nr 2

t	n	n_1	n_2	r	c	σ_1	σ_2	e_1	e_2
1	18	10	8	0	1	0	0	0	0
2	17	9	8	2	0	0	2	1.06	0.94

Raportul de subiecţi sub observaţie în grupul 1 este $\frac{9}{17}$ iar in grupul 2 , $\frac{8}{17}$.

Astfel evenimentele aşteptate sunt $e_1 = 2 \cdot \frac{9}{17} = 1.06$ şi $e_2 = 2 \cdot \frac{8}{17} = 0.94$.

Luna 3

t	n	n_1	n_2	r	c	σ_1	σ_2	e_1	e_2
1	18	10	8	0	1	0	0	0	0
2	17	9	8	0	0	0	2	1.06	0.94
3	`15	9	6	1	1	1	0	0.6	0.4

Raportul de subiecţi sub observaţie în grupul 1 este $\frac{9}{15}$ iar in grupul 2 , $\frac{6}{15}$,

evenimentele aşteptate sunt $e_1 = \frac{9}{15} = 0.6$ şi $e_2 = \frac{6}{15} = 0.4$.

Tabelul următor prezintă situatia cu toţi timpii de supravieţuire:

t	n	n_1	n_2	r	c	σ_1	σ_2	e_1	e_2
1	18	10	8	0	1	0	0	0	0
2	17	9	8	0	0	0	2	1.06	0.94
3	15	9	6	1	1	1	0	0.6	0.4
4	13	8	5	1	1	0	1	0.62	0.38
5	11	7	4	2	0	2	0	1.27	0.73
6	9	5	4	0	3	0	0	0	0
7	6	4	2	3	1	2	1	2	1
8	2	1	1	1	0	1	0	0.5	0.5
Total						6	4	6.05	3.95

Aplicăm testul χ^2 cu un grad de libertate (având două grupuri)

$$T = 1.06 + \frac{(1.06)}{0.94} + \frac{(0.4)^2}{0.6} + 0.4 + 0.62 + \frac{(0.62)^2}{0.38} + \frac{(0.73)^2}{1.27} + 0.73 +$$

$$+ 0.5 + 0.5 = 6.7032$$

	p value											
df	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.0025**	**0.001**	**0.0005**
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.59	6.74	7.78	9.49	11.14	11.67	13.23	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.33	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.53	14.45	15.03	16.81	13.55	20.25	22.46	24.10
7	9.04	5.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87

Din tabelul testului χ^2 obținem că $P \approx 0.01 < 0.05$ și astfel ipoteza nulă nu este valabilă, ceea ce se traduce că există deosebiri între timpii de supraviețuire ale grupurilor 1 și 2.

10. Exemplu

Reluăm experimentele cu răul de mare, pentru a vedea dacă există diferenţe semnificative între timpii de supravieţuire a celor două grupuri:

In primul experiment avem 5 evenimente(rău de mare) în minutele: 30,50,51,82,92.

In al doilea experiment avem 14 evenimente(rău de mare): câte unul în minutele 5,13,24,63,65,78,102,115 şi câte două în minutele 11,69,82.

Pentru cele două evenimente combinate avem 16 valori distincte de timp, şi vom obţine 16 intervale de timp:

(0,5'], (5',11'], (11',13'], (13',24'], (24',30'], (30',50'], (50',51'], (51',63'], (63',65'], (65',78'], (78',82'], (82',92'], (92',102'], (102',115'], (115',120']

Aplicăm algoritmul de calcul prezentat în exemplul nr 12, utilizând aceleaşi notaţii:

t	n	n_1	n_2	r	c	σ_1	σ_2	e_1	e_2
t_1	42	21	21	1	0	0	1	0.5	0.5
t_2	41	21	20	2	1	0	2	1.03	0.97
t_3	38	21	17	1	0	0	1	0.55	0.45
t_4	37	21	16	1	0	0	1	0.55	0.45
t_5	36	21	16	1	0	0	1	0.58	0.42
t_6	35	21	15	1	1	1	0	0.57	0.43
t_7	33	18	15	1	0	1	0	0.56	0.44
t_8	32	17	15	1	0	0	1	0.53	0.47
t_9	31	17	14	1	0	0	1	0.55	0.45
t_{10}	30	17	13	2	1	0	2	1.86	1.14
t_{11}	27	16	11	1	0	0	1	0.59	0.41

t_{11}	27	16	11	1	0	0	1	0.59	0.41
t_{12}	26	16	10	3	0	1	2	1.86	1.14
t_{13}	23	15	8	1	0	1	0	0.65	0.35
t_{14}	22	14	8	1	0	0	1	0.64	0.36
t_{15}	21	14	7	1	0	0	1	0.67	0.33
t_{16}	20	14	6	0	20	0	0	0	0

$$T = \sum_{j=1}^{16} \sum_{i=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 20.3261$$

Din tabelul testului χ^2 obținem $P \approx 0$ și astfel respingem ipoteza nulă.

Modelul Cox

Metoda clasică de analiza supraviețuirii nu poate fi utilizată pentru a explora efectul simultan asupra supraviețuirii, a mai multor variabile.

În acest caz, nu se poate utiliza direct metoda regresiei multiple din cel puțin două motive:
(*a*) variabila ce descrie timpul de supraviețuire nu este de cele mai multe ori normal repartizată (de obicei este exponențială sau Weibull) și
(*b*) analiza supraviețuirii utilizează așa-numitele *date cenzurate*, adică unele observații sunt incomplete.

Analiza supraviețuirii utilizează mai multe metode regresive specializate, din care amintim:

- Analiza regresivă a hazardurilor proporționale –modelul Cox (*Cox proportional hazards regression analysis*), pe scurt -modelul Cox al hazardului proporțional (*Cox proportional hazard model*);
- Modelul Cox al hazardului proporțional cu covariante dependente de timp (*Cox proportional hazard model with time-dependent covariates*);
- Modelul regresiv exponențial (*Exponential regression model*);
- Modelul liniar regresiv log-normal (*Log-normal linear regression model*);
- Modelul liniar regresiv normal (*Normal linear regression model*).

Modelul Cox al hazardului proporţional

Funcţia de supravieţuire (*survival function*) este definită prin probabilitatea:

$$S(t) = P\{T > t\},$$

unde t reprezintă timpul în general, iar T este timpul până la deces.

Repartiţia duratei de viaţă (*lifetime*) este dată de

$$F(t) = 1 - S(t),$$

unde $f(t) = \frac{d}{dt} F(t)$ reprezintă rata de deces pe unitatea de timp.

Funcţia hazard (*hazard function*) este definită de formula:

$$\lambda(t) = P(t < T < t + dt) = \frac{f(t)dt}{S(t)} = -\frac{S'(t)dt}{S(t)}$$

Funcţia hazard reprezintă deci riscul de a deceda într-un interval foarte scurt de timp *dt* după un timp dat *T*, presupunând evident supravieţuirea până la acel moment.

Curba de supravieţuire reprezintă grafic supravieţuirea cumulată în funcţie de timp, în timp ce derivata curbei de supravieţuire este rata survenirii decesului într-un interval scurt de timp, iar acesta este hazardul.

De exemplu, dacă ne aşteptăm ca 20% dintre pacienţii cu un anume tip de cancer să moară în acest an, atunci hazardul este 20% pe an.

Aşadar, hazardul este un risc, iar raportul hazardelor (hazard ratio) este un risc relativ.
Dacă raportul hazardelor este 0,5, atunci riscul de a muri într-un grup (cel tratat) este jumătate din riscul de a muri în celălalt grup (placebo).

Regresia hazardului proporţional (modelul lui Cox) utilizează metodele de regresie pentru a prezice riscul relativ pe baza uneia sau mai multor variabile X.

Hazardul (riscul) se poate modifica în timp:

- el poate creşte în timp; *de exemplu* la pacienţii cu cancer, cu cât trece mai mult timp de la momentul diagnosticului, cu atât riscul de deces ca urmare a cancerului este mai mare.
- hazardul poate scădea în timp: tot la pacienţii cu cancer, la care s-a obţinut o remisie prin tratament, cu cât trece mai mult timp fără o recădere, cu atât riscul de recădere este mai mic.

Se numeşte "hazard proporţional" deoarece se bazeaza pe presupunerea că, dacă hazardul poate să varieze în timp, raportul hazardelor rămâne constant (sau altfel spus, cele două funcţii ale hazardului sunt proporţionale una cu cealaltă).

Presupunerea "hazardului proporţ ional" este rezonabilă în multe situaţii clinice dar nu întotdeauna; *de exemplu* atunci când comparăm un tratament medical cu unul chirurgical, al doilea are de obicei un hazard mai mare perioperator, pe când la cel medical hazardul creşte după un timp mai lung, motiv pentru care regresia hazardului proporţional nu este indicată.

Modelul lui Cox ne furnizează o estimare a efectului tratamentului asupra supraviețuirii după ajustarea pentru celelalte variabile independente și ne permite să estimăm hazardul (sau riscul) de deces sau alt eveniment de interes la indivizi, date fiind variabilele lor prognostice.

Chiar dacă grupurile (tratament și martor) sunt similare în privința variabilelor cunoscute ca afectând supraviețuirea, folosirea unui model al lui Cox pentru aceste variabile poate produce o estimare mai precisă a efectului tratamentului.

De exemplu, într-o analiză multivariată trebuie să includem atât de mulți pacienți, încât să avem cel puțin câte 10 efecte (în acest caz decese) pentru fiecare variabilă independentă.

Modelul Cox al hazardului proporţional este foarte general între modelele regresive deoarece nu se bazează pe nicio ipoteză prealabilă privind repartiţia supravieţuirii. El se bazează doar pe presupunerea că hazardul este o funcţie doar de variabilele independente (predictive, covarianţe) $Z_1, Z_2, ..., Z_k$, adică:

$$h(t; Z_1, Z_2, ..., Z_k) = h_0(t) \cdot \exp\left(b_1 \cdot Z_1 + b_2 \cdot Z_2 + ... + b_k \cdot Z_k\right),$$

sau, logaritmând,

$$\ln\left(\frac{h(t, Z_1, ..., Z_k)}{h_0(t)}\right) = b_1 \cdot Z_1 + ... + b_k \cdot Z_k$$

fiind deci un model semi-parametric.

$h(t,Z)$ poate fi considerată o densitatea de repartiţie Weibull deoarece:

$$h(t,Z) = \lambda \cdot p \cdot t^{p-1},$$

unde

$$\lambda = \exp(b_1 \cdot Z_1 + \ldots + b_k \cdot Z_k)$$

$$h_0(t) = p \cdot t^{p-1}$$

Termenul $h_0(t)$ se numeşte hazardul de bază (*baseline hazard*, *underlying hazard function*) reprezentând hazardul pentru un anumit individ atunci când toate variabilele independente sunt egale cu zero. $h_0(t)$ depinde doar de timp -*t*.

Exponenţiala se referă doar la $Z = (Z_1,\ldots,Z_k)$ şi variabilele Z_i sunt independente de timp.

Exemple de variabile independente de timp (în domeniul medical):

- variabile care nu se schimbă în timp, de exemplu: sexul. Valorile sunt setate la $t = 0$;
- variabile ce nu par a se schimba în timp, *de exemplu* statutul de fumător;
- variabile ce sunt considerate că nu depind de timp: vârsta, greutatea.

Modelul Cox este foarte utilizat, fiind o alegere sigură în multe situații.

Hazardurile estimate sunt nenegative. în plus chiar dacă $h_0(t)$ nu este specificat, putem estima coeficienții β_i și apoi calcula rata hazardului (rata de risc - *hazard ratio*), notată *HR*.

In analiza supraviețuirii este preferat modelul Cox celui logistic, deoarece modelul Cox lucrează și cu date cenzurate.

hazard ratio-*HR* este definită de formula:

$$HR = \frac{h(t, Z^*)}{h(t, Z)} \ ,$$

unde $Z = (Z_1^*, \ldots, Z_k^*)$ și $Z = (Z_1, \ldots, Z_k)$

Pentru simplificarea interpretării luăm $HR \geq 1$, adică $h(t, Z^*) \geq h(t, Z)$ considerând Z^* un grup cu risc (hazard) mai mare, de exemplu grupul placebo şi Z un grup cu risc mai mic, cum ar fi grupul căruia i se aplică un anume tratament.

Astfel :

$$HR = \frac{h_0(t) \cdot \exp(\sum_{i=1}^{k} \beta_i \cdot Z_i^*)}{h_0(t) \cdot \exp(\sum_{i=1}^{k} \beta_i \cdot Z_i)} = \exp(\sum_{i=1}^{k} \beta_i \cdot (Z_i^* - Z_i))$$

Să menţionăm că, totuşi, trebuie luate în consideraţie două condiţii:

(*a*) trebuie să existe o relaţie multiplicativă între $h_0(t)$ şi funcţia log-liniară a covarianţelor –*ipoteza proporţionalităţii*, prin prisma hazardului
(*b*) trebuie să existe o relaţie log-liniară între hazard şi variabilele independente.

Remarcă

1. Selecţia variabilelor explicative se face după aceeaşi regulă ca şi în cazul regresiei liniare multiple.

2. Semnul coeficienţilor de regresie b_i trebuie interpretat după cum urmează. Un semn pozitiv indică un hazard ridicat deci, în consecinţă, un prognostic negativ pentru individul cu o valoare ridicată a acelei variabile. În schimb, un semn negativ pentru un anumit coeficient indică un hazard scăzut relativ la acea variabilă.

3. La fel ca şi în cazul regresiei liniare multiple/logistice şi în acest caz putem utiliza valoarea predictivă a modelului, folosind indicele prognostic definit de $IP = b_1 \cdot Z_1 + b_2 \cdot Z_2 + ... + b_k \cdot Z_k$. Se poate calcula astfel funcţia de supravieţuire $S(t) = \exp\left[-H_0(t)\right]^{\exp(IP)}$, unde $H_0(t)$, numit şi *hazardul de bază cumulat*, este o funcţie scară în variabila timp.

1. În cazul unei singure variabile $Z \in \mathbf{R}$, cu parametrul $\beta \in \mathbf{R}$, rata hazardului este:

$$h(t, Z) = h_0(t)\exp(\beta Z).$$

Rata hazardului pentru doi subiecţi corespunzând lui Z_1 şi Z_2 este:

$$\exp(\beta(Z_1 - Z_2)).$$

Exemplu

Vârsta dependenţilor de droguri.

Intr-un program de dezintoxicare ce a cuprins 600 subiecţi, se notează cu a_i vârsta subiectului i la intrarea în program.

Rata hazardului a fost calculată ca fiind:

$$h(t,a) = h_0(t) \cdot \exp(-0.013a).$$

Se estimează că fiecare an din vârsta unui dependent multiplică riscul de a relua drogurile cu $e^{-0.013} = 0.99$. Nivelul de semnificaţie fiind $p = 0.07$ nu avem motiv să respingem ipoteza că vârsta nu are efect asupra reluării consumului de droguri.

2. În cazul a două variabile $(Z_1, Z_2) \in \mathbf{R^2}$, cu parametrii $(\beta_1, \beta_2) \in \mathbf{R^2}$, rata hazardului este:

$$h(t, Z_1, Z_2) = h_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2)$$

Dacă există o interacțiune între Z_1 și Z_2, parametrii sunt $(\beta_1, \beta_2, \beta_{12}) \in \mathbf{R^3}$ și rata hazardului este:

$$h(t, Z_1, Z_2) = h_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_{12} Z_1 \cdot Z_2)$$

Exemplu

Numărul de tratamente de dezintoxicare (droguri) şi numărul de tratamente antidepresive.

Se notează b_i numărul de tratamente de dezintoxicare ale subiectului i. b_i ia valori intre 0 şi 40.

Scorul de depresie Beck atribuit subiectului i se notează cu e_i. Acesta ia valori între 0 şi 54 şi anume: între 0 şi 9, depresie inexistentă, între 10 şi 18 depresie uşor moderată, între 19 şi 29 depresie sever moderată şi între 30 şi 54 depresie severă.

Rata hazardului este:

$$h(t,b,e) = h_0(t) \cdot \exp(0.030b + 0.010e)$$

Astfel raportul hazardului pentru un dependent tipic cu depresie severă - 40 şi un dependent cu depresie uşoară - 15 este $e^{25 \cdot 0.01} = 1.27$

Nivelul de semnificaţie p fiind mai mic de 0.05 indică un efect slab, în sensul că dependenti de droguri sever depresivi au mai multe sanse de recidivă.

Raportul hazardului pentru un dependent ce a urmat 20 tratamente de dezintoxicare în raport cu un dependent ce nu a urmat niciun tratament este $e^{20 \cdot 0.03} = 1.8$, ceea ce înseamnă că de 1.8 ori este mai probabil ca primul subiect să se drogheze din nou față de primul. Acest rezultat este foarte semnificativ, deoarece $p = 0.01$

Rata hazardului în cazul interacțiunii dintre numărul de tratamente de dezintoxicare și depresie este:

$$h(t, b, e) = h_0(t) \cdot \exp(0.079b + 0.043e + 0.0012b \cdot e)$$

S-a constatat că este totuși preferabil modelul care nu ia în seamă interacțiunea.

3. Dacă variabila $Z \in \{0,1\}$, (binară) și parametrul $\beta \in \mathbf{R}$, rata hazardului este

$$h(t,Z) = h_0(t)\exp(\beta Z).$$

În particular:

$$h(t,0) = h_0(t) \text{ și } h(t,1) = h_0(t) \cdot e^{\beta}$$

Rata hazardului pentru grupul 1 este de e^{β} ori mai mare decât rata hazardului la grupul 0.

Exemplu

Efectul rasei asupra tratamentului de dezintoxicare.

Subiecţii au fost clasificaţi în albi şi alţii. S-au codificat cu albii cu 0, alţii cu 1 şi s-a notat cu f_i rasa subiectului i.

Rata hazardului este

$$h(t, f) = h_0(t)\exp(-0.29 f)$$

şi rata hazardului pentru alţii este 0.75 din rata hazardului pentru albi.

Acest rezultat este foarte semnificativ, deoarece $p < 0.01$, ceea ce înseamnă că tratamentul are mai mult succes la alţii decât la albi.

4. Dacă variabilele sunt $Z_1 \in \{0,1\}$ şi $Z_2 \in \mathbf{R}$, cu parametrii $(\beta_1, \beta_2) \in \mathbf{R^2}$, şi respectiv $(\beta_1, \beta_2, \beta_{12}) \in \mathbf{R^3}$ în cazul în care există o interacţiune între Z_1 şi Z_2.

Pentru modelul fără interacţiuni, rata hazardului este:

$$h(t, Z_1, Z_2) = h_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2)$$

În particular

$$h(t, Z_1 = 0, Z_2) = h_0(t)\exp(\beta_2 Z_2)$$

$$h(t, Z_1 = 1, Z_2) = h_0(t)\exp(\beta_1 + \beta_2 Z_2)|$$

Pentru modelul cu o interacţiune, rata hazardului este:

$$h(t, Z_1, Z_2) = h_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_{12} Z_1 \cdot Z_2)$$

În particular

$$h(t, Z_1 = 0, Z_2) = h_0(t)\exp(\beta_2 Z_2 + \beta_{12} Z_1 \cdot Z_2)$$

$$h(t, Z_1 = 1, Z_2) = h_0(t)\exp(\beta_1 + \beta_2 Z_2 + \beta_{12} Z_1 \cdot Z_2)$$

Exemplu

rasa şi numărul de tratamente de dezintoxicare.

Fiecare individ este codificat $Z_1 = 0$, alb şi $Z_1 = 1$, pentru non-alb.

Pentru modelul fără interacţiune, funcţia hazard este:

$$h(t, f, Z_2) = h_0(t)\exp(-0.26Z_1 + 0.027Z_2)$$

în timp ce pentru modelul cu interacţiune, funcţia hazard este:

$$h(t, f, Z_2) = h_0(t)\exp(-0.25Z_1 + 0.027Z_2 - 0.001Z_1 \cdot Z_2)$$

11. Exemplu

Datele următoare reprezintă supraviețuirea, în zile, a unor pacienți cu limfom histiocitar difuz, pacienți ce fac parte din două grupuri distincte. Vor fi comparate două grupuri distincte: pacienți aflati în stadiul al 3-lea, respectiv în stadiul al 4-lea.

Stadiul 3: 6, 19, 32, 42, 42, 43*, 94, 126*, 169*, 207, 211*, 227*, 253, 255*, 270*, 310*, 316*, 335*, 346*

Stadiul 4: 4, 6, 10, 11, 11, 11, 13, 17, 20, 20, 21, 22, 24, 24, 29, 30, 30, 31, 33, 34, 35, 39, 40, 41*, 43*, 45, 46, 50, 56, 61*, 61*, 63, 68, 82, 85, 88, 89, 90, 93, 104, 110, 134, 137, 160*, 169, 171, 173, 175, 184, 201, 222, 235*, 247*, 260*, 284*, 290*, 291*, 302*, 304*, 341*, 345*

>>b = coxphfit(X,y) returnează vectorul coloană de dimensiune p ale cărui componente sunt coeficienții regresiei lui Cox a hazardelor (riscurilor) prognozate pentru răspunsurile y , pentru cei p predictori din matricea X.
X este o matrice cu p coloane şi n linii (n fiind numărul observaţiilor).

```
>> x1=[6, 19, 32, 42, 42, 43, 94, 126, 169, 207, 211, 227, 253, 255, 270,
310, 316, 335, 346];
>> x1cens=[1 1 1 1 1 0 1 0 0 1 0 0 1 0 0 0 0 0];
>> x2=[4, 6, 10, 11, 11, 11, 13, 17, 20, 20, 21, 22, 24, 24, 29, 30, 30, 31,
33, 34, 35, 39, 40, 41, 43, 45, 46, 50, 56, 61, 61, 63, 68, 82, 85, 88, 89,
90, 93, 104, 110, 134, 137, 160, 169, 171, 173, 175, 184, 201, 222, 235,
247, 260, 284, 290, 291, 302, 304, 341, 345];
>> x2cens=[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 0 0 1
1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0];
>> b1=[1]; for i=2:19 b1=[1 b1];end
>> b2=[2]; for i=2:61 b2=[1 b2];end
>> b=[b1 b2]';x=[x1 x2]';
>> xcens=[x1cens x2cens]';
>> A=[b x xcens];
>> b = coxphfit(A,x)
b =
    3.8410
   -3.2932
    0.0994
```

Aşadar rata hazardului este:

$$h(t,b,x,xcens) = h_0(t)\exp(3.8410b - 3.2932x + 0.0994xcens)$$

12. Exemplu

Vom genera aleator date cu repartiție Weibull:

```
>> x = 4*rand(100,1);
>> A = 50*exp(-0.5*x); B = 2;
>> y = wblrnd(A,B);
```

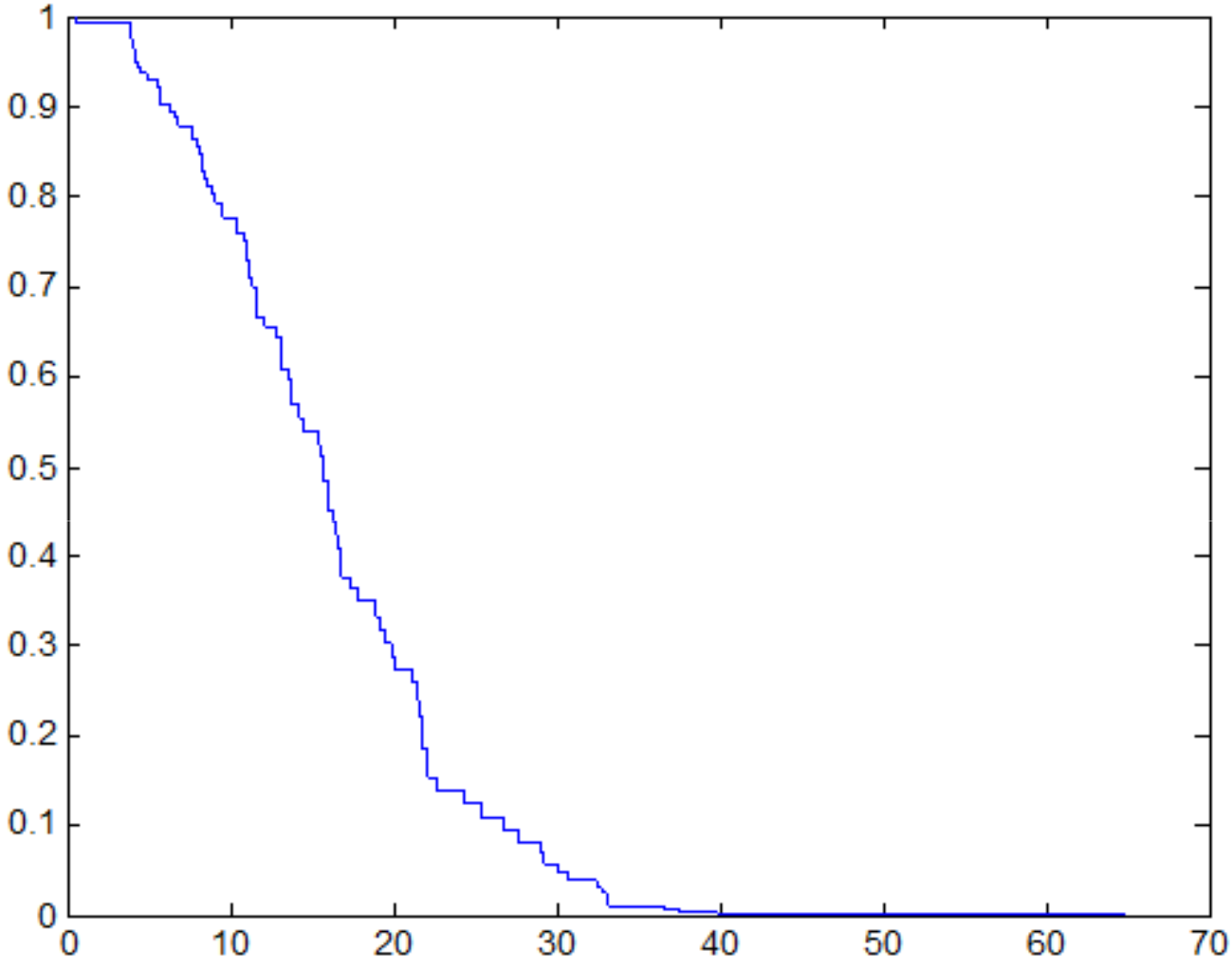
Vom implementa modelul Cox:

```
>> [b,logL,H,stats] = coxphfit(x,y);
```

Vom obține b =0.9409 și p = 6.9462e-014

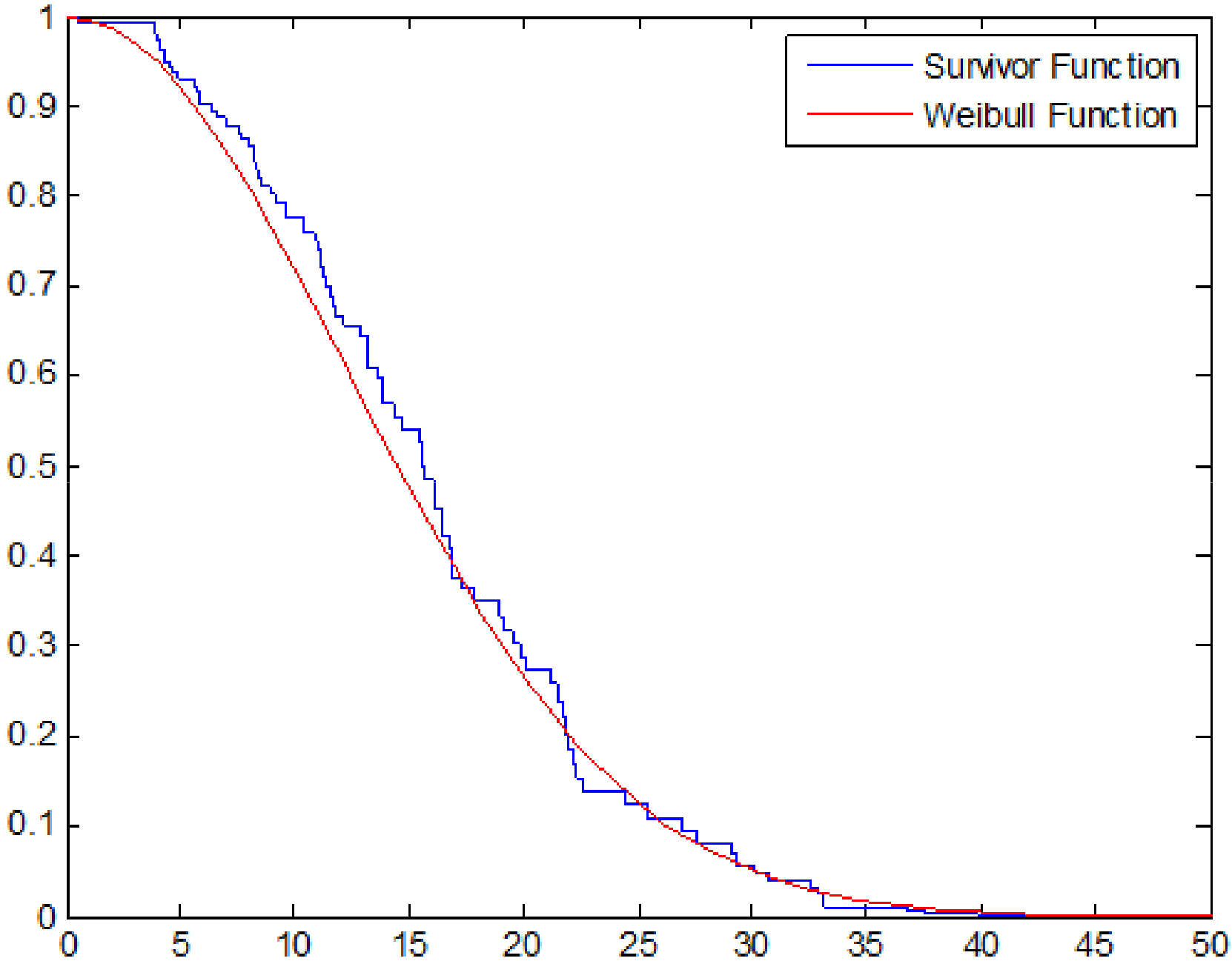
Vom desena funcția Cox estimată:

```
>>stairs(H(:,1),exp(-H(:,2)))
```


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70

Vom desena funcţia Cox estimată în acelaşi sistem de axe cu repartiţia Weibull:

```
>>xx = linspace(0,100);
>>line(xx,1-wblcdf(xx,50*exp(-0.5*mean(x)),B),'color','r')
>>xlim([0,50])
>>legend('Survivor Function','Weibull Function')
```


Survivor Function
Weibull Function
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
25
30
35
40
45
50

Modele aditive

Presupunem, la fel ca şi în cazul regresiei multiple, că avem o variabilă dependentă Y şi k variabile explicative (predictoare) $X_1, X_2, \ldots, X_k$.

Spre deosebire de cazul modelelor liniare, cazul modelului aditiv implică posibila legătură între variabila dependentă şi predictori sub forma:

$$Y = f_1(X_1) + f_2(X_2) + \ldots + f_k(X_k) + \varepsilon$$

unde f_j, $j = 1, 2, \ldots, k$ sunt, în general, funcţii de clasă C^∞, în unele cazuri şi funcţii de clasă C^1, iar ε este o variabilă aleatoare repartizată normal standard $N(0, 1)$.

Este uşor de observat că un model aditiv reprezintă generalizarea modelului regresiei liniare multiple (pentru $\varepsilon = 0$).

Cu alte cuvinte, în loc de un singur coeficient *per* variabilă explicativă, la modelele aditive găsim o funcţie nespecificată *per* fiecare predictor, care va trebui să fie estimată în vederea prognozei optime a valorilor variabilei dependente.

Ipoteza aditivităţii $\sum f_i(X_i)$ este o restricţie a cazului general al unui model predictiv de tipul $Y = f(X_1, X_2, ..., X_n)$.

Funcţiile parametru ale modelului aditiv sunt estimate până la o constantă aditivă.

Modele aditive generalizate

Un *model liniar generalizat* este reprezentat prin ecuaţia:

$$Y = g(b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + ..b_k \cdot X_k) \ ,$$

unde g este o funcţie de clasă C^{∞}, nedeterminată.

Dacă notăm formal g^{-1} inversa funcţie g, funcţie numită *funcţie legătură* (*link function*), atunci putem scrie ecuaţia de mai sus sub forma uşor modificată:

$$g^{-1}(E[Y]) = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + ..b_k \cdot X_k \ ,$$

unde $E[Y]$ reprezintă media variabilei dependente Y.

Combinând un model aditiv cu un model liniar generalizat, vom obține ecuația modelului, cunoscut sub denumirea de *model aditiv generalizat*, sub forma:

$$g^{-1}(E[Y]) = f_1(X_1) + f_2(X_2) + ... + f_k(X_k) .$$

Problema de bază la aceste modele este estimarea funcțiilor parametri f_i ale modelului.
Cea mai cunoscută metodă de evaluare a funcțiilor f_i este reprezentată de o interpolare pe bază de diagrame de dispersie (*scatterplot smoother*), utilizând funcții spline cubice.

De exemplu, în cazul unui model simplu cu doar două funcţii f_1 şi f_1, având forma:

$$Y = f_1(X_1) + f_2(X_2) + \varepsilon \,,$$

folosind aproximarea spline, se obţine forma celor două funcţii:

$$f_1(X) = \delta_1 + X \cdot \delta_2 + \sum_{j=1}^{q_1-2} R(X, X_j^*) \cdot \delta_{j+2} \,,$$

$$f_2(X) = \gamma_1 + X \cdot \gamma_2 + \sum_{j=1}^{q_2-2} R(X, X_j^*) \cdot \gamma_{j+2} \,,$$

unde δ_j, γ_j sunt parametrii necunoscuţi ai funcţiilor f_1 şi f_2, q_1, q_2 reprezintă numărul parametrilor necunoscuţi, iar x_j^* reprezintă nodurile de interpolare pentru cele două funcţii.