IDENTIFICAREA SISTEMELOR LOGISTICE

Mihai Tertişco

Universitatea Politehnica din Bucureşti

e-mail: tertisco_mihai@yahoo.com

Gabriel-Cristian Ene

Spitalul de Urgență -Târgoviște

e-mail: enegabrielcristian@yahoo.com

Cristian Eremia

Universitatea Politehnica din Bucureşti

e-mail: cristian_valentin2003@yahoo.com

Rezumat: Lucrarea prezintă rezultate ale cercetării ştiinţifice cu privire la modelarea şi identificarea experimentală a proceselor logistice cu evenimente aleatoare binare.

Cuvinte cheie: modelare, identificare, regresii logistice, metoda Monte-Carlo.

Abstract: The paper presents results of the scientific research regarding modelling and experimental identification of logistical processes with binary random events.

Key words: modelling, identification, logistic regression, Monte Carlo methods

1. Particularităţile sistemelor logistice

Sistemele logistice sunt descrise de modele probabilistice care descriu o mulţime omogenă M de cardinalitate NS, formată din două tipuri distincte de entităţi care pot fi separate în două clase. Fiecare entitate din aceasta populaţie este caracterizată de o variabilă dependentă Y (de ieşire) şi una sau mai multe variabile independente (de intrare) X. Variabila Y poate lua numai valori logice: 1 sau 0; DA sau NU; bolnav ori sănătos etc. Variabila independentă fie poate lua valori logice, fie poate lua valori în mulţimea numerelor reale. În majoritatea aplicaţiilor întâlnite în literatura de specialitate aceste variabile iau valori logice, 0 sau 1. Pe baza testării experimentale a fiecărei entităţi din cele N se pot împărţi entităţile în două clase: clasa entităţilor cu Y=1 şi clasa entităţilor cu Y=0. Modelul în care avem doar o singură variabilă independentă x îl vom numi model logistic SISO (single - input - single - output). În cazul mai multor variabile independente modelul se numeşte MISO (multy - input - single - output). Pentru simplitate ne vom referi cu precădere la modele logistice de tip SISO. În biostatistică modelul folosit pentru a analiza relaţia probabilistică între variabila dependentă, de tip binar, şi una sau mai multe variabile independente X este denumit *regresie logistică*. De exemplu, sunt făcute publice (pe INTERNET) datele experimentale din figura 1 privind analiza unei probabile cauze ale dezastrului navetei spaţiale Challenger (1986) în care se prezintă diversele temperaturi la care a apărut sau nu defectarea unei legături mecanice specifice de-a lungul a N=23 de încercări.

x temperatura	Y defect	x temperatura	Y defect	x temperatura	Y defect
66	0	57	1	70	0
70	1	63	1	81	0
69	0	70	1	76	0
68	0	78	0	79	0
67	0	67	0	75	1
72	0	53	1	76	0
73	0	67	0	58	1
70	0	75	0		

Figura 1. Date experimentale privind analiza unei probabile cauze a dezastrului navetei spaţiale Challenger (1986).

Cunoscând rezultatele acestor N încercări ale navetei am putea construi un model cu care să putem răspunde la întrebarea: „***care este probabilitatea ca defecţiunea respectivă să apară (Y=1) la o temperatură dată x ?".*** În acest exemplu clasa de evenimente Y=1 conţine următoarele categorii de evenimente aleatoare determinate de valorile variabilei independente x:

{(Y=1, x=75(Y=1, x=70);(Y=1, x=63);(Y=1, x=58);(Y=1, x=58);(Y=1, x=57)}

2. Structura modelelor logistice

Ecuaţia de regresie obţinută în acest caz este de un tip diferit de celelalte regresii cunoscute, cum ar fi cele continue, monodimensională, multidimensională, liniare şi neliniare etc.

În figura 2 sunt prezentate trei variante ale structurii modelului logistic pentru un sistem SISO (*single input single output*), întâlnite în literatura de specialitate [1]. În prima variantă (2a) mărimea continuă ***p este*** o funcţie neliniară de x şi de doi parametri necunoscuţi: α şi β. Dacă evenimentul y=1, atunci apariţia acestui eveniment are loc cu probabilitatea ***P(y=1)=p***. Acest tip de regresie oferă informaţii despre ***importanţa variabilelor x în diferenţierea claselor,*** şi ***despre clasificarea unei observaţii într-una din clase.*** Spre deosebire de regresia liniară clasică, în cazul regresiei logistice (fig. 2a), în locul variabilei dependente Y, *care poate lua valoarea binară Y=1 → "succes" ori Y=0 → "eşec",* este folosită o variabilă continuă ***p*** care ia valori între 0 şi 1. O valoare a lui ***p*** este interpretată ca ***probabilitatea*** de a obţine un „succes" (Y=1), condiţionată de valoare variabilei independente ***x***. Atunci evenimentul contrar y=0 are probabilitatea de apariţie ***P(y=0)=1-p.***

$$P(y=1|x)=\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$ (a) regresie logistică

$$\frac{p}{1-p}=e^{\alpha+\beta x}$$ (b) model de tip „şansă"

$$\ln\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right)=\alpha+\beta x$$ (c) model de tip „logit"

Figura 2. Trei variante ale modelelor logistice de tip SISO

Acest tip de regresie oferă informaţii despre importanţa variabilelor ***x*** în diferenţierea fiecărei entităţi dintr-o mulţime dată de ***n*** entităţi sortându-le pe categorii***.*** *În domeniul medical aceste entităţi pot fi n pacienţi investigaţi prin testare directă pentru a stabili, spre exemplu, dacă este* ***"bolnav Y=1 şi fumător x=1„*** ori ***„bolnav Y=1 şi nefumător x=0„*** sau dacă este ***„fumător x=1 dar sănătos Y-0"*** ori ***„nefumător x=0 şi sănătos Y=0"***. Prezenţa categoriilor este hotărâtă de variabilele ***x*** *care mai* sunt denumite şi *variabile categoriale* ori *predictoare.* Rezultatul acestor investigaţii experimentale constituie datele necesare estimării parametrilor modelului logistic. Pentru obţinerea unui model mai comod (uşor liniarizabil în parametri) s-a introdus modelul cu structura din figura 2b, numit model de tip „şansă". Modelul de tip „şansă" este exprimat de raporul ***p/(1-p)*** numit ***raport*** de ***şansă*** (***odds report,*** sau pe scurt, ***OR***). Acest raport este o funcţie tot ***neliniară de x şi de parametri*** α şi β, dar ***uşor de liniarizat prin logaritmare***. Rezultatul logaritmării este modelul din figura 2c, numit, în literatura de specialitate, modelul ***logit(p) = ln[p/(1-p)]*** care, spre deosebire de celelalte, este liniar în parametri [4]. Se poate constata imediat că, pentru o observaţie din setul respectiv de observaţii experimentale (date experimentale), dacă $p > 0,5$, atunci este mai probabil ca observaţia să aparţină grupului caracterizat de $y = 1$. Această condiţie este echivalentă cu: ***OR > 1***, adică ***logit > 0.*** În ceea ce priveşte semnificaţia parametrului β el ***exprimă creşterea cantităţii logit***

atunci când x creşte cu o unitate sau, în cazul binar, când devine x = 1.

Structura modelului logit, în cazul multivariabil MISO (multy-input –single-output) este:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \quad (1)$$

în care variabilele categoriale sunt: $x_1 ... x_k$.

3. Estimarea parametrilor regresiei logistice aplicând criteriul verosimilității logaritmice maxime

În cazul identificării experimentale a unui sistem de tip SISO cu evenimente aleatoare binare este utilizat un model logistic. Pentru identificarea unui proces logistic sunt folosite *n* perechi de date experimentale intrare - ieşire în *n puncte*, (*Y1, x1*); (*Y1, x2*);...; (*Y1, xn*). În cazul investigării a *n* pacienți, dintre care unii bolavi de cancer, datele sunt obținute prin observații succesive directe asupra acestei mulțimi de *n* entități în care fiecare entitate ***i*** din cele **n,** este caracterizată de perechea de valori Y_i şi x_i. Se observă că, în funcție de valorile lui x, acest model atribuie valori ale probabilității prezenței cancerului (când Y = 1) fiecărei entități din mulțimea de *n* entități în funcție de valoarea lui *x*. De exemplu, atunci când x este o variabilă dihotomică şi ia valorea 0 (de exemplu *nefumător*i) sau 1 (de exemplu *fumători*) pentru toate cele *n* entități investigate experimental mulțimea se împarte în 2 categorii :

- categoria celor pentru care probabilitatea de prezență a cancerului este

$$\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$$

când x = 1 şi Y = 1;

- categoria celor pentru care probabilitatea de prezență a cancerului este

$$\frac{e^{\beta_0}}{1+e^{\beta_0}}$$

când x = 0 şi Y = 1.

Pentru identificarea unui proces logistc de tip SISO sunt folosite cele *n* perechi de date intrare – ieşire în *n puncte* obținute experimental. Aceste date sunt obținute prin observații succesive directe asupra acelei mulțimi de *n* entități în care fiecare entitate ***i*** este caracterizată de perechea de valori (Y_i şi x_i). Pe baza acestor n perechi de date experimentale trebuie calculate acele valori ale celor doi parametri necunoscuți α şi β încât modelul estimat (având structura 2a) să poată descrie în mod ***optim*** datele experimentale şi să asigure un grad ridicat de generalitate, în sensul de a fi capabil să descrie corect comportarea procesului logistic respectiv şi în alte puncte (Y, x), care nu fac parte din mulțimea inițială de *n* puncte ale datelor experimentale. Printre aceste puncte din setul de date experimentale sunt unele în care Y=1 şi altele în care Y = 0. Având în vedere că ***ieşirea procesului*** este o variabilă logistică, care în cadrul experimentului ia valorile $Y_1, Y_2, ..., Y_n$, iar ***ieşirea modelului*** în unele puncte experimentale sunt probabilitățile $p(Y_i=1|\ x_i)$, iar in celelalte puncte $p(Y_i=0|\ x_i) = 1 - p(Y_i=1|x_i)$,. caracterizarea întregului ansamblu *n* de evenimente independente aleatoare de tip logistic este exprimată de produsul celor *n* probabiliăți aferente evenimentelor aleatoare binare observate:

$$P = \prod_{i=1}^{n} p_i\,(Y, x, parametri) \quad (2)$$

În cadrul acestui produs sunt două tipuri de termeni: termeni corespunzători evenimentelor Y = 1 pentru care $p_i = p = Pr(Y_i=1, x_i, parametri)$ şi termeni aferenți evenimentelor Y = 0

pentru care $p_i = 1-p$. În aceste condiții relația (2) devine:

$$P = \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1-Y_i} \tag{3}$$

Cu notațiile din (1) modelul de regresie logistică, în cazul SISO, când Y = 1, devine

$$p = \Pr(Y = 1 \mid x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{4}$$

şi respectiv pentru evenimentul contrar Y = 0 va fi:

$$1\text{-}\, p = \Pr(Y = 0 \mid x) = \frac{1}{1+e^{\beta_0+\beta_1 x}} \tag{5}$$

În statistică, funcția de probabilitate (2) se notează *L (date, parametri)* şi este denumită ***funcție de veromsimilitate*** (**likelihood function**) deoarece ea exprimă cât de ***verosimil*** este modelul propus pentru întreg şirul de *n* evenimente experimentale date. Dacă în toate punctele de tip (4) şi de tip (5) modelul este verosimil, funcțiile (4) şi (5) iau valori maxime în aceste puncte, iar dacă modelul este verosimil pe tot ansamblul celor *n* puncte experimentale, funcția de verosimilitate (3) capătă de asemenea valoarea maximă. Funcția de verosimilitate depinde de parametrii regresiei logistice şi de datele experimentale, având următoarea expresie în cazul modelului logistic pentru procesul cu evenimente aleatoare binare de tip SISO:

$$\begin{aligned} L\left((\beta_0,\beta_1);Data\right) &= \prod_{i=1}^{n}\left(\frac{e^{(\beta_0+\beta_1 X_i)}}{1+e^{(\beta_0+\beta_1 X_i)}}\right)^{Y_i}\left(\frac{1}{1+e^{(\beta_0+\beta_1 X_i)}}\right)^{1-Y_i} \\ &= \prod_{i=1}^{n}\frac{(e^{(\beta_0+\beta_1 X_i)})^{Y_i}}{(1+e^{(\beta_0+\beta_1 X_i)})} \end{aligned} \tag{6}$$

În cazul identificării proceselor logistice problema constă în găsirea acelor valori pentru parametrii modelului care vor asigura maximul funcției de verosimilitate. Aceste valori sunt notate ($\hat{\beta}_0$ *şi* $\hat{\beta}_1$) şi ele constituie aşa numitele *estimații ale parametrilor modelului în sensul verosimilității maxime.* Problema estimațiilor de ***verosimilitate maximă (VM)*** pentru parametrii β_0 şi β_1 constă în determinarea acelor valori $\hat{\beta}_0$ *şi* $\hat{\beta}_1$ care maximizează funcția de verosimilitate $L((\beta_0, \beta_1);Data)$ care are expresia (6).

Observație: întrucât funcția *ln(x)* este monoton crescătoare în tot domeniul de definiție, maximul oricărei funcții L(β) este şi maximul funcției ln[L(β)].

Rezolvarea problemei de căutare a valorilor ($\hat{\beta}_0$ *şi* $\hat{\beta}_1$) care maximizează (6) este destul de dificilă datorită prezenței produsului funcțiilor exponențiale. Aplicând ***logaritmul natural*** funcției de VM din (6) rezultă funcția de ***logverosimilitate(VL) pentru evenimente binare aleatoare cu model logistic SISO***. Această funcție notată ***VL= $l(\beta_o, \beta_1)$*** are expresia:

$$l(\beta_o,\beta_1) = \ln[L(\beta_o,\beta_1)] = \sum_{i=1}^{n} Y_i(\beta_o + \beta_1 X_i) - \sum_{i=1}^{n} \ln[1+e^{(\beta_o+\beta_1 X_i)}] \tag{7}$$

Funcțiile ***$L((\beta_o, \beta_1)$*** şi ***$l(\beta_o, \beta_1)$*** au maximul în planul parametrilor ***β_o, β_1,*** în acelaşi punct de coordonate: ($\hat{\beta}_0$ *şi* $\hat{\beta}_1$), care reprezintă estimațiile de verosimilitate maximă ale parametrilor modelului logistic SISO.

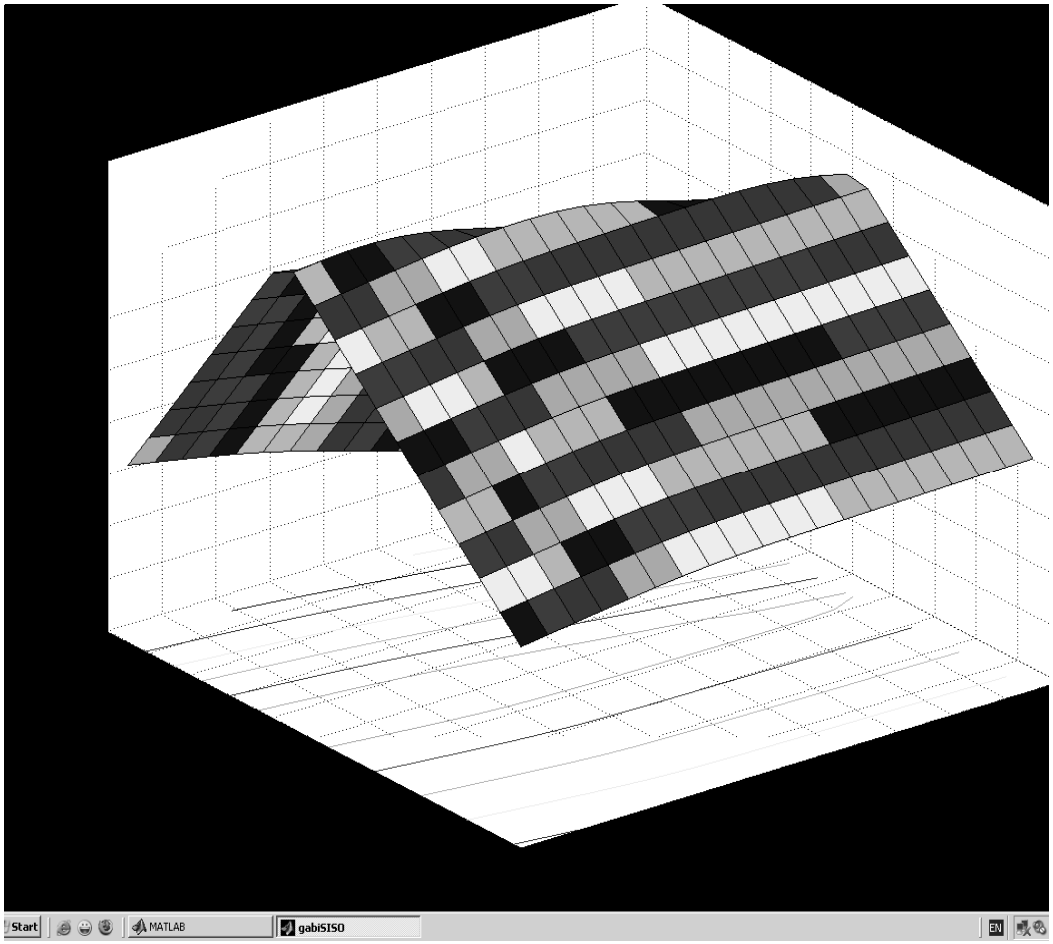

Figura 3. Imaginea grafica a suprafeței descrise în spațiul parametrilor $\hat{\beta}_0$ şi $\hat{\beta}_1$ de către funcția de verosimilitate logaritmică (VL)

Pentru rezolvarea problemei de maxim avem la dispoziție o mulțime de **n** perechi de date experimentale, observații intrare – ieşire, **data={ (X_1,Y_1), (X_2,Y_2),…, (X_n,Y_n) }**, care pentru cazul concret din figura 1 devin:

$$\text{data}=\{(X_1=66,Y_1=0),(X_2=70,Y_2=1),\ldots,(X_{22}=76,Y_{22}=0),(X_{23}=76,Y_{23}=1)\} \quad (8)$$

Folosind aceste date a fost elaborat un program MATLAB care a construit imaginea grafică din figura 3 a verosimilității logaritmice (7) pentru exemplul din figura1. În figura 3 se poate observa platoul din zona de extrem, care a fost pus în evidență, şi figura 4 care conține imaginea liniilor de izonivel ale aceleiaşi funcții de verosimilitate logaritmică. Aceste linii de izonivel au fost trasate prin intermediul aceluiaşi program MATLAB menționat mai sus. Având în vedere aspectele menționate cu privire la geometria suprafeței de verosimilitate logaritmică în cazul datelor experimentale, am apelat la o metodă de tip Monte Carlo pentru găsirea punctului de maxim.

Metoda propusă constă în testarea aleatoare a suprafeței verosimilității, folosind două semnale aleatoare de testare Sb_1 şi Sb_0 , câte unul pentru fiecare parametru. Aceste semnale au distribuții de probabilitate uniforme în banda de investigare din planul parametrilor [3].

Semnalele aleatoare sunt obținute de la două generatoare de numere aleatoare simulate în matlab.

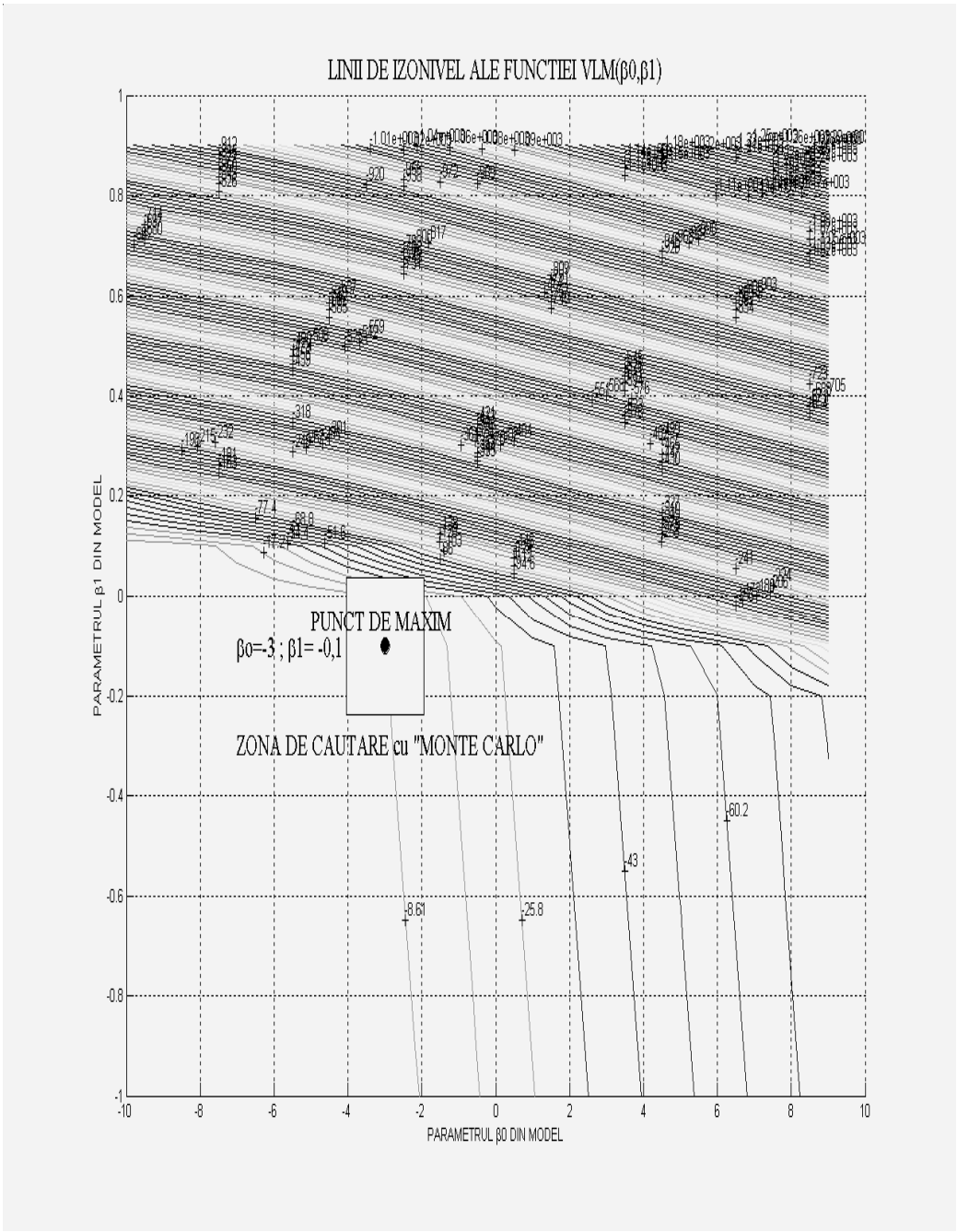

Figura 4. Liniile de izonivel ale suprafeței funcției de verosimilitate logaritmică prezentată în figura 3

Algoritmul Monte Carlo de căutare aleatoare a maximului verosimilității logaritmice ***VL(β_o, β_1)*** presupune execuția a trei paşi descrişi în figura 5.

Pasul 1: Se generează o pereche de numere aleatoare [Sb0(k=1),Sb1(k=1)]: şi cu aceste valori şi datele experimentale existente se calculează verosimilitatea şi se memorează,

VL(Sb0(1),Sb1(1),date)=VL(1)→M

Pasul 2: Se incrementează cu o unitate variabilă de contorizare k =k+1 a numărului de testări şi se generează o nouă pereche de numere aleatoare cu care se calculează,

VL(Sb0(2),Sb1(2),date)=VL(2)

Pasul 3: Se compară VL(2) cu M de la pasul precedent:

DACA

VL(2) > M

ATUNCI

se înlocuieşte vechiul conținut din M cu VL(2): VL(2)→M se revine la Pasul 2.

ALTFEL

se revine la Pasul 2, păstrând în M valoarea precedentă.

Figura 5. Algoritmul Monte Carlo de căutare aleatoare a maximului verosimilității logaritmice

Limitele care determină zona de căutare, *b0min, b0max, b1min, b1max* sunt fixate prin tatonări prealabile efectuate în colțurile unui pătrat şi în centru, ca în figura 4 în care coordonatele celor 5 puncte sunt:

1) centrul experimentului de testare a zonei de cautare $\beta_{1centru}, \beta_{0centru}$

2) $\beta_{\min 1}, \beta_{0centru}$

3) $\beta_{\max 1}, \beta_{0centru}$

4) $\beta_{0\min}, \beta_{1centru}$

5) $\beta_{0\max}, \beta_{01centru}$

În figura 4 este reprezentată zona de căutare determinată prin tatonări experimentale prealabile pe suprafața funcției logaritmice de verosimilitate. **Condiția de oprire** a algoritmului MONTE-CARLO se exprimă fie prin fixarea numărului maxim de încercări Kmax, fie prin impunerea numărului paşilor consecutivi de succes Ks, executați în zona de căutare delimitată experimental. Aplicarea metodei Monte Carlo pentru datele din figura 1 a localizat punctul de maxim al funcției logaritmice de verosimilitate în planul parametrilor (punctul de coordonate, β_0=-0,1, β_1 =-3, din figura 4).

4. Concluzii

Lucrarea scoate în evidenţă particularităţile modelelor proceselor logistice cu evenimente aleatoare binare şi prezintă tehnica identificării acestor procese, care implică pentru estimarea parametrilor modelului aplicarea criteriului statistic al verosimilităţii maxime. Este propusă o metodă MONTE CARLO pentru estimarea parametrilor modelului logistic.

BIBLIOGRAFIE

1. **BĂDULESCU, F., F. GORUNESCU:** Informatica oncologică: metode statistico-informatice în oncologie, Ed. Didactică şi Pedagogică 2003.
2. **GORUNESCU, M.:** Optimal Hospital Beds Allocation Using Queuing Systems. Proceedings 3rd Romanian Conference on Artificial Intelligence and Digital Communications – AIDC2003, Research Notes in Artificial Intelligence and Digital Communications (Universitaria Publishing House), Craiova, 103, ISBN 978-8419-71-9, pp. 131-136.
3. **TERTŞCO, M., P. STOICA:** Identificarea asistată de calculator. Editura Tehnică, 1980.
4. **ENE, G., E. FIROIU, S. ANGELESCU:** Sistem informatic de procesare evaluare şi monitorizare a datelor screening-ului în cancerul colului uterin. A III-a conferinţă naţională de oncologie medicală, Mamaia 9-12 septembrie 2004.
5. **ENE, G.:** Aplicarea în domeniul oncologiei medicale a metodei VaR de evaluare a riscului. În: Revista Română de Informatică şi Automatică nr. II/2007, pp. 63-68.