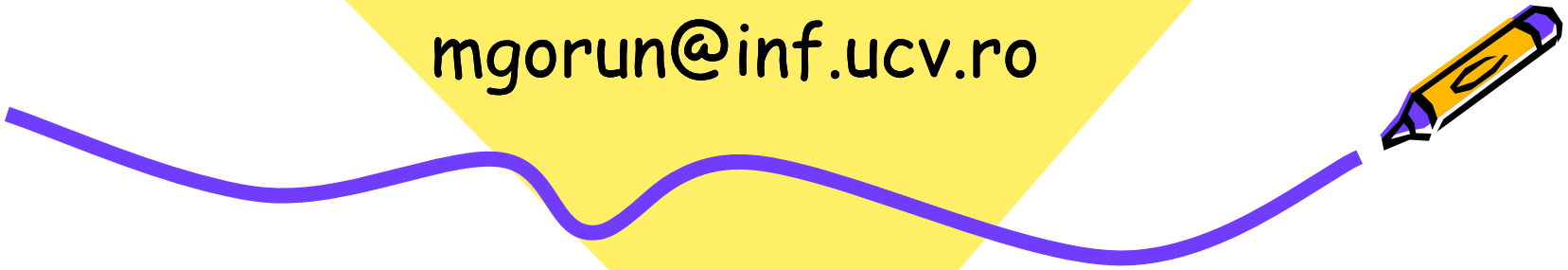
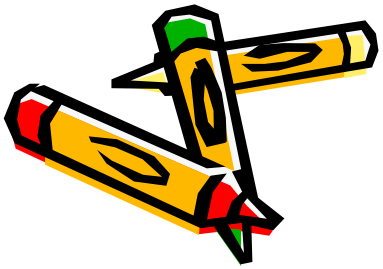
Regresia logistica

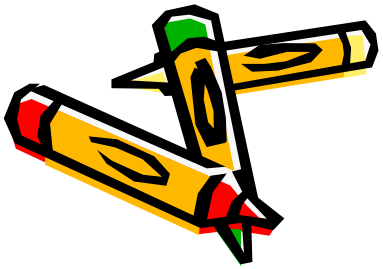
Marina Gorunescu
mgorun@inf.ucv.ro



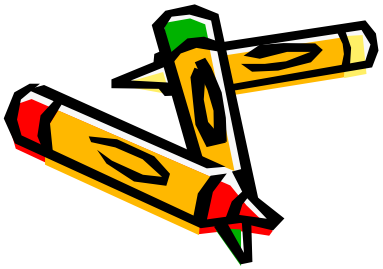
O problemă binecunoscută în multe arii de cercetare este aceea care presupune existența unui set de date privind două sau mai multe variabile aleatoare, scopul modelării fiind descrierea relației dintre ele în vederea prognozării valorilor uneia în raport cu valorile celeilalte sau celorlalte.



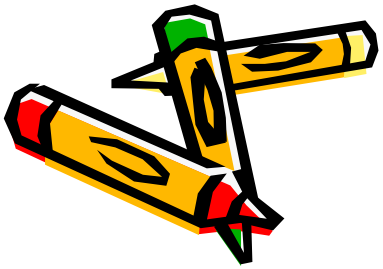
Această problemă se pune atunci când între variabilele aleatoare considerate există o legătură consistentă, bazată pe natura intimă a fenomenelor care stau la baza lor.



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

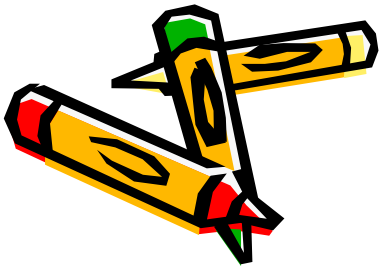


În concluzie, fără cunoaşterea naturii intime a fenomenelor care stau la baza datelor, este hazardat de a întreprinde o analiză regresivă.

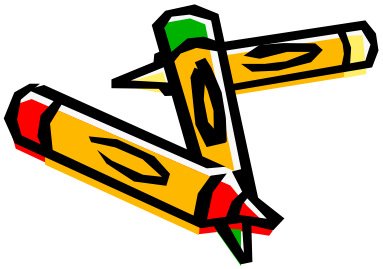


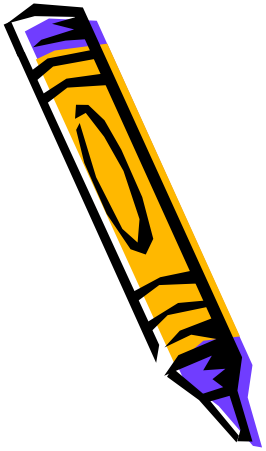
metoda regresiei

Metoda care se foloseşte pentru a descrie relaţia între valorile a două sau mai multe variabile aleatoare se numeşte *metoda regresiei* (noţiune introdusă de Pearson în 1908).

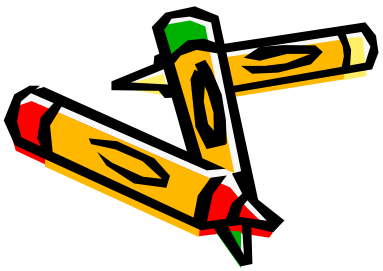


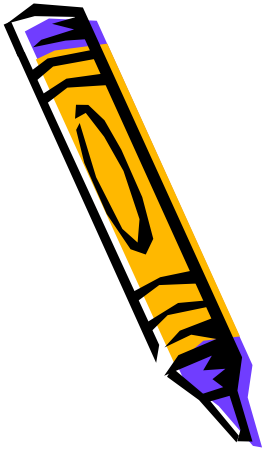
Dacă relația care stabilește legătura între variabila dependentă și variabilele independente este una liniară, se vorbește despre *regresia liniară*, în celălalt caz fiind vorba de *regresia neliniară* (polinomială, exponențială, logaritmică etc.).



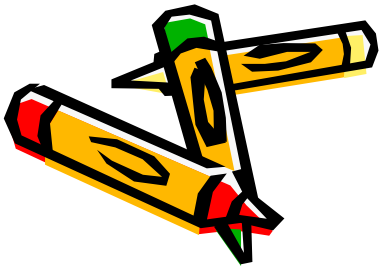


Vom prezenta pe scurt bazele regresiei liniare, începând prin a pune în evidență mecanismele cu care se pot evidenția legăturile între secvențe de date, provenind de la două sau mai multe variabile aleatoare.

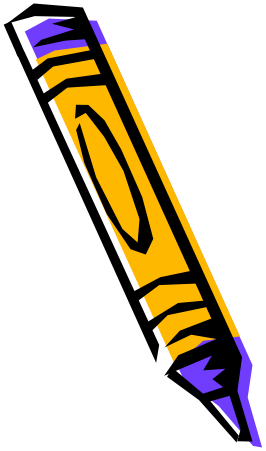




Cele mai multe studii statistice se referă la compararea a două sau mai multe grupuri de subiecți/obiecte sau la stabilirea unor legături existente între aceste grupuri.

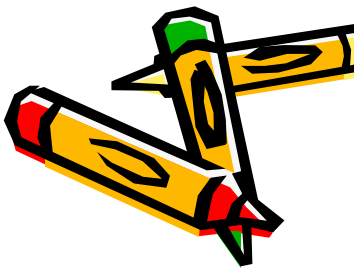


exemple

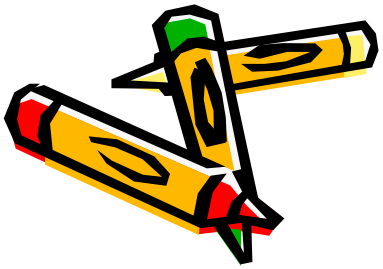


Analiza modului în care sunt corelate sau nu anumite valori medicale (colesterol, albumină, trigliceride etc.) în diferite cazuri, de exemplu pentru bărbați sau femei:

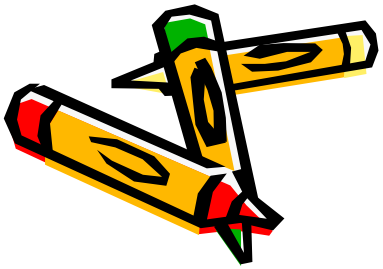
astfel, pe de o parte se identifică posibile legături între aceste caracteristici medicale la fiecare grup în parte şi, pe de altă parte, se pun în evidență anumite deosebiri ce pot exista între grupuri, privitoare la studiul efectuat.



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



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

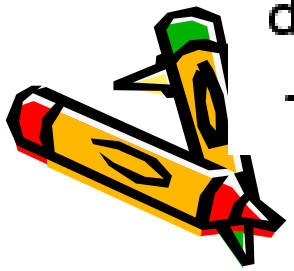


metoda diagramei de imprastiere

Să considerăm mai întâi 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 subiecți/obiecte.

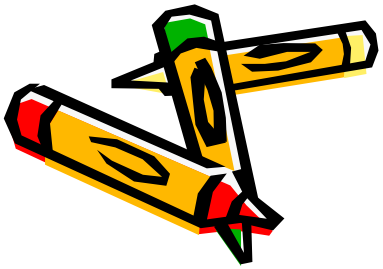
Plecând de la cele două serii, putem considera seria cuplurilor de observații $\{(x_i, y_i)\}_{1\le i\le n}$ definite de cele două variabile statistice pe acelaşi individ/obiect *i*.

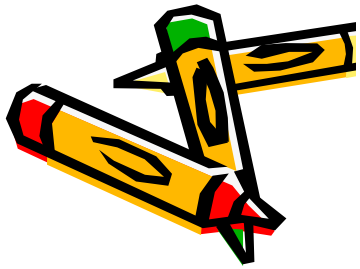
Cel mai obişnuit mod de reprezentare grafică al acestui cuplu de observații este cel folosind *norul* de puncte definit de reprezentarea bidimensională a punctelor (x_i, y_i) - aşa numita *diagramă de împrăştiere*.



exemplu

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





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

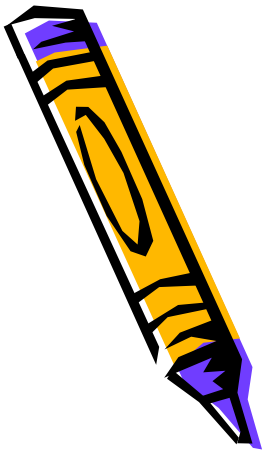
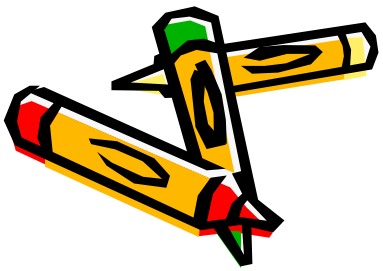


Diagrama ‚norului' de împrăștiere a cuplurilor (AST, gGT)
pentru lotul de 25 subiecți:

```
» x=[22 24 10 32 24 53 24 58 13 30 22 29 18 30 16 12 33 33 9
  21 17 54 32 19 23];
» y=[11 40 10 6 20 15 9 34 22 10 4 13 3 7 24 16 27 9 5 11 10 15
  8 5 4 ];
»  plot(x,y,'kO')
```



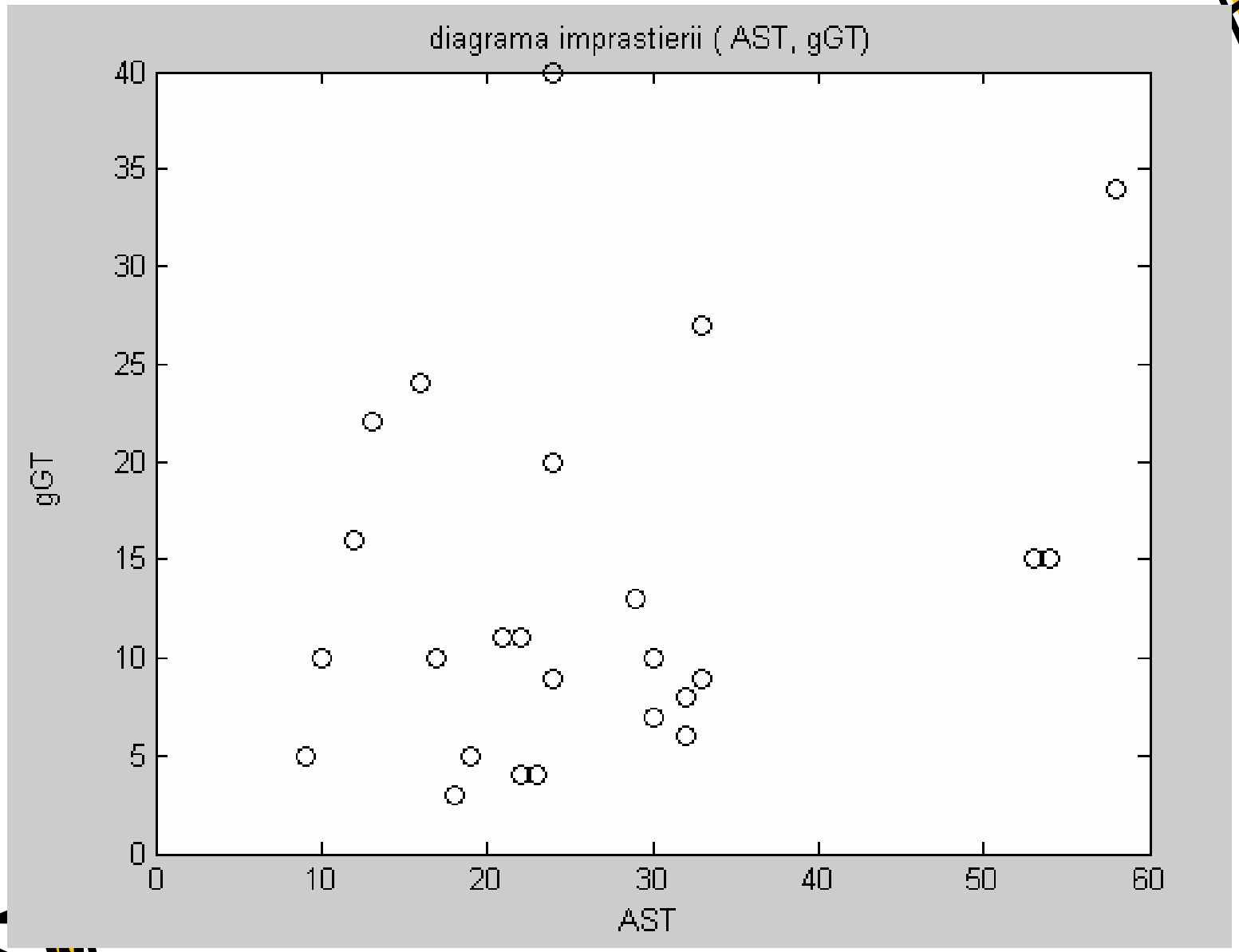
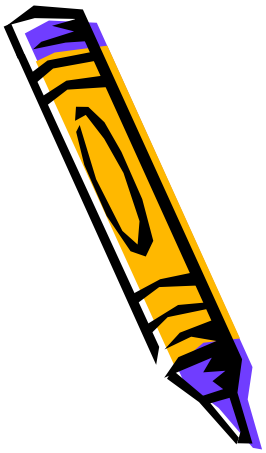


diagrama imprastierii (AST, gGT)
gGT
AST
0
5
10
15
20
25
30
35
40
0
10
20
30
40
50
60

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

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

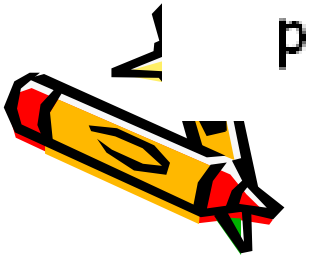


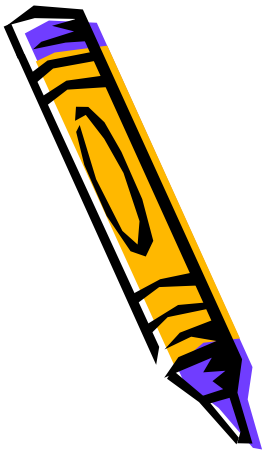
- norul are forma unei elipse mai mult sau mai puțin alungite, paralelogram alungit, figură geometrică alungită simetrică față de o axă, fapt ce implică o legătură liniară între variabile;
- norul are forma unui cerc sau pătrat, fapt ce implică independența variabilelor.

metoda prezentarii numerice a datelor multiple

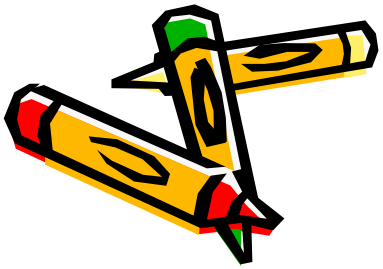
Prezentarea este făcută prin tabele, existând câteva reguli pentru obținerea unui efect semnificativ.

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



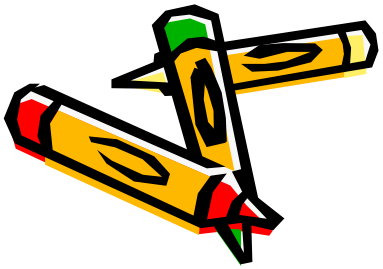


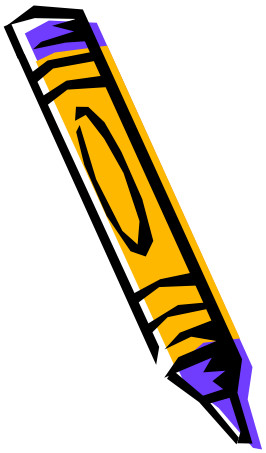
O mare parte a studiilor statistice uzuale se ocupă cu analiza relației între două variabile statistice ce corespund aceluiași grup de subiecți/obiecte.
Cel mai cunoscut exemplu se referă la relația ce există între înălțimea și greutatea unui individ ce corespunde unor anumite standarde geografice, rasiale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



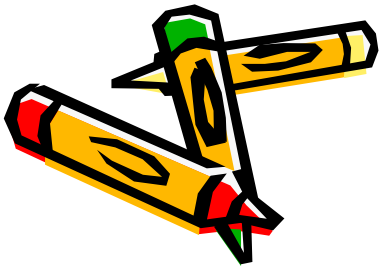
Pentru a o identifica, se studiază relația dintre cele două caracteristici măsurate pe indivizii dintr-un anumit lot.

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

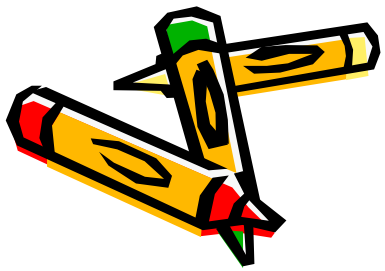




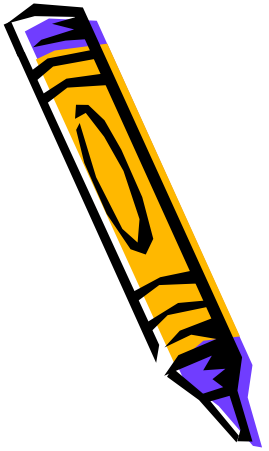
1. Descrierea relației care ar putea exista între cele două variabile, analizând legătura între cele două serii de observații.
   Concret, se analizează dacă tendința ascendentă a uneia implică o tendință ascendentă, descendentă sau nici o tendință a celeilalte;



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

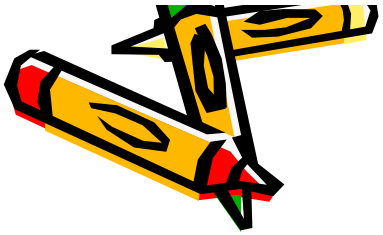


Metoda corelatiei



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

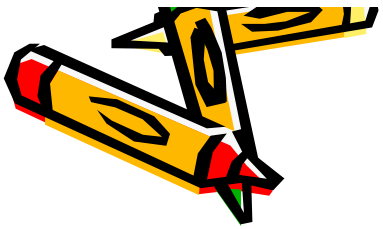
Metoda prin care analizăm posibilele asociații între valorile a două variabile statistice continue prelevate de la același grup de subiecți, este cunoscută ca metoda *corelației* și are ca indice *coeficientul de corelație*.



coeficientul de corelatie

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

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

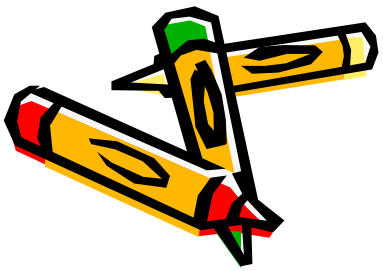


2. cel puțin una din variabile să aibă o repartiție aproximativ normală, ideal fiind ca ambele să fie normal repartizate.

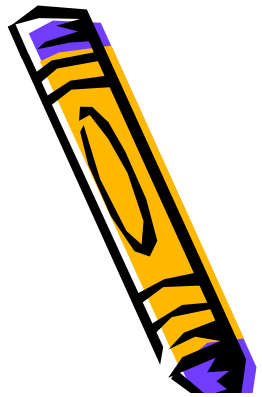
Dacă datele nu au o repartiție normală (cel puțin una din variabile) se procedează fie la transformarea lor pentru normalizare, fie la considerarea unor coeficienți de corelație ne-parametrici.

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

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



Coeficientul de corelatie

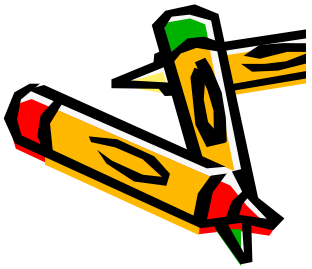


Să considerăm două serii statistice $\{x_i\}_{1\le i\le n}$ şi $\{y_i\}_{1\le i\le n}$ corespunzătoare variabilelor statistice X şi Y, generate de un grup de subiecți/obiecte.

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

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

(folosită în programele de computer):



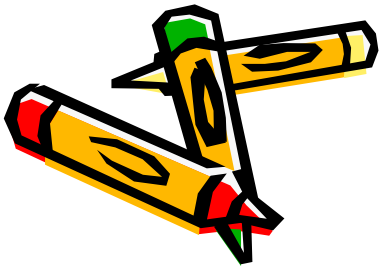
Pentru calcule concrete (manuale, cu ajutorul calculatorului de buzunar) se foloseşte formula de mai sus, scrisă sub forma:

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

exemplu

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

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



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

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

```
»G=[15.3 10.7 8.1 19.5 7.2 5.3 9.3 11.1 7.5 12.2 6.7 5.2 19 15.1
6.7 8.6 4.2 10.3 12.5 16.1 13.3 4.9 8.8 9.5];
»V=[1.76 1.34 1.27 1.47 1.27 1.49 1.31 1.09 1.18 1.22 1.25 1.19
1.95 1.28 1.52 1.65 1.12 1.37 1.19 1.05 1.32 1.03 1.12 1.70];
» A=[G' V'];
»corrcoef(A)
   ans =
      1.0000    0.4189
      0.4189    1.0000
```

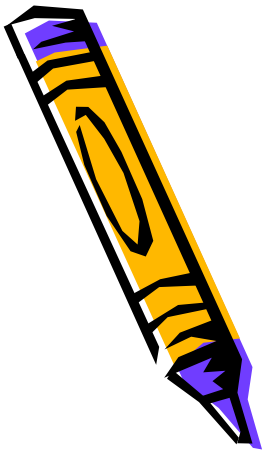
Coeficientul de corelație a celor două variabile este 0.4189.

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

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

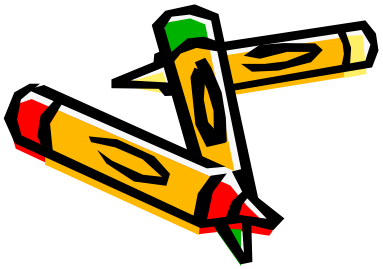
este normal repartizată, rezultă că intervalul de încredere 95% pentru z are forma (z_1, z_2) unde:

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

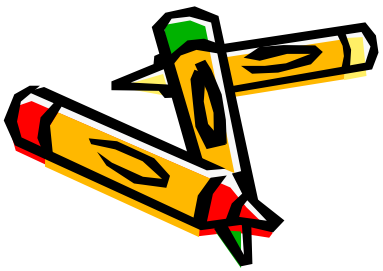


de unde rezultă că, aplicând transformarea inversă, obținem intervalul de încredere 95% pentru r, dat de:

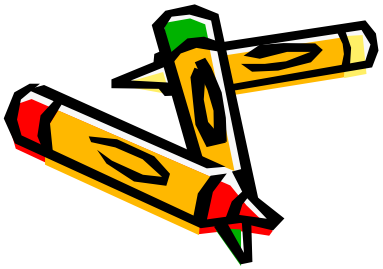
$$\left(\frac{e^{2z_1} - 1}{e^{2z_1} + 1}, \frac{e^{2z_2} - 1}{e^{2z_2} + 1} \right) .$$

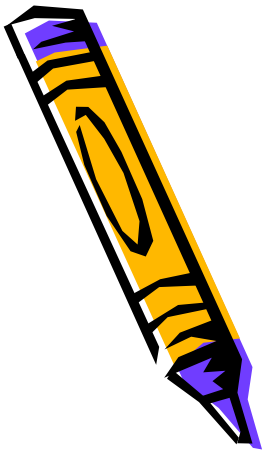


Coeficientul de corelație *r* (Pearson) ia valori cuprinse între -1 și +1, trecând deci și prin valoarea 0 care indică O asociație neliniară între cele două variabile (independență liniară):

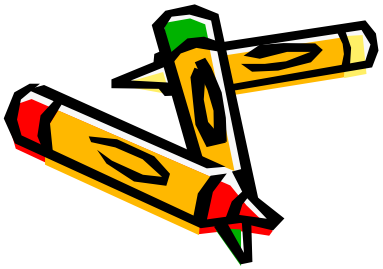


- o valoare a lui *r* apropiată de -1 indică o corelație negativă puternică, adică tendința unei variabile de a scădea puternic când cealaltă variabilă crește;
- o valoare a lui *r* apropiată de +1 indică o corelație pozitivă puternică, adică tendința de creștere puternică a unei variabile atunci când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





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



În acest sens indicăm

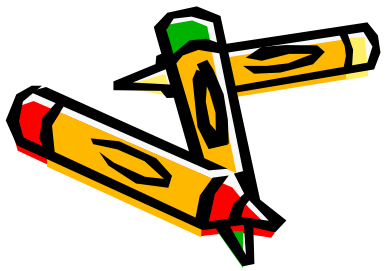
- fie un prag definit de inegalitatea

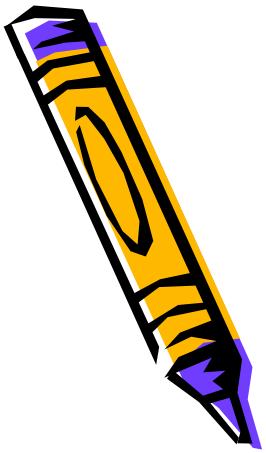
$$|r| \cdot \sqrt{n-1} \geq 3,$$

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

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

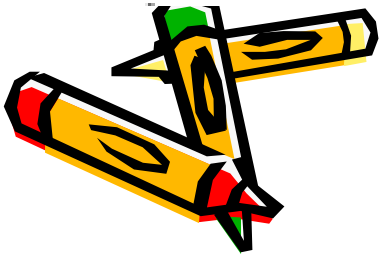
„dacă p < 0.05 *atunci legătura este semnificativă".*

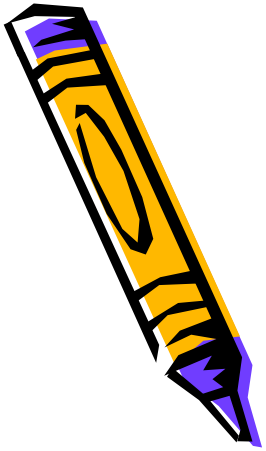




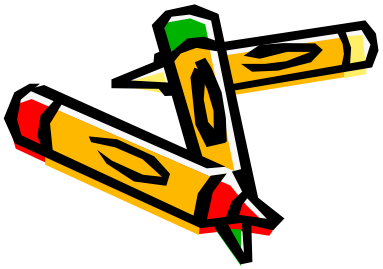
Un coeficient de corelație important nu implică totdeauna în mod necesar o legătură naturală, intrinsecă, între caracteristicile ce definesc cel două variabile statistice analizate.

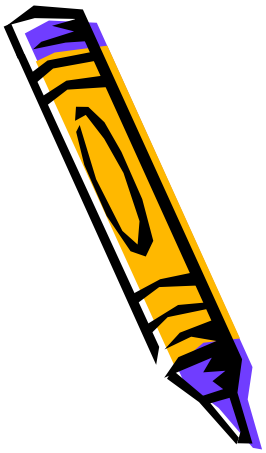
De exemplu, în medicină, aceiași valoare redusă a coeficientului de corelație poate fi importantă în epidemiologie dar nesemnificativă din punct de vedere clinic





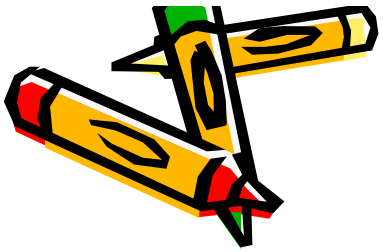
În concluzie, coeficientul de corelație este o măsură a legăturii liniare, „aritmetice", dintre cele două variabile, care poate fi câteodată și întâmplătoare, fără relevanță reală.

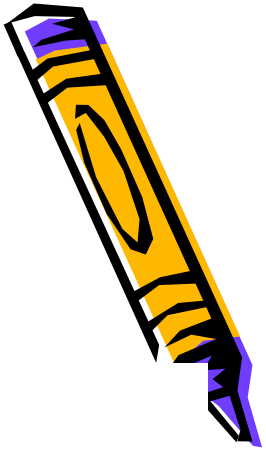




Presupunând că legătura dintre cele două variabile, reliefată de coeficientul de corelație, nu este întâmplătoare, există trei posibile explicații:

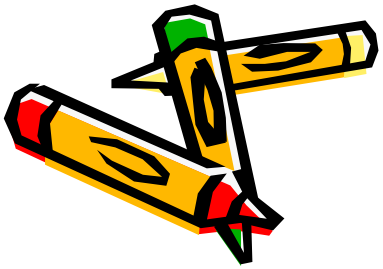
- Variabila *X influențează* (*cauzează*) variabila *Y*;
- Variabila *Y influențează* (*cauzează*) variabila *X*;
- Ambele variabile *X* și *Y sunt influențate* de același fenomen din fundal, diferit de ele.





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

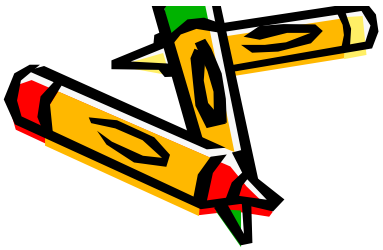
- se prezintă mai întâi diagrama de împrăștiere a norului de puncte;
- când se prezintă coeficientul de corelație *r*, valoarea sa trebuie să aibă două zecimale și să fie însoțită de nivelul de semnificație *p* și de intervalul de încredere corespunzător, dacă este posibil.
- trebuie menționat și numărul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

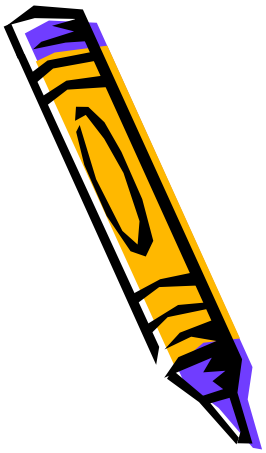


Covarianța poate fi privită ca „momentul" corelației
și, amintindu-ne și de formula sa probabilistă dată de:

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

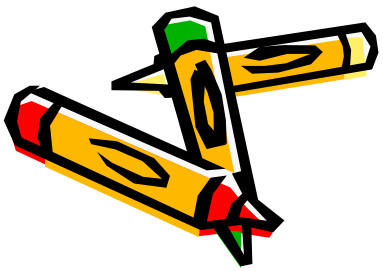
observăm că ea este nulă dacă variabilele care generează
cele două serii statistice sunt (liniar) independente.





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

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



regresia liniara

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

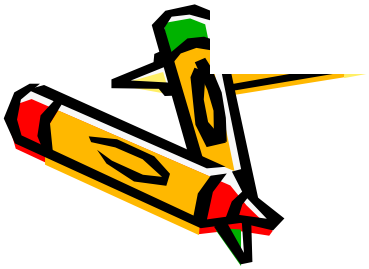
În acest scop, se consideră una dintre variabile ca *variabilă independentă* sau *variabilă predictor*, iar cealaltă variabilă ca *variabilă dependentă* sau *variabilă răspuns*.

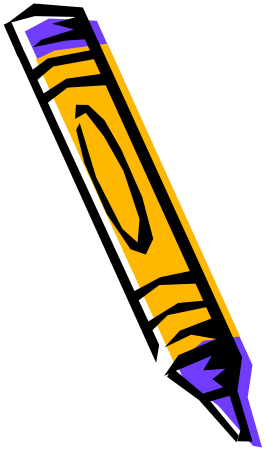
Legătura liniară dintre cele două variabile este descrisă de o ecuație liniară, așa-numita *ecuația de regresie*.

regresia multipla

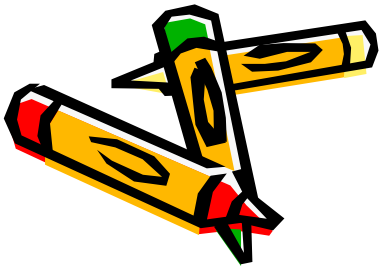
Cazul cel mai general se referă la situația în care dispunem, pe de o parte, de *n* variabile predictive sau explicative și, pe de altă parte, de o variabilă explicată, numită și *răspuns* sau *efect*, pe care vrem să o deducem, cunoscând valorile celorlalte variabile.

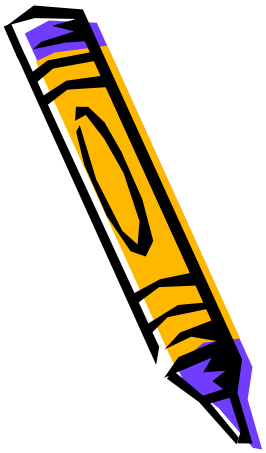
Acesta este cazul regresiei multiple.





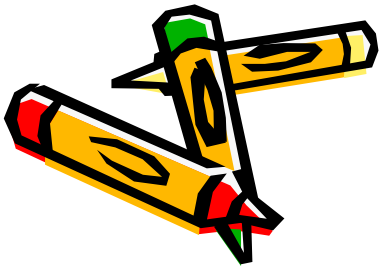
Regresia liniară (simplă sau multiplă), ce va fi prezentată în partea dedicată prognozei, procesează variabilele modelului în ipoteza că variabila dependentă (prognozată) este o variabilă numerică continuă.



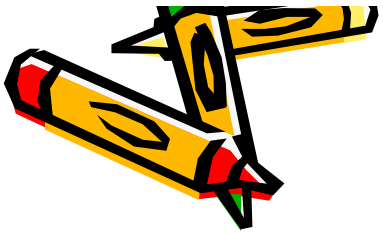


Ce se întâmplă însă în cazul în care una din variabile este categorială?

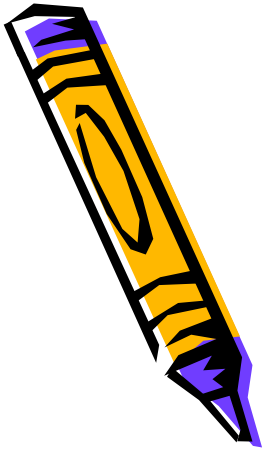
În multe studii, variabila dependentă (variabila răspuns), care trebuie dedusă din variabilele explicative, este o variabilă categorială.



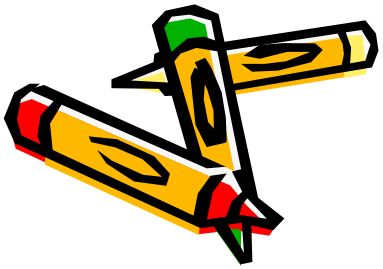
De exemplu, în medicină, această variabilă poate reprezenta diagnosticarea unei boli (DA sau NU) a unui individ, pe baza unor date prelevate de la acesta. Astfel, în cursul analizei epidemiologice în studiul cancerului, se colectează date privind un anumit număr de variabile care ar putea influența riscul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

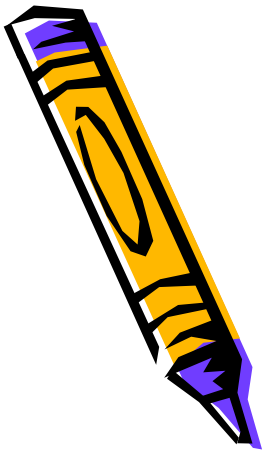


exemple

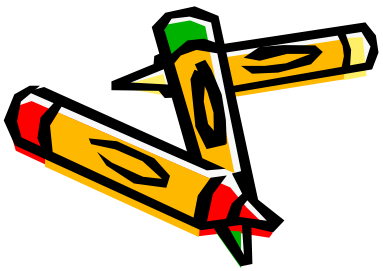


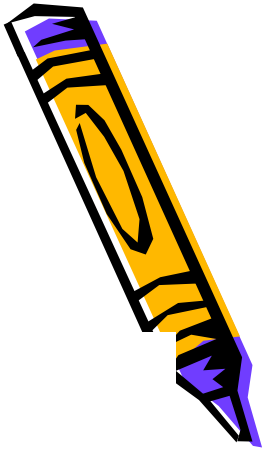
Dacă vom considera cazul cancerului pulmonar, ne propunem să estimăm riscul de îmbolnăvire a unui subiect de sex masculin, în vârstă de 58 ani, care a muncit timp de 40 ani ca muncitor siderurgist și care a fumat în medie 20 țigări pe zi, începând cu vârsta de 16 ani.



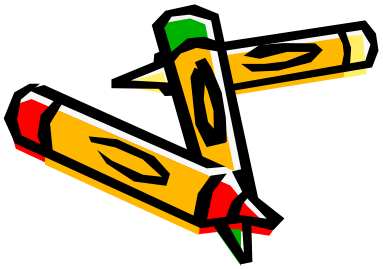


Altă situație se referă la încadrarea în categoria „fraudă" a unor tranzacții comerciale, pe baza anumitor caracteristici ale lor.



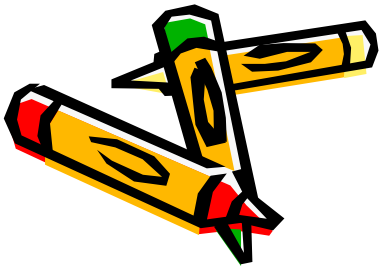


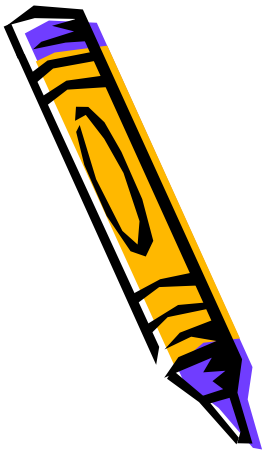
Spre deosebire de cazul modelelor regresive liniare multiple, în care valorile ce trebuie estimate sunt numere, în această situație în care variabila dependentă este reprezentată de o variabilă categorială (posibilitatea de a face cancer, sau de a produce o fraudă), având valorile: DA sau NU, problema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



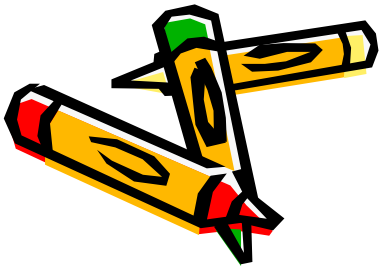
regresia logistica

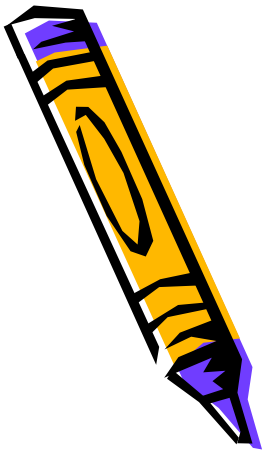
Plecând de la aceeași filosofie ca și în cazul regresiei multiple clasice, vom construi un model ușor diferit, cunoscut sub numele de *regresie logistică liniară multiplă*, pe scurt *regresie logistică*.



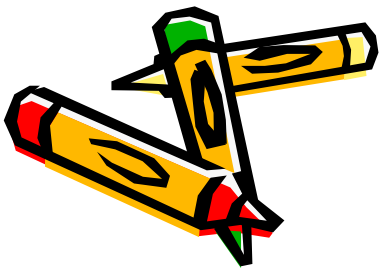


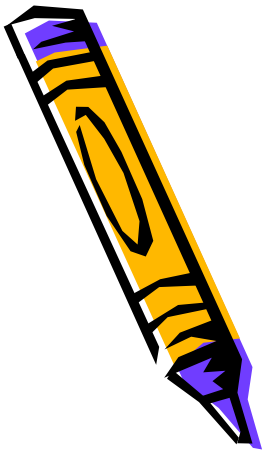
Spre deosebire de regresia clasică, aici una sau mai multe variabile predictive/explicative (variabile independente) pot fi categoriale, obligatoriu variabila dependentă, deci în cazul acestui nou model este vorba de caracterul nenumeric, calitativ (categorial) al unora dintre variabilele sale.



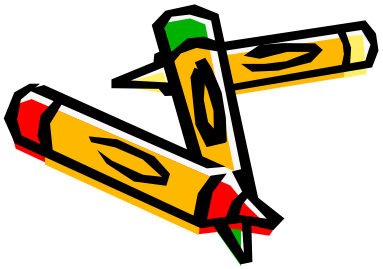


Principiul de bază rămâne acelaşi ca şi la regresia multiplă, diferenţa constând în faptul că dacă în primul caz estimăm valoarea variabilei dependente pe baza valorilor predictorilor, în acest caz estimăm o *transformare* a variabilei dependente.

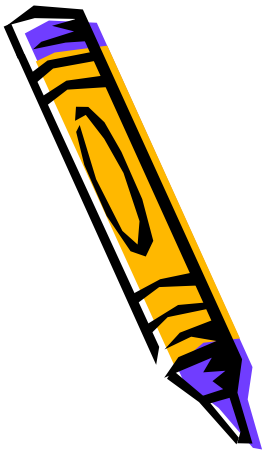




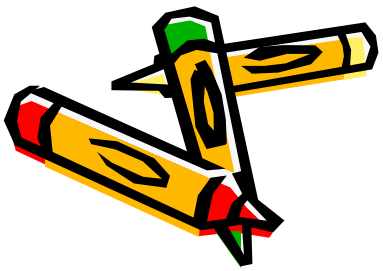
Astfel, dacă variabila dependentă are ca valori binare afirmațiile DA și NU, spre exemplu îmbolnăvire sau nu, răspuns pozitiv sau negativ la un tratament, fraudă sau nu etc., care codate au valorile 1 și 0, atunci media acesteia reprezintă proporția indivizilor din populație cu caracteristica respectivă.



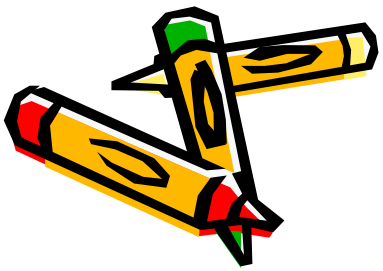
exemplu

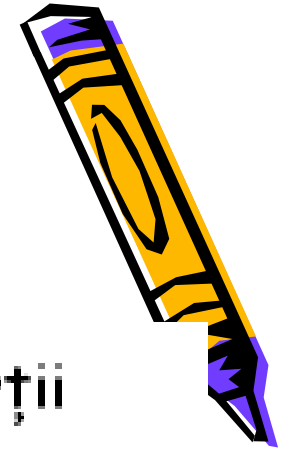


să presupunem că avem pentru 100 de indivizi,
65 răspunsuri DA şi 35 răspunsuri NU, adică 65 de 1
şi 35 de 0.
Media acestor valori va fi $(65 * 1 + 35 * 0)/100 = 65/100$
adică 65% de DA.



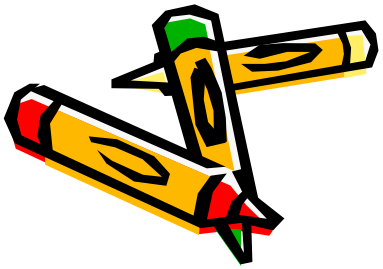
În general, va rezulta că modelul regresiv logistic va estima proporția subiecților/obiectelor care vor avea aceeași caracteristică de interes pentru studiul statistic, sau echivalent, probabilitatea ca un subiect/obiect oarecare din populație să aibă o anumită caracteristică, de exemplu cancer pulmonar,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

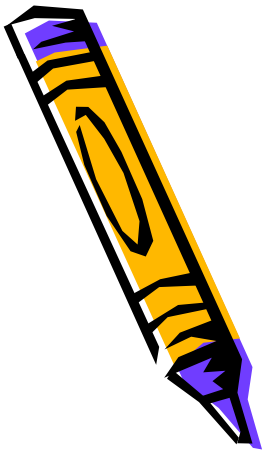




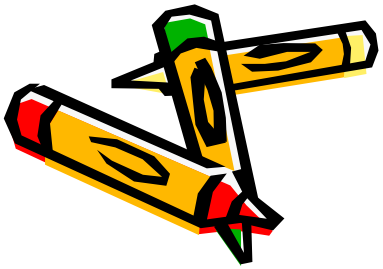
Din punct de vedere practic, în locul acestei proporții se ia în considerație o transformare a ei.

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

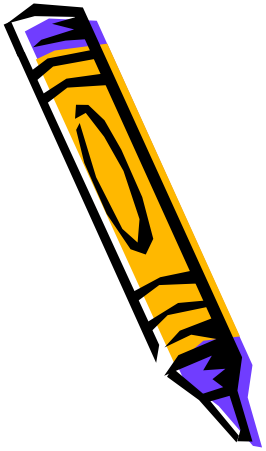




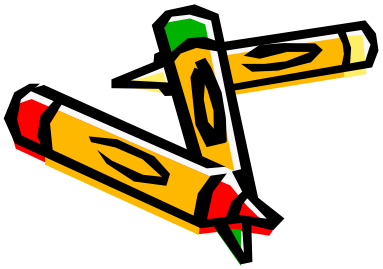
Cu alte cuvinte, se procedează la o ‚normare' a intervalului valorilor produse de combinația predictorilor, adică o transformare a acestuia în intervalul $[0, 1]$, și astfel valorile variabilei răspuns pot fi considerate probabilități.

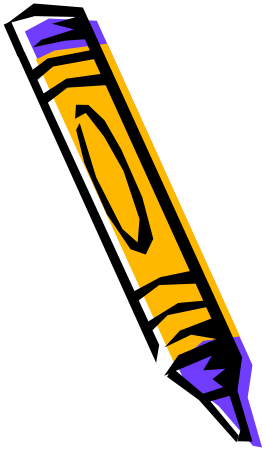


logit

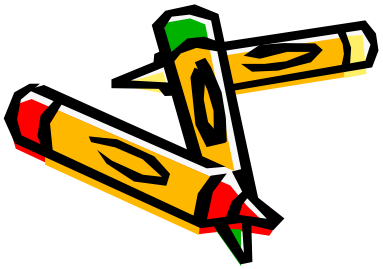


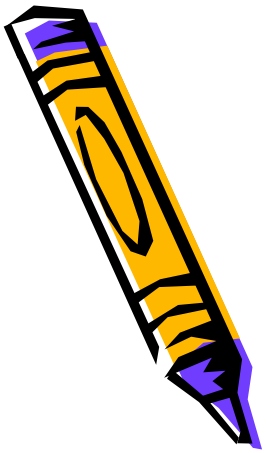
Metoda regresiei logistice constă în folosirea transformării *logit* scrisă ca *logit*(*p*).
Concret, aici *p* reprezintă probabilitatea ca un individ/obiect oarecare să aibă caracteristica de interes cerută şi deci (1 - *p*) va reprezenta probabilitatea ca acesta să nu o aibă.





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

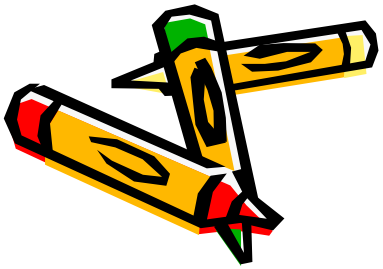


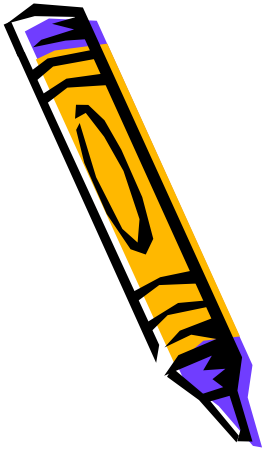


Raportul $\frac{p}{1-p}$ se numeşte *şansă*,

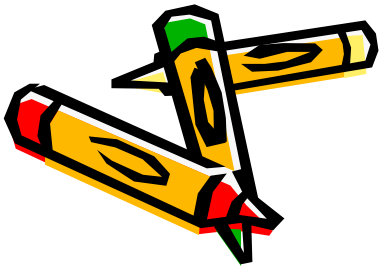
iar transformarea $logit(p) = \ln \frac{p}{1-p}$

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

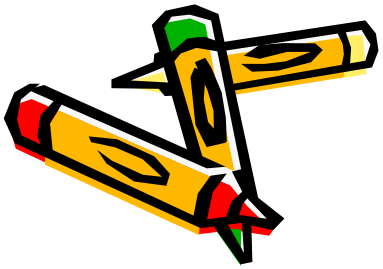


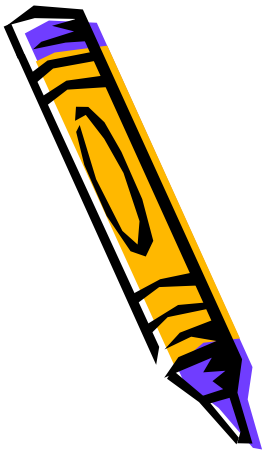


Această mărime este importantă şi deoarece
ajută la compararea predicțiilor între subiecții/obiectele
care au sau nu o anumită caracteristică,
de exemplu pentru subiecții dintr-un lot cu vârsta mai mică
de 40 ani față de cei peste 40 ani.



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





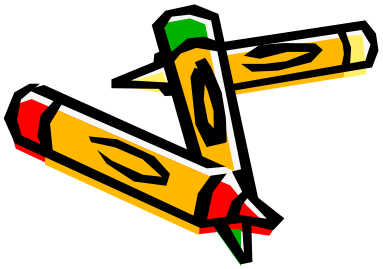
Dacă vom considera:

$l_1 = logit(p_1)$

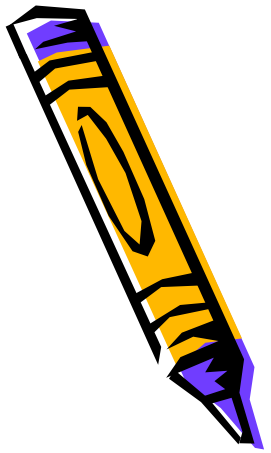
log-ul şansei pentru primul grup şi

$l_2 = logit(p_2)$

log-ul şansei pentru al doilea grup, atunci:

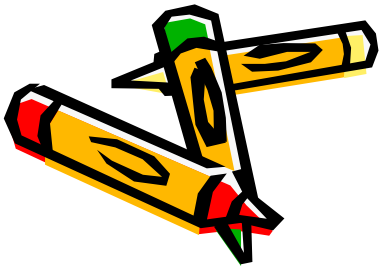


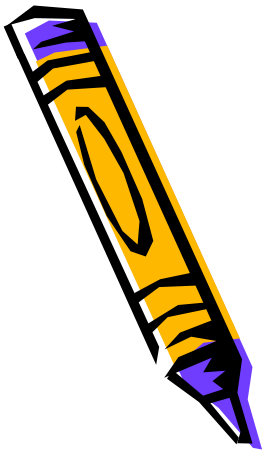
logaritmul riscului relativ



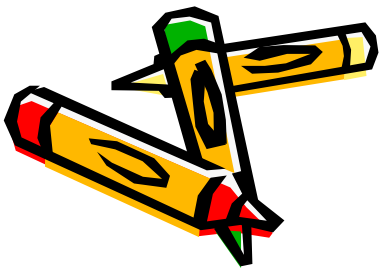
$$l_1 - l_2 = logit(p_1) - logit(p_1) = \ln \frac{p_1 \cdot (1 - p_2)}{p_2 \cdot (1 - p_1)}$$

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





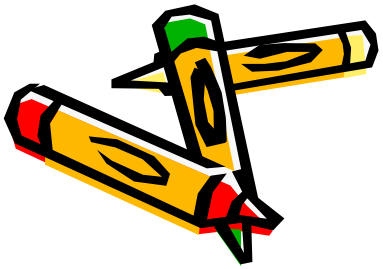
Acesta este folosit uzual de exemplu în studiile epidemiologice, pentru a stabili conexiunea între maladie şi expunere.



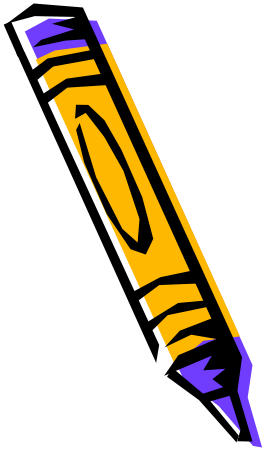
În concluzie, odată găsit *logit*(*p*), pe baza predictorilor, dat de ecuația de regresie logistică:

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

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

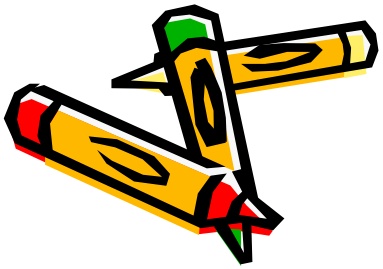


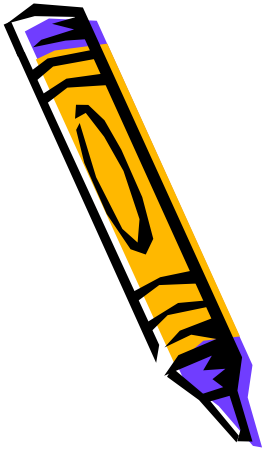
exemplu



Pentru stabilirea legăturilor existente între apariția hipertensiunii 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.

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

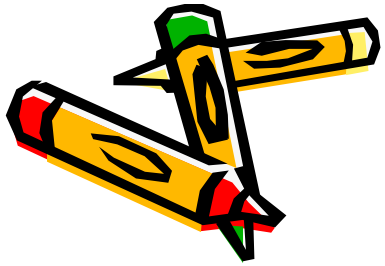




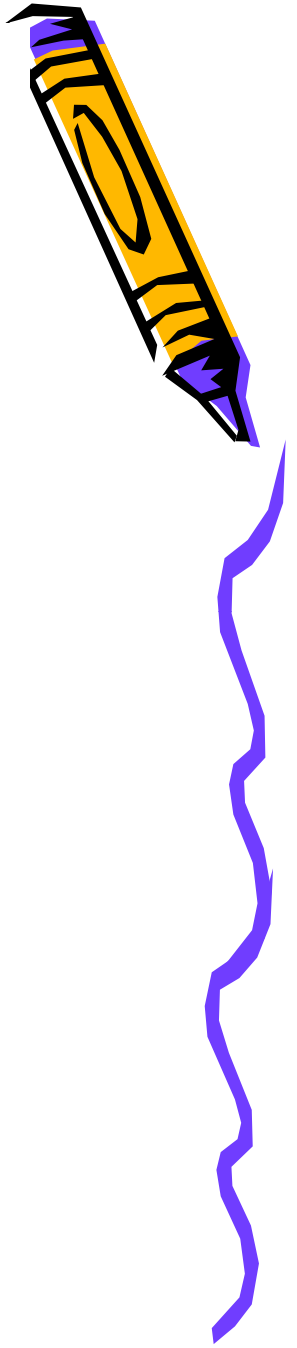
fumat	obezitate	vârstă	număr subiecți	număr subiecți cu hipertensiune (%)
0	0	0	60	5(8%)
1	0	0	17	2(11%)
0	1	0	8	1(13%)
0	0	1	187	35(19%)
1	1	0	2	0(0%)
1	0	1	60	5(8%)
0	1	1	51	15(29%)
1	1	1	23	8(35%)
			Total=433	total=79 (18%)

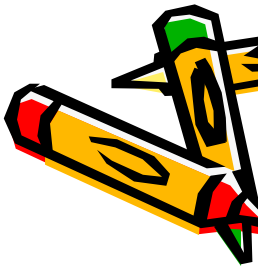
Plecând de la acest tabel general, vom construi un tabel de lucru, care rezumă, sintetic, toate posibilitățile existente.

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



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





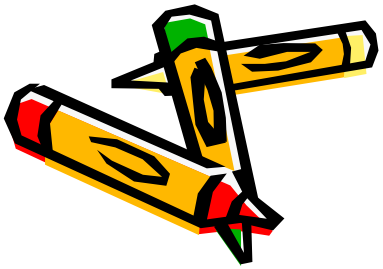
În cazul a 3 variabile explicative avem 16 combinații posibile în total, iar dacă vom considera 4 variabile explicative, vom avea 64 de cazuri posibile în total.
Toate variabilele explicative au fost codate (0 = Nu și 1 = DA).
Calculul nu se poate face ‚manual' din cauza volumului mare de muncă.

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

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

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

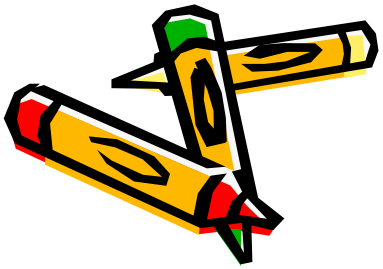
Am obținut şi o clasificare a importanței variabilelor explicative asupra variabilei răspuns (ierarhizarea acestora), ierarhie obținută pe baza nivelului de semnificație *p*.



Ecuația de regresie logistică corespunzătoare este dată de:

$$logit(p) = -2.378 - 0.068 \cdot fumat + 0.695 \cdot obezitate +$$

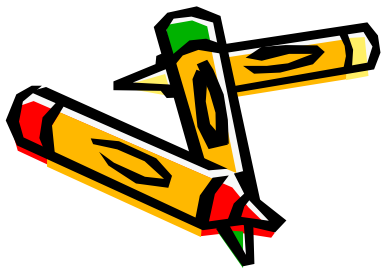
$$+ 0.872 \cdot varsta$$



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

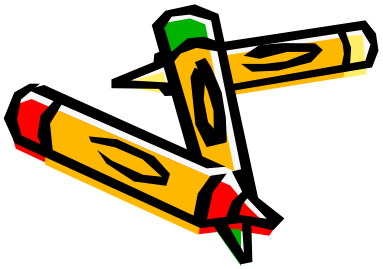
Se consideră valorile concrete ale acelui individ privind cei trei factori de risc enumerați mai sus, care se vor introduce în ecuația de regresie.

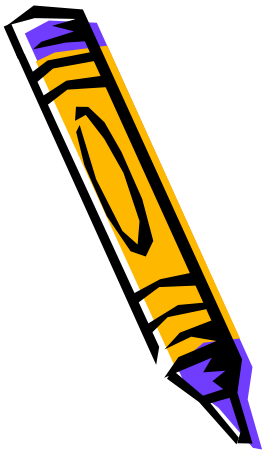
Va rezulta valoarea *logit*(*p*) corespunzătoare, care va estima riscul ca individul respectiv să fie sau nu predispus la hipertensiune.



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

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

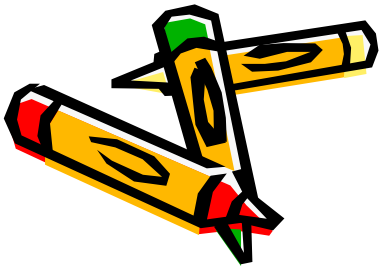




$$logit(p_{varsta>40}) = -2.378 - 0.068 \cdot fumat + 0.695 \cdot obezitate +$$

$$+ 0.872$$

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



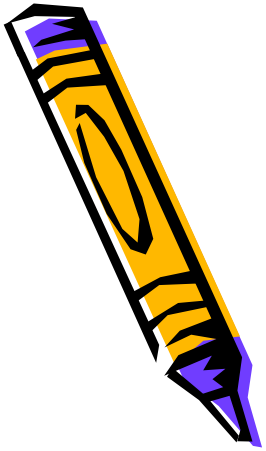
Se obține astfel că:

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

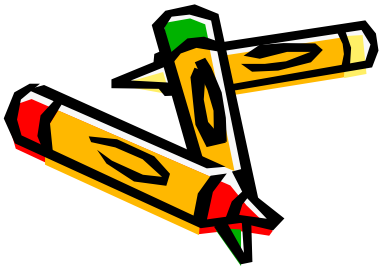
Rezultă că raportul şanselor hipertensiunii asociat cu nivelul de *vârstă* considerat mai sus (de 40 ani) este $e^{0.872} = 2.3917$, valoare care poate fi interpretată astfel:

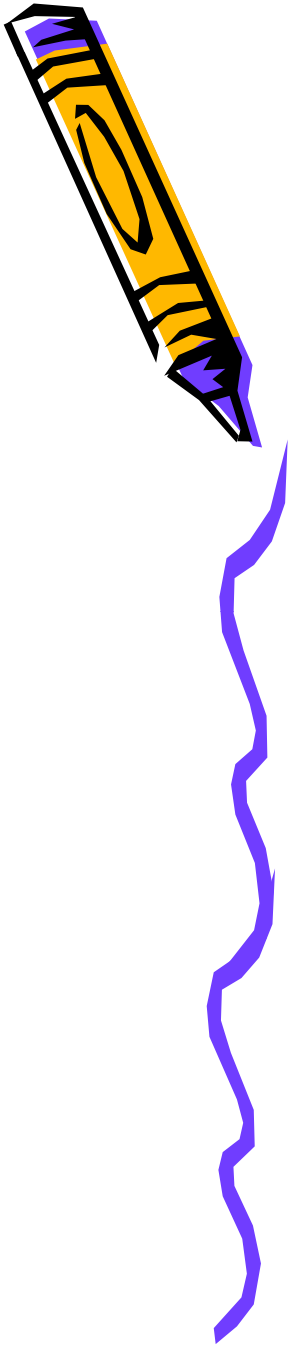
„riscul de a face hipertensiune peste vârsta de 40 de ani este de $2,39$ de ori mai mare decât sub această vârstă."

exemplu

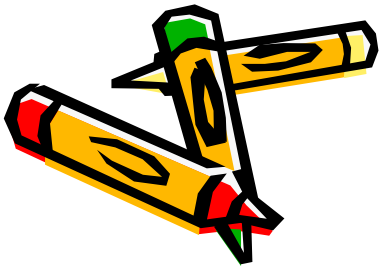


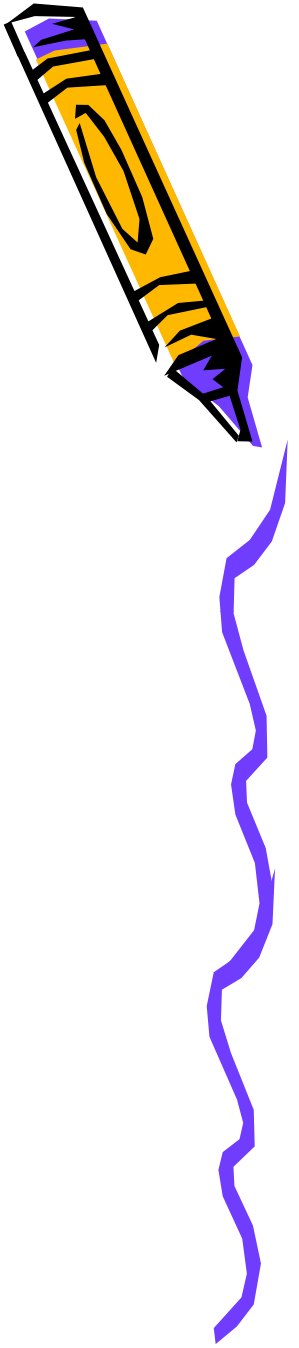
Studiul lui Tuyns et al., 1977 în departamentul Ille-et-Vilaine din Bretania este un studiu caz-control cuprinzând în lotul caz 200 bărbați diagnosticați cu cancer esofagian iar în lotul control un număr de 775 adulți aleși aleator din aceiași localitate.
În acest studiu au fost aleși ca factori de risc consumul de alcool și de tutun măsurate în g/zi.



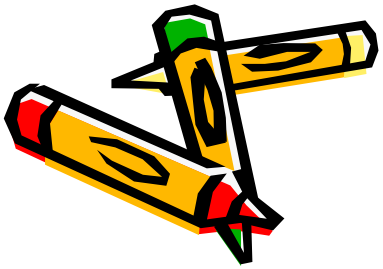


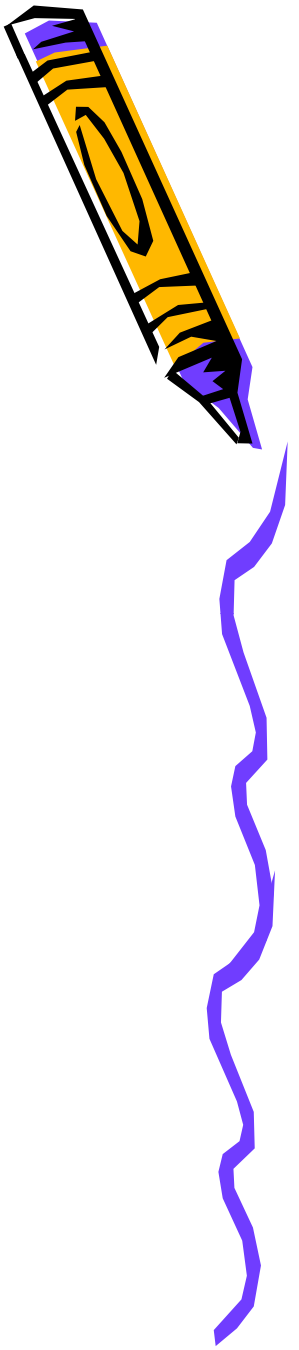
	caz	control
Vârstă (ani)		
25-34	1	115
35-44	9	190
45-54	46	167
55-64	76	166
65-74	55	106
75 +	13	31
medie	60,0	50,2
dev. Std	9.2	14.3



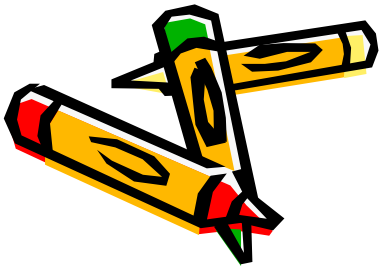


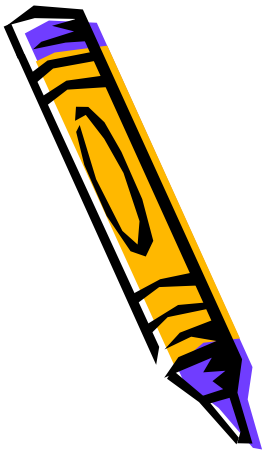
	caz	control
Alcool (g / zi)		
0-39	29	386
40-79	75	280
80-119	51	87
120 +	45	22
medie	84.9	44.4
dev. std.	48.4	31.9





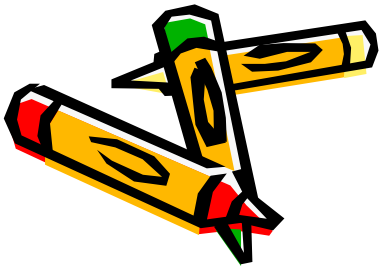
	caz	control
Tutun (g / zi)		
0-9	78	447
10-19	58	178
20-29	33	99
30 +	31	51
medie	16.7	10.5
dev. std.	12.9	11.9



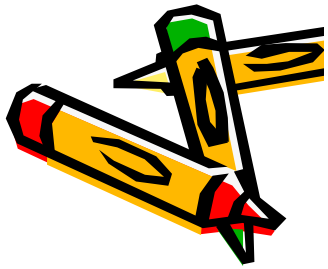


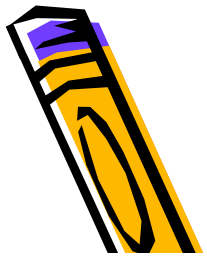
Avem şase categorii de vârstă iar factorul de risc principal este consumul de alcool.

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



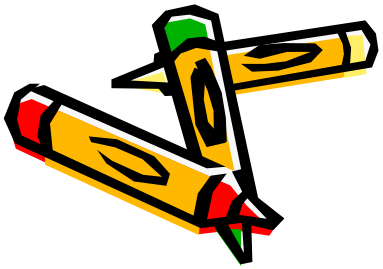
Strat vârstă	expunere	caz	Total(caz + control)
1	1	1	10
1	0	0	106
2	1	4	30
2	0	5	169
3	1	25	54
3	0	21	159
4	1	42	69
4	0	34	173
5	1	19	37
5	0	36	124
6	1	5	5
6	0	8	39



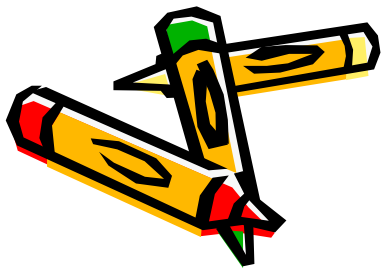


În urma procesării computerizate (STATISTICA),
s-a obținut riscul relativ (raportul șanselor) corespunzător consumului de alcool egal cu 5.31 iar intervalul corespunzător de încredere de nivel 95% este (3.66, 7.71),
de unde tragem următoarea concluzie:

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



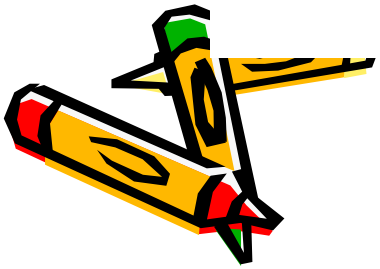
Se pot face analize regresive mai complexe, care să țină seama și de acțiunea concertată a doi sau mai mulți factori de risc.

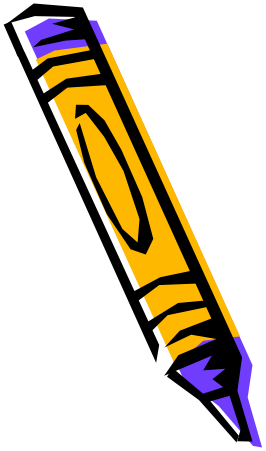


analiza discriminant

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

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



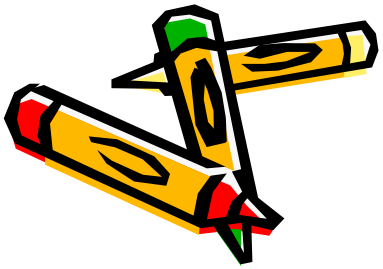


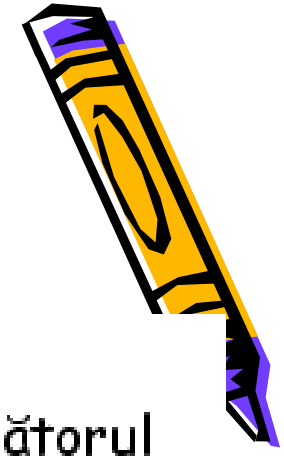
Putem utiliza modelul regresiv logistic pentru calcularea *indicelui pronostic* (sau *index diagnostic* în medicină).

Definind *L* ca fiind log-ul şansei *p*, rezultă că:

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

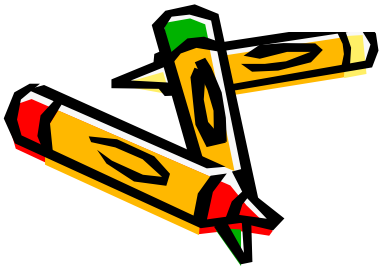
unde avem *k* variabile explicative în model.



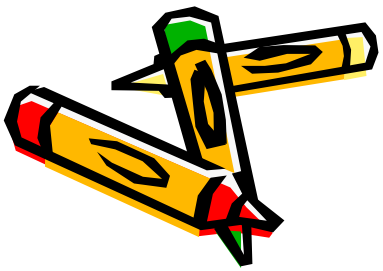


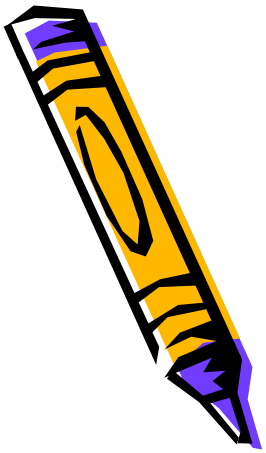
Astfel, în cazul primului exemplu, considerând doar factorii predictivi *obezitatea* şi *vârsta*, obţinem următorul tabel cu valorile lui *L*, precum şi proporţiile corespunzătoare de pacienţi.

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



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





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

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

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