Statistică multivariată

Lucrarea nr. 10 — Regresia logistică - SPSS

A. Noțiuni teoretice

Regresia logistică

Regresia logistică modelează relația dintre o mulțime de variabile independente x_i (categoriale, continue) şi o variabilă dependentă dihotomică (nominală, binară) Y. O astfel de variabilă dependentă apare, de regulă, atunci când reprezintă apartenența la două clase, categorii – prezență/absență, da/nu etc.

Ecuația de regresie obținută, de un tip diferit de celelalte regresii discutate, oferă informații despre:

- importanța variabilelor în diferențierea claselor,
- clasificarea unei observații într-o clasă.

De remarcat că diagrama de împrăştiere a valorilor nu oferă nici un indiciu în privinta dependențelor. În asemenea cazuri, regresia liniară clasică nu oferă un model adecvat.

Presupunem că valorile *y* (variabilă binară) sunt codificate 0/1, valoarea 1 exprimând în general apariția unui anumit eveniment, astfel încât ceea ce se caută este o estimare a probabilității de producere a respectivului eveniment în funcție de valorile variabilelor independente.

Cazul unei singure variabile independente

Modelul este

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

sau

$$\ln\left(\frac{P(y=1\,|\,x)}{1-P(y=1\,|\,x)}\right)=\alpha+\beta\,x$$

Cantitatea din partea stângă este numită (***transformarea***) **logit** a probabilității $P(y=1|x)$.

Semnificația expresiei $P(y=1|x)$ este evidentă: probabilitatea de realizare a valorii $y=1$ condiționată de valoarea x. Cu alte cuvinte, probabilitatea de clasare a observației x în clasa $y=1$, sau probabilitatea ca valoarea x să fie asociată cu producerea evenimentului $y=1$. In continuare se notează $P(y=1|x)$ cu p, conform notației de la modelul probabilist binomial (probabilitatea de "succes").

Transformarea logit este necesară pentru a proiecta probabilitatea p din intervalul (0,1) în intervalul $(-\infty, +\infty)$, fapt necesar în procesul de estimare a parametrilor. Modelul este legat direct de noțiunea de *odds* (raport de şanse), notat OR (*odds report*):

$$\mathrm{OR}=\frac{p}{1-p}$$

care reprezintă raportul dintre probabilitatea de « succes » şi probabilitatea de « insucces ».

Modelul se mai poate scrie

$$\frac{p}{1-p} = e^{\alpha+\beta x}$$

de unde interpretarea coeficientului β:

- creşterea cantităţii logit atunci când *x* creşte cu o unitate sau
- OR creşte de e^{β} ori atunci când *x* creşte cu o unitate.

Testarea ipotezei β = 0 se realizează prin testul Wald, corespunzător testului t de la regresia liniară, statistica testului fiind

$$\chi 2 = \frac{b^2}{\text{Var}\,(b)}$$

care este repartizată χ^2 cu un singur grad de libertate.

Intervalul de încredere pentru β este, potrivit rezultatelor de la analiza ecuaţiei de regresie,

$$\left(e^{\,b - z_{1-\frac{\alpha}{2}} SE(b)}, e^{\,b + z_{1-\frac{\alpha}{2}} SE(b)} \right),$$

unde *b* este estimaţia lui β (din ecuaţia de regresie estimată) iar SE(*b*) este abaterea standard a repartiţiei de sondaj a lui *b*.

Se observă imediat că, pentru o observaţie, 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.

Cazul mai multor variabile independente

Modelul general este

$$\ln\left(\frac{p}{1\text{-}p}\right) = \beta_0 + \beta_1 x_1 \; + \; \beta_2 x_2 \; + \; \ldots \beta_k x_k \,,$$

unde *p* este $P(y = 1 \mid x_1, x_2, \ldots, x_k)$. Se poate obţine imediat şi forma exponenţială echivalentă.

Interpretarea coeficienţilor β_i este evidentă: creşterea cantităţii logit (logaritm din OR) atunci când x_i creşte cu o unitate (celelalte variabile *x* rămânând constante). Pentru interpretări mai sofisticate rescriem modelul sub forma:

$$P(y = 1 \mid x_1, x_2, \ldots, x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 \; + \; \beta_2 x_2 \; + \; \ldots \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 \; + \; \beta_2 x_2 \; + \; \ldots \beta_k x_k)}$$

Se obţine atunci, după calcule imediate,

$$\exp(\beta_0) = \frac{P(y = 1 \mid x_1 = x_2 = \ldots = x_k = 0)}{1 - P(y = 1 \mid x_1 = x_2 = \ldots = x_k = 0)} =$$

$$= \frac{P(y = 1 \mid x_1 = x_2 = \ldots = x_k = 0)}{P(y = 0 \mid x_1 = x_2 = \ldots = x_k = 0)}$$

adică OR în situaţia de bază $x_1 = x_2 = \ldots = x_k = 0$.

Pentru coeficientul β_i se obţine :

$$\exp(\beta_i) = \frac{P(y=1 \mid x_i = 1,\ x_j = 0 \ \text{pentru} \ \ j \neq i)}{1 - P(y=1 \mid x_i = 1,\ x_j = 0 \ \text{pentru} \ \ j \neq i)} \times \frac{1}{\text{OR}_{\text{baza}}} =$$

$$= \frac{\text{OR}_{x_i=1,\ x_j=0 \ \text{pentru} \ \ j \neq i}}{\text{OR}_{\text{baza}}}.$$

Se ajunge astfel, din caracterul multiplicativ al modelului logistic,

$$\text{OR}_{x_1, x_2, \ldots, x_k} = \exp(\beta_0) \times \exp(\beta_1 x_1) \times \ldots \times \exp(\beta_k x_k),$$

la interpretarea utilă că fiecare β_i exprimă contribuția factorului x_i la explicarea probabilității (sub forma OR) de producere a evenimentului $y = 1$. Astfel, fixând $x_i = 1$, $\exp(\beta_i)$ va reprezenta factorul multiplicativ constant indiferent de valorile celorlalte variabile independente.

Dacă $\beta_i = 0$, factorul corespunzător nu are nici un efect, (înmulțirea cu 1). Dacă $\beta_i < 0$ prezența factorului reduce probabilitatea evenimentului $y = 1$, $\beta_i > 0$ mărind această probabilitate.

Construirea modelului se poate realiza şi prin metode forward sau backward, testarea semnificației coeficienților realizându-se prin testul Wald sau prin testul raportului de verosimilitate (LR, *likelihood-ratio*).

Testul Wald este prezentat la modelul logistic cu un singur factor. Testul LR se bazează pe statistica obținută ca raport între maximul funcției de verosimilitate sub ipoteza nulă şi maximul funcției de verosimilitate în condiții mai largi. Lema Neyman-Pearson arată că acesta este cel mai puternic test la un prag α fixat. Pentru cazul regresiei logistice, se calculează raportul între valoarea maximă a funcției de verosimilitate pentru modelul complet (L_1) şi cea pentru modelul mai simplu (L_0). Statistica LR este $-2\log(L_0/L_1)$, repartizată χ^2. Testul LR este recomandat în cazul construirii modelului pas cu pas, verificând dacă variabila eliminată din model este semnificativă, deci dacă modelul poate fi simplificat.

Observație. O mai bună imagine intuitivă asupra raportului de verosimilitate este dată în continuare. presupunem că se doreşte distingerea între două ipoteze H_0 şi H_1 (o contrară a lui H_0). Fie p_0 probabilitatea ca datele observate să apară în ipoteza H_0 adevărată şi p_1 probabilitatea ca datele observate să apară în ipoteza H_1 adevărată. Raportul p_1/ p_0 este raportul de verosimilitate (LR) şi măsoară OR (odds report) ca H_1 să fie adevărat ca opusă lui H_0 adevărată.

Deoarece unele simulări arată că datorită datelor "rare" (*sparse*) statistica prin care se compară două modele nu este repartizată χ^2 şi, din acest motiv, s-a dezvoltat testul Hosmer-Lemeshow. De notat că testul este recomandat pentru variabile independente continue şi mai m ult de 400 de observații. Testul constă în clasificarea în decile a probabilităților prognozate (10 grupuri bazate pe rangul percentilic) şi calcularea statisticii χ^2 care compară frecvențele observate cu cele prognozate (în tabelul 2×10). Valori mici ale statisticii (deci acceptarea nediferențierii dintre cele două şiruri de frecvențe) arată o bună potrivire a datelor prognozate, deci o adecvanța modelului.

În regresia logistică nu există un indicator absolut similar coeficientului R^2 din regresia liniară. S-au dezvoltat însă indicatori similari. Astfel în SPSS există Cox & Snell Pseudo-R^2 definit prin

$$R^2 = 1 - \left[\frac{-2LL_{null}}{-2LL_k} \right]^{2/n}$$

unde LL_{null} este logaritm din maximul funcției de verosimilitate pentru modelul constant, iar LL_{k} este logaritm din maximul funcției de verosimilitate pentru modelul cu variabile independente incluse. Se poate astfel observa că se merge pe varianta de comparare a cantităților -2*LL* prin intermediul raportului lor şi nu a împărțirii lor (ca la LR). Acest R^2 nu atinge 1 şi a fost introdusă de Nagelkerke o modificare prin care se atinge 1. Formula pentru Nagelkerke Pseudo-R^2 este

$$R^2 = \frac{1 - \left[\frac{-2LL_{null}}{-2LL_k}\right]^{2/n}}{1 - \left(-2LL_{null}\right)^{2/n}}$$

Alți indicatori sunt:

- AIC (*Akaike's Information Criterion*) definit ca $-2LL_k + 2k$, unde k este numărul de parametri estimați.
- BIC (*Bayesian Information Criterion*) definit ca $-2LL_k + k*\log(n)$ unde k este numărul de parametri estimați iar n este numărul de observații. BIC mai este referit şi drept criteriul Schwartz (care l-a argumentat).

Vor fi preferate modelele pentru care criteriile (AIC sau BIC) au valori mai mici. Se observă că ambele criterii "recompensează" buna potrivire a modelului dar şi "penalizează" numărul de parametri estimați, astfel încât să se obțină un model bun dar cu un număr minim de parametri. În BIC, penalizarea lui k este mai puternică decât în AIC. Ambii indicatori necesită condiția ca erorile (reziduurile) să fie normal distribuite.

Regresia logistică multinomială

Modelul regresional logistic multinomial (cunoscut şi ca regresia logistică politomică – *polytomous logistic regression* – sau ca model de alegere discretă – *discrete choice model* – în econometrie) este o generalizare a modelului logistic acceptând ca variabila dependentă Y să aibă mai mult de două valori.

Să presupunem că variabila Y are ca valori posibile elementele mulțimii neordonate {1, ... , g}. Modelul logistic multinomial presupune că probabilitatea ca Y să fie egal cu s în observația i depinde de valorile variabilelor x_{i1}, ... , x_{ip} prin

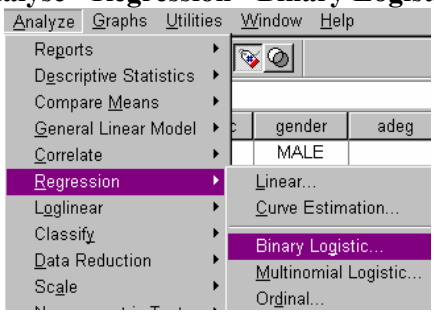
$$P(Y_i = s) = \frac{e^{\eta_{is}}}{\sum_{t=1}^{g} e^{\eta_{it}}}$$

unde $\eta_{is} = \sum_{k=1}^{p} x_{ik}\beta_{ks}$ este o funcție liniară. În această formulare a modelului, este de remarcat că există coeficienți de regresie β_{ks} diferiți pentru fiecare k şi, mai ales, s. Prin urmare, fiecare valoare posibilă Y are un model asociat.

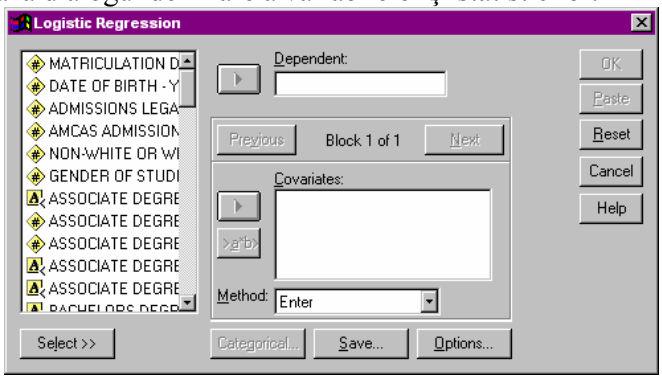
Modelul astfel definit este supraparametrizat, ceea ce impune o reducere prin fixarea unei valori Y, de exemplu Y = 1, drept categorie de referință (adică β_{11}, ... , β_{p1} sunt egali cu zero). Alegerea categoriei de referință poate facilita interpretarea.

B. Instrumente SPSS

Comanda este **Analyse - Regression - Binary Logistic**.

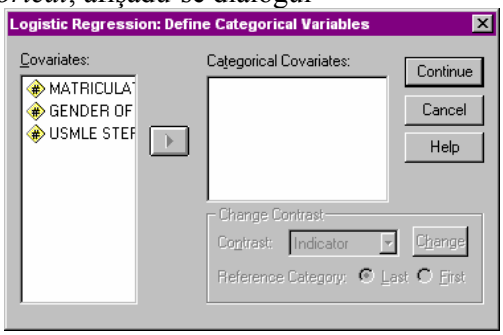


Se afişează dialogul de fixare a variabilelor şi statisticilor.

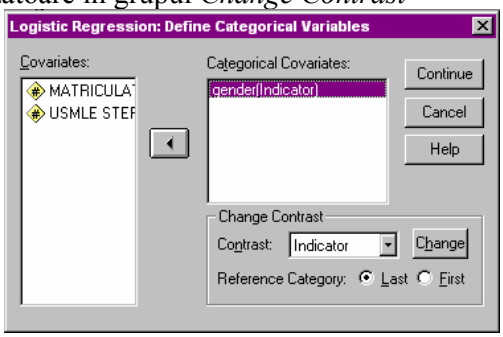


Se mută variabila dependentă (binară) în *Dependent*. Variabila independentă sau variabilele independente (în cazul multivariat) sunt mutate în lista *Covariates*.

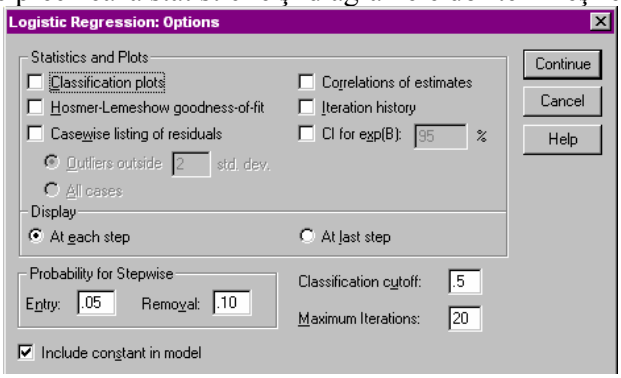
Pentru a indica variabilele independente care este categoriale (discrete), se va acționa butonul *Categorical*, afişâdu-se dialogul



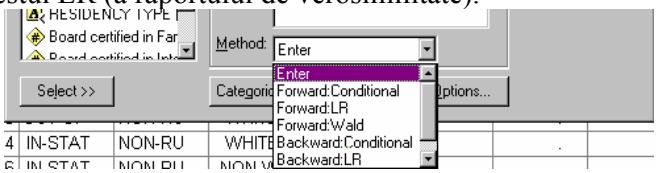
Fiecare variabilă trecută în lista *Categorical Covariates* poate fi caracterizată prin selecții corespunzătoare în grupul *Change Contrast*



Acționând butonul *Options* din dialogul principal, se deschide dialogul sinonim în care se precizează statisticile şi diagramele dorite în ieşire.



În dialogul principal *Logistic Regression*, se poate alege metoda utilizată pentru introducerea variabilelor la estimarea regresiei. De reţinut metoda *Enter* în care variabilele sunt introduse în bloc (se estimează o singură ecuaţie) sau metode de selectare pas cu pas (ca la regresia liniară multiplă), cum ar fi *Forward: LR*. Aceasta înseamnă că modelul este construit ascendent, criteriul de introducere a unei noi variabile fiind testul LR (a raportului de verosimilitate).



Prin acţionarea butonului *Save* în dialogul principal se pot preciza noile variabile care pot fi create din ieşirea procedurii, ca şi la regresia multiplă.

Informaţiile care apar în fişierul de ieşire SPSS sunt explicate în continuare.

Un prim tabel cu informaţiile generale (număr de observaţii valide etc.).

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	370	33.1
	Missing Cases	747	66.9
	Total	1117	100.0
Unselected Cases		0	.0
Total		1117	100.0

a. If weight is in effect, see classification table for the total number of cases.

Un tabel în care se precizează codificările variabilelor categoriale (inclusiv cea dependentă). Pentru variabilele categoriale independente are loc o recodificare cu considerarea categoriei de referinţă: aceasta este recodificată 0.

Dependent Variable Encoding

Original Value	Internal Value
0 no	0
1 yes	1

Categorical Variables Codings

		Frequency	Paramete (1)
GENDER GENDER OF STUDENT	1 FEMALE	109	1.000
	2 MALE	261	.000

Ieşirea diferă ca structură după metoda de selectare a variabilelor, dar conține un prim bloc de informații care se referă la modelul simplu (doar cu termenul constant). De remarcat structura: clasificare, variabile în ecuație, variabile candidate.

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted		
			Board certified in FM, IM, Peds or Ob-Gyn		Percentage Correct
			no	yes	
Step 0	Board certified in FM, IM, Peds or Ob-Gyn	no	309	0	100.0
		yes	61	0	.0
	Overall Percentage				83.5

a. Constant is included in the model.

b. The cut value is .500

Tabelul de clasificare este construit prin considerarea probabilității de clasificare prognozate de modelul curent pentru fiecare observație. după principiul că OR>1 clasează observația în grupul codificat 1. Un model bun trebuie să numere cele mai multe observații pe diagonala principală a tabelului.

Tabelul care urmează, referitor la model, este explicat şi se interpretează potrivit celor spuse la tabelul similar dintr-un pas intermediar afişat ceva mai departe în lucrare.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-1.622	.140	134.090	1	.000	.197

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	MATYR	78.766	1	.000
		GENDER(1)	.865	1	.352
		USMLE2	18.236	1	.000
	Overall Statistics		91.356	3	.000

Informațiile oferite pentru faza finală sunt după structura

Block 1: Method = Forward Stepwise (Likelihood Ratio)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	70.688	1	.000
	Block	70.688	1	.000
	Model	70.688	1	.000
Step 2	Step	10.984	1	.001
	Block	81.672	2	.000
	Model	81.672	2	.000
Step 3	Step	6.862	1	.009
	Block	88.534	3	.000
	Model	88.534	3	.000

Se observă că în fiecare pas al estimării modelului se testează dacă trecerea de la precedent este semnificativă (se respinge ipoteza nulității variabilei sau variabilelor adăugate).

Indicatorii similari coeficientului de determinare din regresia multiplă sunt în tabelul care urmează.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	260.573	.174	.294
2	249.589	.198	.335
3	242.727	.213	.360

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	23.800	7	.001
2	11.019	8	.201
3	7.160	8	.519

Testul Hosmer & Lemeshow este explicitat pentru fiecare pas prin raportarea celor 10 frecvențe observate/aşteptate (statistica şi semnificația sunt raportate în tabelul precedent).

Contingency Table for Hosmer and Lemeshow Test

		PRIMCARE Board certified in FM, IM, Peds or Cb-Gyn = 0 no		PRIMCARE Board certified in FM, IM, Peds or Ob-Gyn = 1 yes		
		Observed	Expected	Observed	Expected	Total
Step 3	1	37	36.817	0	.183	37
	2	37	36.535	0	.465	37
	3	37	36.102	0	.898	37
	4	37	35.507	0	1.493	37
	5	35	34.348	2	2.652	37
	6	31	32.848	6	4.152	37
	7	28	30.777	9	6.223	37
	8	26	28.110	11	8.890	37
	9	26	23.459	11	13.541	37
	10	15	14.496	22	22.504	37

Se raportează de asemenea tabelul de clasificare pentru fiecare pas al procedurii.

Classification Table[a]

			Predicted		
			Board certified in FM, IM, Peds or Ob-Gyn		Percentage Correct
	Observed		no	yes	
Step 1	Board certified in FM, IM, Peds or Ob-Gyn	no	290	19	93.9
		yes	46	15	24.6
	Overall Percentage				82.4
Step 2	Board certified in FM, IM, Peds or Ob-Gyn	no	299	10	96.8
		yes	47	14	23.0
	Overall Percentage				84.6
Step 3	Board certified in FM, IM, Peds or Ob-Gyn	no	296	13	95.8
		yes	41	20	32.8
	Overall Percentage				85.4

a. The cut value is .500

În tabelul referitor la variabilele din model se raportează:

- coeficienții B
- Exp (B) cu interpretarea, dată în partea teoretică, că reprezintă modificare OR a variabilei dependente la modificarea cu o unitate a variabilei independente, deci Exp (B) $\approx$ 1 pentru variabilele nesemnificative.
- informații asociate testul Wald de semnificație a fiecărui coeficient.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MATYR	-.568	.083	46.714	1	.000	.566
	Constant	1129.854	165.478	46.619	1	.000	.
Step 2[b]	MATYR	-.561	.086	42.232	1	.000	.571
	USMLE2	-.026	.008	10.387	1	.001	.974
	Constant	1120.518	171.865	42.507	1	.000	.
Step 3[c]	MATYR	-.598	.090	44.344	1	.000	.550
	GENDER(1)	.957	.366	6.855	1	.009	2.605
	USMLE2	-.029	.008	12.294	1	.000	.971
	Constant	1193.796	178.716	44.621	1	.000	.

a. Variable(s) entered on step 1: MATYR.

b. Variable(s) entered on step 2: USMLE2.

c. Variable(s) entered on step 3: GENDER.

În tabelul următor (apare doar pentru anumite metode de selectare a variabilelor) se prezintă informațiile necesare pentru a testa ce s-ar întâmpla daca o variabilă din model este exclusă. Pentru un model care se construieşte ascendent, acestea pot sugera prezența unor variabile care au devenit nesemnificative prin includerea altor variabile.

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	MATYR	-165.630	70.688	1	.000
Step 2	MATYR	-156.540	63.491	1	.000
	USMLE2	-130.287	10.984	1	.001
Step 3	MATYR	-155.605	68.483	1	.000
	GENDER	-124.795	6.862	1	.009
	USMLE2	-127.933	13.138	1	.000

Pentru variabilele care nu sunt în model, se prezintă testele care decid necesitatea prezenței lor. La pasul următor, va fi introdusă în model variabila cu scorul cel mai mare (scor calculat potrivit metodei selectate).

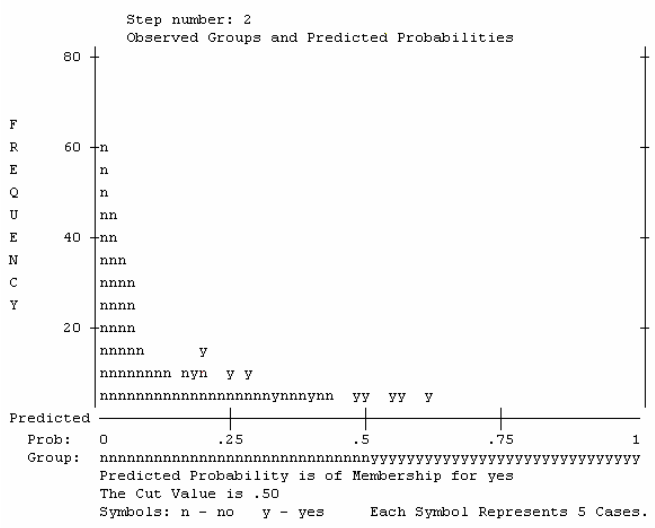
Variables not in the Equation

			Score	df	Sig.
Step 1	Variables	GENDER(1)	4.885	1	.027
		USMLE2	10.948	1	.001
	Overall Statistics		17.728	2	.000
Step 2	Variables	GENDER(1)	7.140	1	.008
	Overall Statistics		7.140	1	.008

Diagrama de clasificare (afişată în continuare) este alcătuită:

- Axa X este probabilitatea prognozată (de la 0 la 1) de a fi clasificat în grupul codat "1". Sub axă sunt diferențiate zonele de clasificare prin simbolurile care codifică grupul 1 (Yes) şi grupul 2 (No). Se observă pragul de 0.5 care schimbă clasificarea.
- Axa Y este frecvență (număr de cazuri).
- Coloanele care apar în diagramă sunt alcătuite din marcaje (fiecare reprezintă un număr de cazuri pentru simplificare) care reprezintă clasificarea observată a cazurilor.
- Examinarea diagramei constă în analiza faptului dacă marcajele corespund la acelaşi semn (Yes/No) situat sub axa X. Prin urmare
    - semnele Y care corespund la valori Y de pe axa OX (şi semnele N care corespund la valori N de pe axa OX) reprezintă clasificări prognozate corect de model.
    - celelalte marcaje (semnele Y care corespund la valori N de pe axa OX, precum şi semnele N care corespund la valori Y de pe

axa OX) reprezintă cazuri clasate eronat, deci observații pentru care modelul estimat nu funcționează.



C. Lucrarea practică

1. Un studiu care urmăreşte de cine depinde gustul brânzeturilor de tip cheddar a prelevat probe şi a determinat concentrația unor compuşi chimici. Fiecare probă a fost supusă unui proces de degustare şi a primit o notă. Unele valori au fost transformate în prealabil (***Acetic*** şi ***H2S*** sunt obținute prin logaritmarea valorilor măsurate).
   Fişierul de date este **www.infoiasi.ro/~val/statistica/CheeseData.txt**
   Variabilele sunt
   i. ***Taste***: nota obținută în urma combinării notelor acordate de mai mulți degustători
   ii. ***Acetic***: logaritm natural din concentrația de acid acetic
   iii. ***H2S***: logaritm natural din concentrația de H_2S.
   iv. ***Lactic***: concentrația de acid lactic
   - Să se modeleze variabila ***Taste*** cu ajutorul celorlalte trei variabile.
   - Să se analizeze modelul obținut.
2. Date privind un număr de companii au fost selectate din lista Forbes 500 pentru anul 1986 (printr-un sondaj sistematic 1/10 din lista alfabetică a companiilor). Studiul urmăreşte volumul de vânzări al companiei.
   Fişierul de date este **www.infoiasi.ro/~val/statistica/ForbesData.txt**
   Variabilele sunt:
   i. ***Company***: numele companiei
   ii. ***Assets***: bunurile companiei (milioane $)
   iii. ***Sales***: volumul de vânzări (milioane $)
   iv. ***Market_Value***: valoarea de piață a companiei (milioane $)
   v. ***Profits***: profitul (milioane $)
   vi. ***Cash_Flow***: volumul tranzacțiilor (milioane $)
   vii. ***Employees***: numărul de angajați (mii persoane)
   viii. Sector: domeniul de activitate a companiei.

- Să se modeleze volumul de vânzări în funcție de celelalte variabile. Să se analizeze modelul obținut.
- Să se determine transformările prealabile necesare pentru unele variabile şi să se refacă modelarea.

3. Se va deschide fişierul **Employee Data.sav** din setul de fişiere test oferite de SPSS. Să se decidă dacă faptul că un angajat aparține minorității (*minority* = 1) este reflectat de variabilele *educ, prevexp, jobcat* şi *gender*.
   Pentru aceasta se va estima şi se va analiza o regresie logistică în care variabila dependentă este *minority*, restul variabilelor fiind considerate independente.