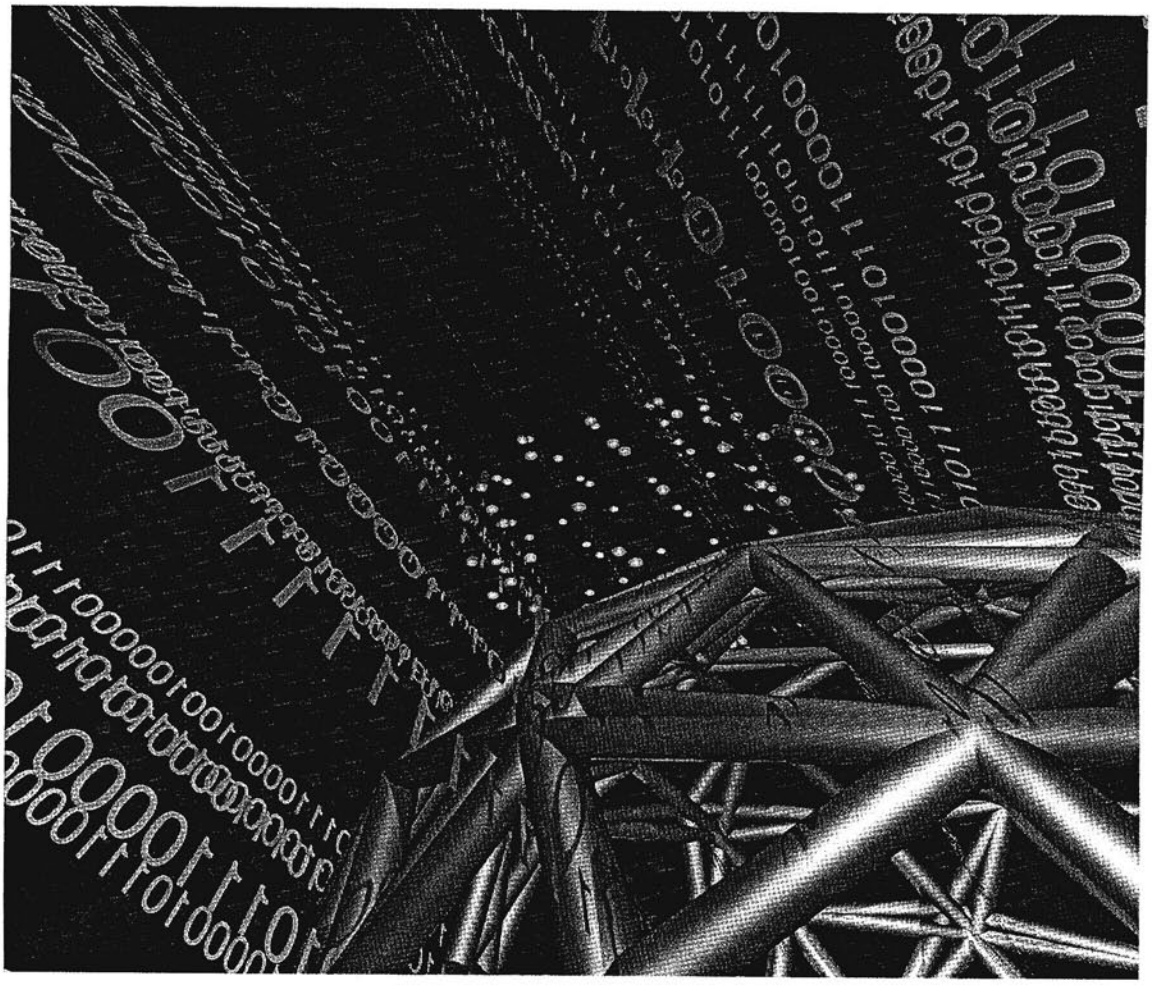
Celostátní seminář – sborník přednášek

Analýza dat 2006/II
Statistické metody pro praxi

Lázně Bohdaneč 24. - 27. 10. 2006



Pokročilé statistické metody
pro řízení jakosti, technologickou, výzkumnou a zkušební praxi

TriloByte® Ltd. STATISTICAL SOFTWARE

TriloByte Statistical Software

CQR Výzkumné centrum pro jakost a spolehlivost

Centrum pro jakost a spolehlivost ve výrobě
Center of Quality and Reliability of Production

Editor: Karel Kupka

Celostátní seminář – sborník přednášek
Analýza dat 2006/II
(Pokročilé statistické metody
pro řízení jakosti, technologickou, výzkumnou a zkušební praxi)

Lázně Bohdaneč 24. - 27. 10. 2006

Vydal (c) 2007

TriloByte® Ltd. STATISTICAL SOFTWARE

TriloByte Statistical Software, Jiráskova 21
530 02 Pardubice, Czech Republic
info@trilobyte.cz, http://www.trilobyte.cz

ISBN 978-80-239-8998-4

Základy logistické regrese

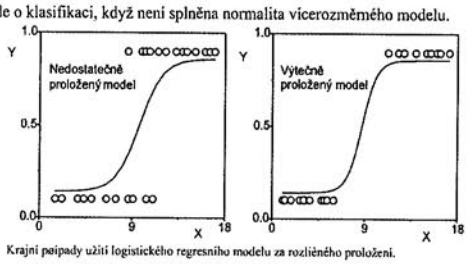
Milan Meloun
Univerzita Pardubice

Logistická regrese LR

1. Zaměření metody LR

Navržena v 60tých letech jako alternativní postup k MNČ, když je závislc proměnná binární (medicina), kde y představuje přítomnost (1) nebo nepřítomnost (0) choroby.

Jde o klasifikaci, když není splněna normalita vícerozměrného modelu.



Krajní paipady užití logistického regresního modelu za rozličného proložení.

Příklad 8.2 ***Významnost sledovaných znaků ovlivňujících ústup leukemie***
Lee (1980) publikoval data o leukemii pacientů a o ovlivnění jejího ústupu. Cílem je nalézt znaky, které jsou v navrženém logistickém regresním modelu statisticky významné k ovlivnění ústupu leukemie.

○ *Data:* Pacientů (řádků) $n = 27$ a znaků (sloupců) $m = 7$.

Závisle proměnnou je *REMISS*: zda se objeví (1) či neobjeví (0) ústup leukemie.
Nezávisle proměnnými jsou u pacientů naměřené hodnoty 6 znaků:
CELL značí celuliritu, buněčnost sraženiny kostní dřeně,
SMEAR značí skvrnu diferenčního procenta napadení,
INFIL značí procento infiltrátu kostní dřeně buňkou leukemie,
LI je procento označeného indexu leukemických buněk kostní dřeně,
BLAST je absolutní počet napadení v periferní krvi,
TEMP značí nejvyšší naměřenou teplotu před začátkem léčby.

Index	*REMI*	*CELL*	*SMEAR*	*INFIL*	*LI*	*BLAST*	*TEMP*
1	1	0.8	0.83	0.66	1.9	1.1	0.996
..	...	...	...	...	...	...	...
27	0	1	0.73	0.73	0.7	0.398	0.986

○ *Řešení:* Byly užity programy NCSS2000 [67], MINITAB [86] a STATISTICA [102].

Podmínky výpočtu:

Závisle proměnná: *REMISS*

Model obsahuje **6 nezávisle proměnných:** *CELL* |*SMEAR* |*INFIL* |*LI* |*BLAST* |*TEMP*.

Objektů (řádků): 27

Znaků (sloupců): 7

Volba proměnných a výstavba modelu:

Třída závisle proměnné může nabývat hodnot, 0 a 1.
Počet je součet četností závisle proměnné pro každou třídu.
Řádky přináší počet objektů v každé třídě tak, jak byl vyčíslen z hodnot nezávisle proměnných.
Prior apriorní pravděpodobnost v každé třídě zadaná předem uživatelem.
Aktuálně vs. Predikce, R^2 značí hodnotu R^2 , kterou obdržíme z regrese, kde závisle proměnná ve třídě bude lineární funkcí predikované pravděpodobnosti této třídy.
% Správně klasifikováno přináší procento objektů v této třídě, které byly správně klasifikovány logistickým regresním modelem.

REMISS			*Aktuálně*	*% Správně*	
Třída	*Počet*	*Řádky*	*Prior*	*vs. Predikce,*	*R^2 klasifikováno*
0	18	18	0.66667	0.38775	83.333
1	9	9	0.33333	0.38775	55.556
Total	27	27			74.074

Iterační přiblížení odhadů parametrů:

zvyšuje logit metodou maximální věrohodnosti:

	0. iter.	1. iter.	2. iter.	3. iter.	4. iter.	5. iter.	6. iter.	7. iter.
Úsek	0.0000	-18.7941	-43.8181	-64.4856	-66.9072	-59.9205	-58.1082	-58.0387
LI	0.0000	0.7895	0.4785	-1.7390	-8.4324	-19.9753	-24.3896	-24.6605
CELL	0.0000	5.8190	7.8289	6.5724	-0.7276	-14.0188	-18.9941	-19.2925
TEMP	0.0000	-5.9853	-8.4942	-7.5008	0.0725	14.0932	19.2909	19.6001
SMEAR	0.0000	-2.1173	-2.8763	-3.3905	-3.6410	-3.7879	-3.8874	-3.8959
INFIL	0.0000	-0.0030	-0.0152	-0.0980	-0.1186	-0.1456	-0.1513	-0.1511
BLAST	0.0000	20.6740	47.3696	71.2305	80.4766	84.6743	87.2300	87.4331
Logit	-18.7150	-12.3083	-11.3620	-11.0574	-10.9449	-10.8812	-10.8753	-10.8753

Logit představuje rozhodčí kritérium, pomocí kterého se rozhodne, zda dotyčný parametr logistický model zlepší nebo zhorší.

Průběh postupného krokového zavádění parametrů do logistického regresního modelu

R^2 modelu je odhad koeficientu determinace R^2 pro logistický model proložený daty.
Změna v R^2 udává hodnotu, která se přičte k celkové R^2, když se tento parametr přidá do modelu. Změny R^2 mají vesměs kladné znaménko, zavedení dalšího parametru do modelu způsobí zvýšení hodnoty R^2 čili zlepšení modelu až na konečnou 36.719 %.

R^2 je však pouze přibližná hodnota dle vzorce

$$R^2 = \chi^2(df)/[\chi^2(df) + n - p - 1] ,$$

kde df značí stupně volnosti u χ^2-testu a jsou rovny počtu nezávisle proměnných.

Vysvětlení: Z počátečního logitu ln $L_{(1)}$ = -17.18588 až do terminace ln $L_{(1)}$ = - 10.87533 pro *REMISS* = 0.

Krok výstavby	Parametr zaveden	Logit	Dosažená R^2	Změna v R^2
1	*Úsek*	-17.18588	0.00000	0.00000
2	*LI*	-13.03648	0.24144	0.24144
3	*CELL*	-12.17036	0.29184	0.05040
4	*TEMP*	-10.97669	0.36130	0.06946
5	*SMEAR*	-10.92900	0.36407	0.00277
6	*INFIL*	-10.87752	0.36707	0.00300
7	*BLAST*	-10.87533	0.36719	0.00013

Test postaveného logistického regresního modelu

vychází z H_0: „Směrnice všech parametrů jsou nulové“
proti alternativní H_A: „.... nejsou nulové“.

Kritérium $G = D$ (*model bez proměnné*) $- D$ (*model s proměnnými*) =
$[(-2) \times (-17.18588)] - [(-2) \times (-10.87533)] = 12.621$
pro 6 stupňů volnosti dává vypočtenou hladinu významnosti $P = 0.049463$
menší než $\alpha = 0.05$,
a proto je H_0 o nulovosti všech parametrů v modelu zamítnuta.

Vysvětlení: rozdíl deviancí G testuje,
zda **všechny** regresní parametry β_i kromě úseku β_0 **jsou rovny nule**.

Slovy: D je dvojnásobek rozdílu mezi logaritmem pravděpodobnosti nalezeného logistického regresního modelu a modelu, ve kterém je pouze úsek.

Spočtená hladina významnosti $P = 0.049463$ je menší než $\alpha = 0.05$ a ukazuje, že směrnice parametrů b_i v modelu nejsou rovny nule a navržený model je proto statisticky významný.

R^2 modelu	*df*	Odchylka D	Spočtená hladina významnosti P
0.3869	6	12.62	0.049463

Odhad regresních parametrů a test jejich významnosti:

Parametr b_i značí název β_i pro nezávisle proměnnou, úsek je značen β_0.
Odhad b_i značí směrnici regresního parametru β_i, vyčíslený minimalizací.
Směrodatná odchylka je odhadem přesnosti odhadovaného parametru.
χ^2 pro nulovost parametru $\beta = 0$ udává Pearsonovo testační kritérium χ^2 pro 1 stupeň volnosti k testování H_0: $\beta_i = 0$ vs. H_A: $\beta_i \neq 0$, vyčíslí se dle Waldova kritéria $W^2_{a,i} = [b_i / s(b_i)]^2$.
P značí spočtenou hladinu významnosti pro nalezené odhady: je-li P menší než zvolená $\alpha = 0.05$, je nulovost $\beta_i = 0$ zamítnuta a odhad parametru b_i je statisticky významný.

Vysvětlení: až na *BLAST* a *LI* jsou všechny znaky statisticky významné.

Parameter b_i	Odhad b_i	Směrodatná odchylka	Waldovo Z pro $\beta_i = 0$	Spočtená hladina P	Dolní 95% mez	Horní 95% mez
B0: *Úsek*	-58.03849	1.90413	-30.480	0.00000	61.77051	-54.30647
B2: *CELL*	-24.66154	1.81441	-13.592	0.00000	-28.21772	-21.10537
B3: *INFIL*	19.60126	1.33668	14.664	0.00000	16.98142	22.22110
B4: *LI*	-3.89596	2.35303	-1.656	0.09778	-8.50782	0.71589
B5: *SMEAR*	-19.29357	1.40647	-13.718	0.00000	-22.05020	-16.53695
B6: *TEMP*	87.43390	1.89150	46.225	0.00000	83.72663	91.14117
B1: *BLAST*	-0.15109	1.85979	-0.081	0.93525	-3.79620	3.49402

Nalezený logistický regresní model:.

Dle statistické významnosti odhadů parametrů byl stanoven logistický regresní model pro *REMISS* = 0:

$$-58.04 - 24.66\ CELL + 19.60\ INFIL - 3.90\ LI - 19.29\ SMEAR + 87.43\ TEMP + 0.15\ BLAST.$$

Pearsonův test dobré shody kvality těsnosti proložení, porovnává dané a nalezené četnosti.

Data jsou roztříděna dle svých odhadnutých pravděpodobností od nejmenší do největší z 10 tříd:

1. třída obsahuje data s nejmenší vyčíslenou pravděpodobností a
10. třída s největší pravděpodobností.

Vysvětlení: Model zde prokládá data dobře, dané a vypočtené četnosti jsou si podobné.

Hosmerova-Lemeshowova statistika (= Pearsonovo χ^2-kritérium) nabývá hodnoty spočtené hladiny významnosti $P = 0.477$, větší než $\alpha = 0.05$ a proto dobré proložení.

Třída	1	2	3	4	5	6	7	8	9	10	Celkem
Nízký, nezměněný puls (1)											
Dáno	0	0	0	0	1	0	2	1	3	2	9
Nalezeno	0	0	0.1	0.3	0.6	0.9	0.8	1.6	2.1	2.6	
Vysoký, změněný puls (0)											
Dáno	2	3	3	2	2	3	0	2	0	1	18
Nalezeno	2	3	2.9	1.7	2.4	2.1	1.2	1.4	0.9	0.4	
Celkem	2	3	3	2	3	3	2	3	3	3	27

ROC graf (tabulka)

K určení prahového bodu P_C při klasifikaci objektů do dvou tříd, v události 1 a v neudálosti 0, se použijí křivky *ROC*.

Objekt se zařadí do té třídy, kam se nejvíce přibližuje svou vyčíslenou pravděpodobností.

Z tabulky *ROC* je zřejmé, co nastane, když vybereme rozličné hodnoty pro prahový dělicí bod pravděpodobnosti P_C.

Při zařazování objektů do třídy může nastat jedna ze čtyř možností:
Sloupec A, značený $N(1|1)$: objekt je správně zařazen do této třídy 1 a pochází ze třídy 1.
Sloupec C, značený $N(0|1)$: objekt je nesprávně zařazen do opačné třídy 0 a pochází přitom ze třídy 1.
Sloupec B, značený $N(1|0)$: objekt je nesprávně zařazen do třídy 1 a pochází přitom z opačné třídy 0.
Sloupec D, značený $N(0|0)$: objekt je správně zařazen do opačné třídy 0 a pochází přitom z opačné třídy 0.

Tabulka *ROC* pro *REMISS* = 0

P_C	$N(1\|1)$ A	$N(1\|0)$ B	$N(0\|1)$ C	$N(0\|0)$ D	Senzitivita A/(A+C)	Specificita D/(B+D)	Senzitivita +Specificita	Podíl správně zařazených
0.05	18	9	0	0	1.000	0.000	1.000	0.667
0.10	18	8	0	1	1.000	0.111	1.111	0.704
0.15	17	8	1	1	0.944	0.111	1.056	0.667
0.20	17	8	1	1	0.944	0.111	1.056	0.667
0.25	17	7	1	2	0.944	0.222	1.167	0.704
0.30	17	6	1	3	0.944	0.333	1.278	0.741
0.35	17	4	1	5	0.944	0.556	1.500	0.815
0.40	16	4	2	5	0.889	0.556	1.444	0.778
0.45	15	4	3	5	0.833	0.556	1.389	0.741
0.50	15	4	3	5	0.833	0.556	1.389	0.741
0.55	15	3	3	6	0.833	0.667	1.500	0.778
0.60	15	2	3	7	0.833	0.778	1.611	0.815
0.65	15	1	3	8	0.833	0.889	1.722	0.852
0.70	15	1	3	8	0.833	0.889	1.722	0.852
0.75	12	1	6	8	0.667	0.889	1.556	0.741
0.80	11	0	7	9	0.611	1.000	1.611	0.741
0.85	8	0	10	9	0.444	1.000	1.444	0.630
0.90	8	0	10	9	0.444	1.000	1.444	0.630
0.95	8	0	10	9	0.444	1.000	1.444	0.630

Plocha pod křivkou *ROC*: 0.87963

Počet zařazených objektů do obou tříd, do třídy 0 nebo do třídy 1, je vyčíslován pro různé hodnoty prahového dělicího bodu P_c.

Senzitivita je podíl správně zařazených objektů do první třídy A/(A+C).
Specificita je podíl správně zařazených objektů do opačné třídy D/(B+D).

(Pro optimální hodnotu prahového bodu P_c dosahuje suma senzitivity + specificity maximální hodnoty).

Podíl správně zařazených objektů: vybere se taková hodnota prahového dělícího bodu P_c, pro kterou existuje právě maximální podíl správně zařazených objektů.

Nejspolehlivější pravidlo maximální plochy *AUC* pod křivkou *ROC* ukazuje, že všechny objekty jsou správně zařazeny, čím více se plocha blíží jedné nebo 100 %.

Tabulka *ROC* pro *REMISS* = 1

P_c	$N(1\|1)$	$N(1\|0)$	$N(0\|1)$	$N(0\|0)$	Senzitivita	Specificita	Senzitivita	Podíl správně
	A	B	C	D	A/(A+C)	D/(B+D)	+Specificita	zařazených
0.05	9	10	0	8	1.000	0.444	1.444	0.630
0.10	9	10	0	8	1.000	0.444	1.444	0.630
0.15	9	10	0	8	1.000	0.444	1.444	0.630
0.20	9	7	0	11	1.000	0.611	1.611	0.741
0.25	8	6	1	12	0.889	0.667	1.556	0.741
0.30	8	3	1	15	0.889	0.833	1.722	0.852
0.35	8	3	1	15	0.889	0.833	1.722	0.852
0.40	7	3	2	15	0.778	0.833	1.611	0.815
0.45	6	3	3	15	0.667	0.833	1.500	0.778
0.50	5	3	4	15	0.556	0.833	1.389	0.741
0.55	5	3	4	15	0.556	0.833	1.389	0.741
0.60	5	2	4	16	0.556	0.889	1.444	0.778
0.65	5	1	4	17	0.556	0.944	1.500	0.815
0.70	3	1	6	17	0.333	0.944	1.278	0.741
0.75	2	1	7	17	0.222	0.944	1.167	0.704
0.80	1	1	8	17	0.111	0.944	1.056	0.667
0.85	1	1	8	17	0.111	0.944	1.056	0.667
0.90	1	0	8	18	0.111	1.000	1.111	0.704
0.95	0	0	9	18	0.000	1.000	1.000	0.667

Plocha pod křivkou *ROC*: 0.87963

Klasifikace objektů logistickým modelem:

Predikce je určena z logistického regresního modelu při užití klasifikačního prahového bodu.

Procento správně klasifikovaných objektů je procento z celkového počtu, které padne na diagonálu tabulky.

Dáno závisle proměnnou		Nalezeno predikcí logistickým modelem Ne	Ano	Celkem
Ne	četnost	15	3	18
	řádkové procento	83.33	16.67	100.00
	sloupcové procento	78.95	37.50	66.67
Ano	četnost	4	5	9
	řádkové procento	44.44	55.56	100.00
	sloupcové procento	21.05	62.50	33.33
Celkem	četnost	19	8	27
	řádkové procento	70.37	29.63	
	sloupcové procento	100	100.00	

Procento správně klasifikovaných = 74.07

a) *Predikovaná klasifikace*:

Daná třída značí třídu, do které objekt skutečně patří.
Nalezená třída značí třídu vypočtenou dle parametrů logistického regresního modelu.
Logistické skóre je odhad pravděpodobnosti, že objekt patří do třídy *Ne*.
Reziduum představuje rozdíl mezi vypočtenou hodnotou logistického skóre a binárním indexem skutečné třídy. Vyčíslí se: za index třídy *Ne* se dosadí 0 a za index třídy *Ano* dosadí 1.

Řádek	*Daná třída*	*Nalezená třída*	*Logistické skóre*	*Reziduum*
1	Ano (1)	Ano (1)	0.79	0.209
2	Ano (1)	Ne (0)	0.435	0.565
..	...	...	...	...
27	Ne (0)	Ne (0)	0.279	-0.279

Závěrečná tabulka klasifikovaných objektů
na základě navrženého logistického modelu.

	Nalezeno		
Dáno	0	1	Celkově
0	15	3	18
1	4	5	9
Celkem	19	8	27

Procento správně klasifikovaných 74.1 %

b) *Chybně klasifikované objekty*:

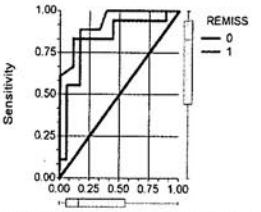
Jsou zde zobrazeny pouze chybně zařazené řádky.

Řádek	*Daná třída*	*Nalezená třída*	*Skóre*	*Reziduum*
2	Ano (1)	Ne (0)	0.435	0.565
..	...	...	...	...
26	Ano (1)	Ne (0)	0.461	0.539

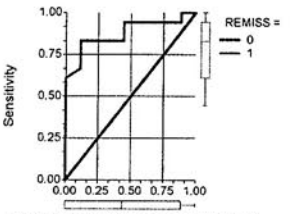
Grafy kvality logistické regresní analýzy.

Graf prahové operační charakteristiky ROC
vystihuje správnost diagnostického testu,
zda logistickým modelem vypočtené Ano nebo Ne je správné.

Na ose *y* se vynáší *senzitivita* a na ose *x* hodnota „1 minus *specificita*".
Křivky pomohou nalézt nejlepší hodnotu dělicího bodu



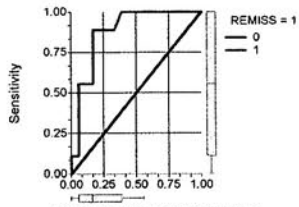
Obr. 8.3a Graf *ROC* pro obě hodnoty *REMISS* = 0 a 1.



Obr. 8.3b Graf *ROC* pro hodnotu *REMISS*= 0.

Graf prahové operační charakteristiky ROC
vystihuje správnost diagnostického testu,
zda logistickým modelem vypočtené Ano nebo Ne je správné.

Na ose y se vynáší *senzitivita* a na ose x hodnota „1 minus *specificita*".
Křivky pomohou nalézt hodnotu dělicího bodu P_c ke klasifikaci objektů.



Obr. 8.3c Graf *ROC* pro hodnotu *REMISS* = 1.

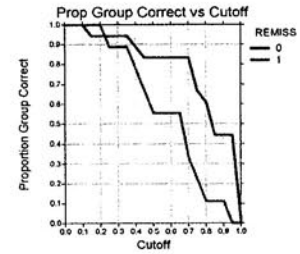
Když leží dělicí bod P_c v levém horním rohu grafu, je dosaženo nejvhodnější hodnoty a dochází k nejlepší klasifikaci objektů.

Kvalita nalezeného logistického modelu se posuzuje také dle plochy *AUC* pod křivkou *ROC*: Čím více se *AUC* blíží jedné nebo 100 %, tím je klasifikace objektů lepší.

Graf podílu správně zařazených objektů v závislosti na P_C

je velice užitečný graf k určení nejlepší hodnoty prahového dělicího bodu P_C.

Na ose y je procento správně zařazených objektů a
na ose x hodnoty pravděpodobnosti dělicího bodu P_C v jednotkách vyčíslované pravděpodobnosti.



Obr. 8.3d Graf závislosti obou křivek podílu správně zařazených objektů na P_c (cutoff).

Nalezený logistický regresní model:.

Dle statistické významnosti odhadů parametrů
byl stanoven logistický regresní model pro *REMISS* = 0:

$$-58.04 - 24.66\ CELL + 19.60\ INFIL - 3.90\ LI$$
$$- 19.29\ SMEAR + 87.43\ TEMP + 0.15\ BLAST.$$

○ *Závěr:*

1. Byl nalezen logistický regresní model znaků *CELL, SMEAR, INFIL, LI, BLAST,* a *TEMP*, které významně ovlivňují znak ústupu leukemie *REMISS*.

2. Znaky *LI* a *BLAST* nejsou statisticky významné.

3. Z *ROC* byl odhadnut prahový dělicí bod pravděpodobnosti, dle kterého se objekty spolehlivě zařadí do dvou tříd ústupu a neústupu leukemie.

Logistická regrese LR se liší od lineární regrese:
predikuje pravděpodobnost události, která se buď stala (1) nebo nestala (0).

Logitová transformace vede na sigmoidální vztah mezi závisle proměnnou y a vektorem nezávisle proměnných x.

Při velmi nízkých hodnotách x se pravděpodobnost proměnné y blíží k nule.
Při vysokých hodnotách x se blíží k jedné.

Logistická regrese používá ***kategorickou závisle proměnnou*** zatímco lineární regrese užívá pouze *spojitou vysvětlovanou proměnnou.*

Logitová transformace vychází z *poměru šancí* či *naděje.*

Dle typu závisle proměnné y se rozlišují:

Binární logistická regrese: binární závisle proměnná nabývá pouze dvou hodnot, například přítomnost-absence, muž-žena. Vektor nezávisle proměnných x obsahuje jednu či více spojitých proměnných (*prediktory*) nebo diskrétních, kategorických (*faktory*).

Ordinální logistická regrese: ordinální závisle proměnná nabývá tři a více možných stavů, např. silný nesouhlas, nesouhlas, souhlas, silný souhlas. Vektor x nezávisle proměnných obsahuje jak *prediktory* tak i *faktory.*

Nominální logistická regrese: nominální závisle proměnná o více než třech úrovních, např. mezi kterými je definována pouze odlišnost. Vektor x může obsahovat jak *prediktory*, tak i *faktory.*

2. Logistický regresní model

V LR potřebujeme vědět, zda se událost stala (1) nebo nestala (0).

Jde o dichotomickou hodnotu 0 - 1 závisle proměnné y, ze které se predikuje odhad pravděpodobnosti, že se událost stala (1) či nestala (0).

Je-li predikovaná pravděpodobnost větší než 0.50, pak se událost stala (1), je-li menší než 0.50, pak se nestala (0).

Postup LR porovnává pravděpodobnost události odehrané $L_{(1)}$ vůči pravděpodobnosti události neodehrané $L_{(0)} = 1 - L_{(1)}$.

Využijeme *pravděpodobnostní poměr* $L_{(1)}/L_{(0)}$, ve kterém pravděpodobnost $L_{(1)}$ je vyjádřena logistickou funkcí

$$L_{(1)} = \frac{1}{1 + e^{c - z}}$$

vděpodobnostní poměr (zvaný "*poměr šanci*") je vyjádřen

$$\frac{L_{(1)}}{L_{(0)}} = e^{a_0 + a_1x_1 + a_2x_2 + \dots + a_px_p}$$

: odhadované koeficienty $a_0, a_1, a_2, \dots, a_p$ jsou míry změny poměru obou pravděpodobností $L_{(1)}/L_{(0)}$.

měr je lineární funkcí diskriminační funkce o p nezávisle proměnných

$$Z = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_p x_p$$

zlogaritmování a úpravě vyjde

$$C - Z = \ln\left(\frac{L_{(0)}}{L_{(1)}}\right)$$

e C je absolutní člen a_0.

Dle klasifikačního postupu je

$$L_{(0)} = P(G = 1 \mid x) \quad \text{a}$$
$$L_{(1)} = P(G = 1 \mid x) = 1 - P(G = 1 \mid x).$$

a po úpravách bude

$$\ln\left(\frac{L_{(1)}}{L_{(0)}}\right) = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$$

kde $b_0 = -C + a_0$, $b_i = a_i$ pro $i = 1, \dots, p$.

Například: ve sportu řekneme, že tým má šanci 3:1. Tvrzení říká, že favorizovaný tým má pravděpodobnost vítězství $\frac{3}{3+1} = \frac{3}{4} = 0.75$.

Platí tedy pravděpodobnostní poměr $\frac{L_{(1)}}{L_{(0)}} = \frac{0.75}{1 - 0.75} = \frac{3}{1}$.

posteriorní pravděpodobnost $P(G = j \mid x)$ zařazení do j-té ategorie: logistický model lze rozšířit na případ K tříd, a předpokládat, že osteriorní pravděpodobnost $P(G = j \mid x)$ zařazení do j-té kategorie bude

$$\ln \frac{P(G = 1 \mid x)}{P(G = K \mid x)} = b_{1,0} + b_1^T x$$

$$\ln \frac{P(G = 2 \mid x)}{P(G = K \mid x)} = b_{2,0} + b_2^T x$$

$$\dots$$

$$\ln \frac{P(G = K-1 \mid x)}{P(G = K \mid x)} = b_{K-1,0} + b_K^T x$$

Po zpětné transformaci vychází

$$P(G = j \mid x) = \frac{\exp(b_{j,0} + b_j^T x)}{1 + \sum_{l=1}^{K-1} \exp(b_{l,0} + b_l^T x)}$$

a

$$P(G = K \mid x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(b_{l,0} + b_l^T x)}$$

Označíme pravděpodobnost

$$P(G = K \mid x) = p_k(\mathbf{x}, \mathbf{b}),$$

aby se zvýraznilo, že jde o funkci regresních parametrů

$$\boldsymbol{b} = [b_{1,0}, \boldsymbol{b}_1, b_{2,0}, \boldsymbol{b}_2, \dots, b_{K-1,0}, \boldsymbol{b}_{K-1}].$$

(Pro $K = 2$ přechází tento model na standardní logistický model pro binární vysvětlovanou proměnnou $y = G$).

Odhady parametrů:

Pro odhad parametrů logistických modelů se používá metoda maximální věrohodnosti.

Přítomnost v první třídě $y = 1$ je pro $G = 1$.
Nepřítomnost v první třídě $y = 0$ je pro $G = 2$ čili přítomnost ve druhé třídě.

Výchozí data: vektor y rozměru $n \times 1$ a matice X rozměru $n \times m$.

Pro i-tý objekt má y_i hodnotu buď 0, nebo 1 a x_i^T je i-tý řádek matice X.

Označme $\quad p(x, b) = p_1(x, b)$ a
$1 - p(x, b) = p_2(x, b)$

a za předpokladu binomického rozdělení y lze zapsat logaritmus věrohodnostní funkce ve tvaru

$$\ln L(b) = \sum_{i=1}^{n} \{y_i \ln p(x_i, b) + (1 - y_i) \ln(1 - p(x_i, b)\}$$
$$= \sum_{i=1}^{n} \{y_i\, b^T x_i - \ln(1 + \exp(b^T x_i\}$$

kde $b^T = \{b_0, b_1\}$ a předpokládá se, že první sloupec matice X obsahuje pouze jedničky (absolutní člen).

Metoda odhadu parametrů:

Pro maximalizaci $\ln L(b)$ se využívá nulity prvních derivací

$$J = \frac{d \ln L(b)}{db} = \sum_{i=1}^{n} x_i\, (y_i - p(x_i, b)) = 0$$

Jde o soustavu $m + 1$ nelineárních rovnic vzhledem k b. . Řešení soustavy nelineárních rovnic využívá Newtonův-Raphsonovův algoritmus, který vyžaduje matici druhých derivací (hessiánu)

$$H = \frac{d^2 L(b)}{db\, db^T} = -\sum_{i=1}^{n} x_i\, x_i^T\, p(x_i, b)\, (1 - p(x_i, b))$$

Newtonova-Raphsonova metoda je iterativní, takže výsledkem j-té iterace je zpřesněný odhad

$$b_{(j+1)} = b_{(j+1)} - H_{(j)}^{-1}\, J_j$$

kde pro vektor pravděpodobnosti p rozměru $n \times 1$ s prvky $p(x_i, b_{(j)})$, a diagonální matici vah W rozměru $n \times m$ s prvky lze psát

$$J_{(j)} = X^T (y - p) \quad \text{a} \quad H = -X^T W X$$

Interpretace regresních koeficientů

Žádné předpoklady o x neexistují a x mohou být jak **diskrétní** (*faktory*) tak i **spojité** veličiny (*prediktory*).

Předpoklad říká, že **logit** $\ln(L_{(1)}/L_{(0)})$ je lineární funkcí nezávisle proměnných.

Pro $\ln(L_{(1)}/L_{(0)})$ se užívá termín **logit**

nebo-li **logit transformace pravděpodobnosti.**

Logistický model se nazývá *vícenásobný logistický regresní model* (krátce *logit*) a koeficienty b_i jsou interpretovány jako regresní koeficienty.

Logit lze ale také upravit: dosazením za $L_{(1)} = (1 - L_{(0)})$ dostaneme

$$L_{(0)} = \frac{1}{1 + \exp[-(b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p)]}$$

Obecně:
Kladné znaménko koeficientu b_i zvyšuje pravděpodobnost $L_{(0)}$, a *záporné znaménko* tuto pravděpodobnost snižuje.

Diskuse koeficientu b_i:

1) Je-li b_i kladné, funkce exp je větší než 1 a pravděpodobnostní poměr $(L_{(1)}/L_{(0)})$ se bude zvyšovat.

   Zvýšení se objeví, když predikovaná pravděpodobnost odehrané události $L_{(1)}$ se zvýší a predikovaná pravděpodobnost neodehrané události $L_{(0)}$ se sníží.
   Proto má model vyšší predikovanou pravděpodobnost odehrané události $L_{(1)}$.

2) Je-li b_i záporné, je funkce exp menší než 1 a pravděpodobnost se **sníží.**

3) Pro koeficient rovný nule vede funkce exp k hodnotě 1 čili k žádné změně pravděpodobnosti.

Test významnosti regresních koeficientů

Logistická regrese umožňuje testovat významnost koeficientů čili ověřit, že regresní koeficient se liší od nuly.

Nula zde značí, že pravděpodobnostní poměr $L_{(1)}/L_{(0)}$ se nemění a pravděpodobnost tím pádem není ovlivněna.

Studentův *t*-test k vyšetření statistické významnosti jednotlivých regresních koeficientů.

Pro velké výběry lze užít *Waldovo testační kritérium* $W_{a,i} = (b_i/s(b_i))^2$, které vyčísluje statistickou významnost pro odhady regresních koeficientů stejně jako ve vícenásobné regresi.

Waldova statistika $W_{a,i}$ má χ^2-rozdělení s 1 stupněm volnosti a představuje čtverec poměru odhadu regresního koeficientu a jeho směrodatné odchylky $W_{a,i} = (b_i/s(b_i))^2$.

Pro kategorické proměnné má $W_{a,i}$ počet stupňů volnosti roven o 1 méně než je počet kategorií.

Waldova statistika W_a má ale jednu nežádoucí vlastnost. Když je absolutní hodnota regresního koeficientu b_i veliká a odhad je ho směrodatné odchylky $s(b_i)$ je také veliký, je výsledkem příliš malá hodnota testačního kritéria $W_{a,i}$, která vede k selhání zamítnutí nulové hypotézy, že regresní koeficient je nulový. Proto, je -li regresní koeficient veliký, neužijeme Waldova kritéria.

Parciální korelace

Je obtížné určit příspěvek jednotlivých proměnných.

Příspěvek každé proměnné závisí také na ostatních proměnných v logistickém modelu.

K vyšetření parciální korelace mezi závisle proměnnou a každou nezávisle proměnnou se užívá **korelační koeficient** R_i, (v intervalu od -1 do +1).

1) **Kladné hodnoty** R_i: když roste hodnota R_i, zvyšuje se pravděpodobnost objektu "v události" $L_{(1)}$.

2) **Záporné hodnoty** R_i: naopak snižuje se pravděpodobnost objektů "v události" $L_{(1)}$.

3) **Malé hodnoty** R_i: proměnná má malý vliv na model.

Kategorické proměnné

Jednou z důležitějších výhod logistického modelu je možnost užívat i kategorické nezávisle proměnné x, zvané **faktory**.

Za **faktor** lze použít numerickou, textovou nebo datumovou hodnotu, zvanou *úroveň* nebo *referenční hladina*.

Interpretace odhadovaných regresních koeficientů je relativní vůči této hladině.

Nejjednodušší situací je **jediný faktor** x se dvěma možnými hodnotami, například, deprese u 143 žen a 101 mužů je ovlivněna pohlavím, kde **faktor pohlaví má dvě úrovně:** pro muže je $x = 0$ a pro ženy je $x = 1$.

objektem žena, pak pravděpodobnostní poměr, že je žena v depresi, je například 40/143. Podobně je tento poměr u mužů například 10/101. Pravděpodobnostní poměr, že **jedinec** je v depresi bude

$$\frac{L_{(1)}}{L_{(0)}} = \frac{40/143}{10/101} = 2.825.$$

oto **šance žen** čili *pravděpodobnostní poměr žen* nacházet se v depresi je 2.825krát větší než šance mužů.

odobně můžeme vyčíslit také šanci "nebýt v depresi":

$$\frac{L_{(0)}}{L_{(1)}} = \frac{143/40}{101/10} = 0.354.$$

mu šance se hodně využívá v biomedikálních aplikacích. Je mírou spojení binární proměnné, jako je faktor risku výskytu dané události, například nemoci.

Kategorická proměnná čili faktor má dvě úrovně, tj. $x = 0$ značící muže a $x = 1$ značící ženy a logistickou rovnici pak bude

$$L_{(1)} = \frac{1}{1 + e^{-a - bx}}$$

odhad parametru $a = -2.313$ a odhad $b = 1.039$.

Odhad b představuje přirozený logaritmus pravděpodobnostního poměru žen a mužů, $1.039 = \ln 2.825$, a proto pravděpodobnostní poměr $e^b = e^{1.039} = 2.825$.

Odhad a je přirozený logaritmus pravděpodobnostního poměru mužů ($x = 0$) nebo $-2.313 = \ln 10/101$. Existuje -li pouze jedna dichotomní proměnná, není potřebné provádět logistickou regresní analýzu.

Přibližný interval spolehlivosti pro pravděpodobnostní poměr jako pro binární proměnnou se vypočte užitím odhadu směrnice b a odhadu je jí směrodatné odchylky.

Například 95% interval spolehlivosti pro pravděpodobnostní poměr se vyčíslí jako $\exp(b \pm 1.96\, s(b))$.

. Volba proměnných

apříklad úloha logistické předpovědi infarktu:

ata jsou z dlouhodobého sledování z počátku zdravých pacientů, u kterých byla dlouhodobě provedena opakovaná měření. Několik jedinců bylo postiženo infarktem, několik ne.

yl sledován **výběr nezávisle proměnných**, které by mohly odhalit blížící se infarkt.

ýběr **účinných** nezávisle proměnných byl předem lékaři vytypován.

astěji však uživatel předem neví nic o nezávisle proměnných.

roměnné x jsou nejprve vyšetřovány, která je nejvíce spjata z dichotomní závisle proměnnou.

Studentův t-test významnosti jednotlivých parametrů: užívá se dostatečně vysoká hladina významnosti, například $\alpha = 0.15$, aby užitečná nezávisle proměnná nemohla být odstraněna.

Vyšetření zredukuje počet nezávisle proměnných na 10 či ještě méně.

Pak nastoupí *kroková logistická regresní analýza*: jde o test, zda proměnná x_i zlepší predikční schopnost modelu. Postupy a jejich kritéria jsou užita k rozhodování, kolik proměnných x_i a které je třeba užít.

Testy v dopředné krokové analýze jsou postaveny na χ^2-statistice: velká hodnota χ^2 nebo malá spočtená hladina významnosti P ukazují, že nezávisle proměnná by měla být zařazena do proměnných.

Nalezená velká hodnota χ^2 ukazuje , že proměnné jsou užitečné.

4. Těsnost proložení logistickým modelem

Před analýzou je třeba posoudit, zda nejsou odlehlé hodnoty. Rozptylové diagramy snadno odhalí **odlehlé body**.

Proměnné nemusí být **normálně** rozděleny.

Regresní diagnostika s analýzou vlivných bodů odhalí O a E.

Logistická křivka má esovitý tvar a vystihuje logistický model, který je vzhledem ke koeficientům b nelineární.

Po **linearizační transformaci** budou koeficienty představovat směrnice u proměnných lineárního regresního modelu.

Mírou těsnosti proložení navrženého modelu daty je hodnota pravděpodobnosti $L_{(1)}$, že se událost uskuteční.

Místo veličiny $L_{(1)}$ se používá tzv. ***odchylka*, deviance** $D = -2 \ln L_{(1)}$ čili

$$D = -2LL.$$

D **představuje** *míru těsnosti proložení dat logistickým regresním modelem:*

1) *Dobrý model* vede k vysoké pravděpodobnosti objektů v události $L_{(1)}$, což přetransformáno do veličiny $-2 \ln L_{(1)}$ poskytne malou hodnotu *blízkou nule*.
2) *Minimální hodnotou* pro $-2 \ln L_{(1)}$ je *nula*, při které je dosaženo naprosto perfektní těsnosti proložení.

Rozdíl v odchylce je definován vztahem

$$G = D(\text{model bez proměnné}) - D(\text{model s proměnnou})$$

čili

$$G = -2 \ln \frac{\text{pravděpodobnost modelu bez proměnné}}{\text{pravděpodobnost modelu s proměnnou}}$$

Veličina G proto odpovídá věrohodnostnímu poměru.

Těsnost proložení: spočívá porovnání experimentálních hodnot E s vypočtenými V:

Pearsonův test dobré shody χ^2 se užije, když model platí:
- Velká hodnota χ^2 indikuje špatné proložení modelu.
- Malé hodnoty vypočtené hladiny významnosti P indikují špatné proložení modelu.

Nejužívanější způsoby posouzení těsnosti proložení:

Nejužívanější způsoby posouzení těsnosti proložení:

a) **Klasický Pearsonův přístup** začíná s identifikováním různých kombinací hodnot proměnných v regresním modelu, tj. *vzorů*.

Například **dvě dichotomní proměnné**,(pohlaví a zaměstnán) vedou na 4 kombinace: *muž zaměstnán, muž nezaměstnán, žena zaměstnána, žena nezaměstnána.*

- Pro každou kombinaci vyčíslíme počet E experimentálních hodnot jednotlivců (objektů) ve třídě I a II.

- Podobně pro každého jednotlivce vypočteme pravděpodobnost, že se nachází ve třídě I a ve třídě II logistickou regresní analýzou.

- Suma těchto pravděpodobností pro daný vzor se označí V.

- Testační statistika testu dobré shody χ^2 se vyčíslí jako

$$\chi^2_{exp} = \sum_{i=1} 2E\left(\ln \frac{E}{V}\right)$$

kde suma se provede přes všechny odlišné vzory.

- Rezidua se sledují právě pro tyto odlišné vzory.

b) **Hosmerův-Lemeshowův test dobré shody** byl navržen v 1982. Pearsonův χ^2-test dobré shody k redukci v logaritmech hodnoty pravděpodobnosti je mírou sledování zlepšení těsnosti zavedením jedné či více nezávisle proměnných.

Základní model, který je podobný výpočtu sumy čtverců při použití pouze průměrů, poskytuje nulovou linii k porovnání.

Vedle χ^2-testu existuje několik R^2-podobných měr k posouzení těsnosti proložení, obdoba koeficientu determinace ve vícenásobné regresi.

"Pseudo R^2" v logistické regresi pro logitový model se vypočte dle

$$R^2_{logit} = \frac{2\ln L_{nul} - (-2\ln L_{model})}{-2\ln L_{nul}} = -\frac{D_{model} + D_{nul}}{D_{nul}}$$

c) **Metoda klasifikačních matic**, vyvinutých v diskriminač analýze slouží k vyhodnocení predikční schopnosti v pojmech zařazení třídy.

Pravděpodobnost zařazení do třídy I je vypočtena pro každého jednotliv (objekt) ve výběru a výsledný počet je uspořádán vzestupně. Pravděpodobnosti jsou pak rozděleny do 10 skupin (decily).

Pro každý naměřený počet jednotlivců ve třídě I je vyčíslen počet E. Užití logické regrese jsou pro jedince v každém decilu vypočteny počty V. P se vyčíslí *Pearsonova χ^2 -statistika testu dobré shody*

$$\chi^2_{exp} = \sum_{i=1}^{n} \frac{(E - V)^2}{V}$$

kde sumace se provede přes obě třídy a 10 decilů.

Velká hodnota χ^2 nebo malá hodnota P indikují, že proložení není dobré.

5. Kvalita vyhodnocení logistickou regresí

Třídíme objekty do tříd, musíme nalézt *prahový bod pravděpodobnosti* P_C: objekt je "v události", když pravděpodobnost události větší nebo rovna hodnotě P_C.

Graf prahové operační charakteristiky ROC k detekci signálu, když signál nebylo vždy možné správně přijmout.

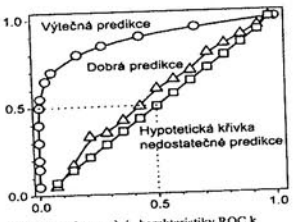
Na ose y je procento správně zařazených objektů "v události" nazvané **pozitivní podíl** (a v lékařském výzkumu nazývané **citlivost**).

Na ose x je procento nesprávně zařazených objektů nazvané **falešný podíl** nebo v lékařském výzkumu **"1 minus specificita"** (v lékařském výzkumu nazývané **senzitivita** zařazených krys pro správně zařazené krysy a **specificita** krys pro falešně zařazené krysy).

1) **Horní křivka v grafu ROC představu je výtečnou predikci:** i pro m hodnoty podílu nesprávně zařazených objektů se získá vysoké proce správně zařazených objektů, které **skutečně jsou "v události"**.

2) **Střední křivka je skutečná křivka při uvažování malého počtu nezáv proměnných,** - třeba dvou. Vysoké procento (80 %) objektů sprá zařazených v události je v poměru k 65 % chybně zařazených v udál na nepřijatelné hladině.

3) **Dolní hypotetická křivka, (přímka) odpovídá nahodilým výsledk** například házení mincí. Blízkost střední křivky k dolní ukazuje , ž potřeba buď volit jinou, anebo přidat ještě další nezávisle proměnn abychom získali lepší model, i když je ale tento model statisti významný na spočtené hladině $P = 0.009$.



Graf prahové operační charakteristiky ROC k detekci signálu, když signál není možné vždy správní pøijmout.

) Vybereme prahový bod na dolní části křivky grafu ROC a nechceme mít příliš mnoho objektů, zařazených jako "v události", bude se nazývat **přísný práh**.
Nevýhoda: je ztráta mnoha objektů, které jsou "v události".

) Vybereme prahový bod na horní části křivky grafu ROC a chceme mít hodně objektů zařazených jako "v události", bude se nazývat **nedbalý práh.**
Nevýhoda: sice velmi málo objektů "v události" bude ztraceno ale mnoho objektů "v neudálosti" bude chybně označeno jako "v události".

Křivky v grafu ROC **musí procházet body (0, 0) a (1, 1).**

Maximální plocha pod křivkou je je dna čili 100%.
Numerická hodnota velikosti plochy bude blízká 1, když predikce modelu bude výtečná.
Když bude **plocha blízká hodnotě 0.5**, bude predikce modelu špatná.

Křivka ROC je proto užitečná při rozhodování, který ze dvou logistických modelů vybrat: lepší model dosáhne větší plochy pod křivkou ROC ale také větší výšky prahového bodu na křivce ROC.

Většina programů vybírá logistický model podle kritéria největší plochy pod křivkou ROC.

. Aplikace logistické regrese

odelu vícenásobné logistické regrese se často užívá k odhadu pravděpodobnosti jisté události, která se přihodí danému objektu.

určení logistického regresního modelu je třeba výběru dat, ve kterém každý objekt, jedinec byl sledován v uvedeném časovém období a hodnoty závažných proměnných byly od začátku pečlivě zaznamenávány.

Výběr dat může být uskutečněn dvojím způsobem:

1. **Výběr cross-validation:** je získán náhodným způsobem a pozorování provedeno v uvedeném časovém období. Z tohoto výběru se vyčlení dva podvýběry: **první podvýběr**, který obsahuje hodně zkušenosti o události, a **druhý podvýběr**, který obsahuje zbylé údaje .

Na datech prvního podvýběru se vyčíslí logistický regresní model, který pak může být aplikován na člena druhého podvýběru.

Předpokládá se, že **původní výběr je v ustáleném stavu**, tzn. neobjevily se žádné podstatné změny, které by pozměnily vztah mezi nezávisle proměnnými a výskytem události.

2. **Případ řídícího výběru:** spočívá v získání dvou náhodných výběrů: **první výběr**, ve kterém se událost objeví, a **druhý výběr**, ve kterém se událost neobjeví.

Hodnoty predikovaných proměnných se musí získat retrospektivním způsobem, z minulých záznamů nebo ze vzpomínek.

Konstanta *a* musí být nastavena tak, aby vyjadřovala pravý poměr objektu v události.

Existují **důležité požadavky,** které je třeba respektovat:

1. Model předpokládá, že logaritmus pravděpodobnostního poměru je lineárně závislý na nezávislých proměnných. Nesplnění by mělo být předem prověřeno buď užitím měr těsnosti proložení, nebo jinými způsoby. To může vyžadovat transformaci dat.

2. Výpočty jsou často časově náročné, a proto by měl uživatel rozumně redukovat počet proměnných.

3. Logistická regrese by se neměla užívat k vyhodnocení faktorů risku v dlouhodobých studiích, ve kterých jsou je dnotlivé studie rozličné délky.

4. Regresní koeficienty pro nezávisle proměnnou v logistickém regresním modelu závisí na ostatních proměnných, zařazených do logistického modelu. Koeficienty pro stejnou nezávisle proměnnou, když se použijí různé výběry proměnných, mohou být zcela odlišné.

5. Je -li užita sehraná analýza, kterákoliv proměnná pro sehrání nemůže být použita jako nezávisle proměnná.

6. Jsou okolnosti, kde metoda maximální věrohodnosti odhadovaných regresních koeficientů neposkytne odhady, tj. nekonvergu je .

PŘÍKLAD 8.5 *Významnost znaků intenzivní péče pacientů na JIP*

Sledováno 200 pacientů na jednotce intenzivní péče JIP. Cílem bylo postavit logistický regresní model k predikci pravděpodobnosti přežití pacienta a vyšetřit významnost rizikových faktorů úmrtnosti.

Zdrojová matice:

20 znaků pro 200 pacientů

ID je kód pacienta.
STA je kód přežití: 0 = přežije, 1 = nepřežije.
AGE vyjadřuje věk v letech.
RACE značí rasu: 1 = bílá, 2 = černá, 3 = ostatní.
SER značí obsluhu na JIP: 0 = běžná lékařská, 1 = operativní, chirurgická.
CAN udává, zda rakovina je akutní problém tohoto pacienta: 0 = není, 1 = ano je.
CRN značí existenci chronické ledvinové poruchy: 0 = není, 1 = ano je.
INF zda pacient dostal infekci až po vstupu na JIP: 0 = ne, 1 = ano.
CPR vyšetření předcházelo přijetí na JIP: 0 = ne, 1 = ano.
SYS udává systolický krevní tlak při přijetí na JIP v mm Hg.
HRA je krevní puls při přijetí na JIP v pulsech/min.
PRE značí předešlá přijetí na JIP v poslední půl roce: 0 = ne, 1 = ano.
TYP značí způsob přijetí na JIP: 0 = dle volby, 1 = v mimořádné nouzi.
FRA je důvod přijetí na JIP jako dlouhá kost, šíje, jednoduchá plocha, zlomenina kyčle: 0 = ne, 1 = ano.
PO2 je uvolněný kyslík z krve: 0 pro $PO2 \geq 60$ a 1 pro $PO2 < 60$.
PH udává pH krve: 0 pro $PH \geq 7.25$ a 1 pro $PF < 7.25$.
PCO značí CO_2 v krvi: 0 pro $PCO \leq 45$ a 1 pro $PCO > 45$.
BIC vyjadřuje diuhličitan v krvi: 0 pro $BIC \leq 18$ a 1 pro $BIC > 18$.
CRE značí creatinin v krvi: 0 pro $CRE \leq 2.0$ a 1 pro $CRE > 2.0$.
LOC značí stav vědomí při přijetí na JIP: 0 = žádné kóma, 1 = hluboké omámení, 2 = kóma.

o *Řešení:* Byly užity programy NCSS2000 [67], MINITAB [86] a STATISTICA [102].

Podmínky výpočtu:

Závisle proměnná: *STA*
Nezávisle proměnné: *AGE|SEX|RACE|SER|CAN|CRN|INF|CPR |SYS|HRA|PRE|TYP|FRA|PO2|PH|PCO|BIC|CRE|LOC*
Objektů (řádků): 200
Znaků (sloupců): 20

Volba proměnných a průběh výstavby modelu:

Po ukončeném iteračním přiblížení.

STA Třída	Řádky	Prior	Aktuálně vs. Predikce, R^2	Správně klasifikováno[%]
0	160	0.80000	0.35837	98.125
1	40	0.20000	0.35837	40.000
Celkem	200			86.500

Výstavba logistického regresního modelu

obsahuje postupné iterační zjemňování směrnic regresních parametrů až po dosažení terminačního kritéria pro $STA = 0$.

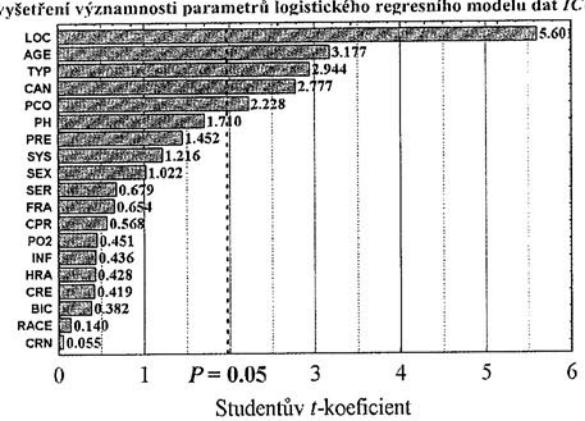
	0. iter.	1. iter.	2. iter.	3. iter.	4. iter.	5. iter.	6. iter.
B0: *Úsek*	0	2.383	3.910	4.962	5.364	5.413	5.414
B1: *AGE*	0	0.018	0.035	0.048	0.052	0.052	0.052
B2: *SEX*	0	0.216	0.388	0.506	0.550	0.555	0.555
B3: *RACE*	0	0.030	0.029	0.012	0.005	0.005	0.005
B4: *SER*	0	0.184	0.359	0.489	0.540	0.545	0.545
B5: *CAN*	0	-1.025	-1.884	-2.488	-2.728	-2.757	-2.758
B6: *CRN*	0	-0.021	0.027	0.076	0.098	0.102	0.102
B7: *INF*	0	-0.099	-0.087	0.004	0.051	0.056	0.056
B8: *CPR*	0	-0.262	-0.601	-0.864	-0.966	-0.978	-0.978
B9: *SYS*	0	0.004	0.007	0.010	0.011	0.012	0.012
B10: *HRA*	0	0.002	0.003	0.004	0.004	0.004	0.004
B11: *PRE*	0	0.434	0.724	0.876	0.924	0.929	0.929
B12: *TYP*	0	0.879	1.733	2.411	2.705	2.744	2.744
B13: *FRA*	0	-0.273	-0.669	-1.007	-1.137	-1.152	-1.152
B14: *PO2*	0	-0.191	-0.308	-0.361	-0.383	-0.387	-0.387
B15: *PH*	0	-0.868	-1.585	-2.141	-2.384	-2.415	-2.415
B16: *PCO*	0	0.951	1.911	2.738	3.122	3.172	3.173
B17: *BIC*	0	0.166	0.444	0.683	0.782	0.794	0.794
B18: *CRE*	0	-0.212	-0.243	-0.237	-0.234	-0.233	-0.233
B19: *LOC*	0	-1.386	-1.986	-2.460	-2.678	-2.706	-2.706
Logit		138.63	74.500	66.632	64.955	64.802	64.800	64.800

Stanovení parametrů logistického modelu pro $STA = 0$ a pro $STA = 1$. Jsou vyznačeny statisticky významné parametry.

		$STA = 0$			$STA = 1$		
	df	*Logit*	χ^2	*P*	*Logit*	χ^2	*P*
B0: *Úsek*					-100.080		
B1: *AGE*	1	-70.726	11.852	0.001	-96.153	7.855	0.005
B2: *SEX*	1	-65.432	1.264	0.261	-96.153	0.001	0.976
B3: *RACE*	1	-64.800	0.000	0.993	-96.127	0.052	0.819
B4: *SER*	1	-65.256	0.913	0.339	-92.303	7.646	0.006
B5: *CAN*	1	68.831	8.062	0.005	92.016	0.574	0.449
B6: *CRN*	1	-64.809	0.018	0.893	-90.761	2.510	0.113
B7: *INF*	1	-64.805	0.011	0.916	-89.572	2.379	0.123
B8: *CPR*	1	-65.274	0.949	0.330	-87.612	3.919	0.048
B9: *SYS*	1	-65.955	2.310	0.129	-84.644	5.936	0.015
B10: *HRA*	1	-64.877	0.154	0.695	-84.146	0.997	0.318
B11: *PRE*	1	-65.856	2.113	0.146	-83.755	0.781	0.377
B12: *TYP*	1	-69.984	10.369	0.001	-77.114	13.284	0.000
B13: *FRA*	1	-65.403	1.206	0.272	-76.905	0.417	0.519
B14: *PO2*	1	64.902	0.205	0.651	76.904	0.002	0.968
B15: *PH*	1	-66.915	4.230	0.040	-76.889	0.031	0.861
B16: *PCO*	1	-68.421	7.242	0.007	-76.195	1.388	0.239
B17: *BIC*	1	-65.194	0.789	0.375	-76.155	0.080	0.777
B18: *CRE*	1	-64.823	0.047	0.829	-76.005	0.299	0.585
B19: *LOC*	1	-76.005	22.411	0.000	-64.800	22.411	0.000

Paretův graf Studentova *t*-koeficientu pro *STA*

vyšetření významnosti parametrů logistického regresního modelu dat *ICU*



Odhad regresních parametrů a test jejich významnosti

Půltučně jsou označeny statisticky významné parametry nalezené Waldovým testem.

Parametr	Odhad b_i	Směrodatná odchylka s	Wald. test W	P	Dolní mez C_D	Horní mez C_H
B0: *Úsek*	5.414	2.122	6.511	0.011	1.255	9.572
B1: *AGE*	-0.052	0.017	9.203	0.002	-0.086	-0.019
B2: *SEX*	0.555	0.501	1.226	0.268	-0.427	1.537
B3: *RACE*	0.005	0.528	0.000	0.993	-1.031	1.040
B4: *SER*	0.545	0.575	0.898	0.343	-0.582	1.673
B5: *CAN*	-2.758	0.980	7.911	0.005	-4.679	-0.836
B6: *CRN*	0.102	0.762	0.018	0.894	-1.392	1.596
B7: *INF*	0.056	0.535	0.011	0.916	-0.992	1.104
B8: *CPR*	-0.978	0.984	0.989	0.320	-2.906	0.950
B9: *SYS*	0.012	0.008	2.161	0.142	-0.004	0.027
B10: *HRA*	0.004	0.009	0.152	0.697	-0.015	0.022
B11: *PRE*	-0.929	0.629	2.184	0.139	-2.162	0.303
B12: *TYP*	-2.744	0.995	7.603	0.006	4.695	0.794
B13: *FRA*	-1.152	0.999	1.329	0.249	-3.111	0.807
B14: *PO2*	-0.387	0.851	0.206	0.650	-2.055	1.281
B15: *PH*	-2.415	1.231	3.849	0.050	-4.828	-0.002
B16: *PCO*	3.173	1.386	5.241	0.022	0.456	5.889
B17: *BIC*	0.794	0.916	0.751	0.386	-1.002	2.590
B18: *CRE*	-0.233	1.075	0.047	0.828	-2.341	1.874
B19: *LOC*	-2.706	0.752	12.955	0.000	-4.180	-1.233

Nalezený logistický regresní model.

Nalezený model obsahuje pouze statisticky významné parametry s jejich odhadnutými směrnicemi pro $STA = 0$:

$$5.414 - 0.052\,AGE + 0.555\,SEX + 0.005\,RACE + 0.545\,SER - 2.758\,CAN + 0.102\,CRN + 0.056\,INF - 0.978\,CPR + 0.012\,SYS + 0.004\,HRA - 0.929\,PRE - 2.744\,TYP - 1.152\,FRA - 0.387\,PO2 - 2.415\,PH + 3.173\,PCO + 0.794\,BIC - 0.233\,CRE - 2.706\,LOC.$$

Kvalita modelu a těsnost proložení dat.

Odchylka (deviance) $D = -2\ln L = -2LL$ je rovna $D = -64.800$

Test rozdílu mezi nalezeným logistickým regresním modelem a modelem úseku se všemi ostatními parametry b_i nulovými je založen na ***rozdílu v odchylce G*** $= D$ (*model bez proměnné*) $- D$ (*model s proměnnou*) $= 70.561$.

Při 19 stupních volnosti je $P < 0.001$, což dokazuje, že alespoň jeden z regresních parametrů je různý od nuly, protože P je nižší než zvolená hladina $\alpha = 0.05$.

Testy dobré shody se svými poměrně vyššími hodnotami spočtené hladiny významnosti $P = 0.040$ a 0.998 ukazují, že navržený model dobře prokládá data.

Použitý test	χ^2	*df*	*P*
Pearsonův	214.487	180	0.040
Rozdílový	129.6	180	0.998
Hosmerův - Lemeshowův	5.726	8	0.678
Brownovy testy: obecný alternativní	1.407	2	0.495
symetrický alternativní	1.063	1	0.302

Klasifikace objektů logistickým modelem.

Tabulka klasifikovaných objektů na základě navrženého logistického regresního modelu.

	Vypočteno		
Dáno	**0**	**1**	Celkově
0	157	3	160
1	24	16	40
Celkem 1	81	19	200

Procento správně klasifikovaných: 86.5 %

Kvalita modelu logistické regrese.

	df	Kritérium	Kritérium/*df*
Test odchylky (deviance)	180	129.600	0.720
Pearsonův χ^2-test	180	214.487	1.192
Logit		-64.800	

Tabulka ROC pro STA = 0

P_c	N(1|1) A	N(1|0) B	N(0|1) C	N(0|0) D	Senzitivita A/(A+C)	Specificita D/(B+D)	Senzitivita +Specificita	Specificita Podíl správně zařazených
0.050	160	34	0	6	1.000	0.150	1.150	0.830
0.100	160	33	0	7	1.000	0.175	1.175	0.835
0.150	159	30	1	10	0.993	0.250	1.243	0.845
0.200	159	30	1	10	0.993	0.250	1.243	0.845
0.250	158	29	2	11	0.987	0.275	1.262	0.845
0.300	158	27	2	13	0.987	0.325	1.312	0.855
0.350	158	26	2	14	0.987	0.350	1.337	0.860
0.400	158	25	2	15	0.987	0.375	1.362	0.865
0.450	158	25	2	15	0.987	0.375	1.362	0.865
0.500	157	24	3	16	0.981	0.400	1.381	0.865
0.550	155	23	5	17	0.968	0.425	1.393	0.860
0.600	153	20	7	20	0.956	0.500	1.456	0.865
0.650	146	18	14	22	0.912	0.550	1.462	0.840
0.700	142	15	18	25	0.887	0.625	1.512	0.835
0.750	135	13	25	27	0.843	0.675	1.518	0.810
0.800	124	8	36	32	0.775	0.800	1.575	0.780
0.850	105	6	55	34	0.656	0.850	1.506	0.695
0.900	88	3	72	37	0.550	0.925	1.475	0.625
0.950	70	2	90	38	0.437	0.950	1.387	0.540

Plocha pod křivkou *ROC*: 0.846

Tabulka ROC pro STA = 1

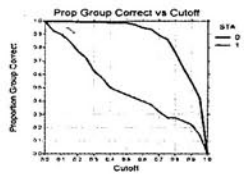
P_c výsek	N(1|1) A	N(1|0) B	N(0|1) C	N(0|0) D	Senzitivita A/(A+C)	Specificita D/(B+D)	Senzitivita +Specificita	Podíl správně zařazených
0.050	38	90	2	70	0.950	0.437	1.387	0.540
0.100	37	72	3	88	0.925	0.550	1.475	0.625
0.150	34	55	6	105	0.850	0.656	1.506	0.695
0.200	32	36	8	124	0.800	0.775	1.575	0.780
0.250	27	25	13	135	0.675	0.843	1.518	0.810
0.300	25	18	15	142	0.625	0.887	1.512	0.835
0.350	22	14	18	146	0.550	0.912	1.462	0.840
0.400	20	7	20	153	0.500	0.956	1.456	0.865
0.450	17	5	23	155	0.425	0.968	1.393	0.860
0.500	16	3	24	157	0.400	0.981	1.381	0.865
0.550	15	2	25	158	0.375	0.987	1.362	0.865
0.600	15	2	25	158	0.375	0.987	1.362	0.865
0.650	14	2	26	158	0.350	0.987	1.337	0.860
0.700	13	2	27	158	0.325	0.987	1.312	0.855
0.750	11	2	29	158	0.275	0.987	1.262	0.845
0.800	10	1	30	159	0.250	0.993	1.243	0.845
0.850	10	1	30	159	0.250	0.993	1.243	0.845
0.900	7	0	33	160	0.175	1.000	1.175	0.835
0.950	6	0	34	160	0.150	1.000	1.150	0.830

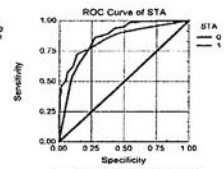
Plocha pod křivkou *ROC*: 0.857

Grafy kvality logistické regresní analýzy.

Graf prahové operační charakteristiky ROC vystihuje správnost diagnostického testu, zda logistickým modelem vypočtené *Ano* nebo *Ne* je správné.

Graf podílu správně zařazených objektů v závislosti na P_c je velice užitečný graf k určení nejlepší hodnoty prahového dělicího bodu P_c.





Obr. 8.8a Graf *ROC* pro *STA* = 0 a 1.

Závěr. V logistickém regresním modelu je z původních 18 znaků nezávisle proměnných statisticky významných 6 znaků (pro proměnnou $STA = 0$), a to *AGE, CAN, TYP, PH, PCO* a *LOC*.

Příklad 4.26 *Volba proměnných k popisu leukemie*

Lee (1980) publikoval data o leukemii pacientů. Závisle proměnnou je binární proměnná *REMISS*, zda se objeví *ústup leukemie y* (1) či neobjeví (0). Nezávisle proměnnými ***x*** jsou:

CELL celulirita, buněčnost sraženiny kostní dřeně,
SMEAR skvrna diferenčního procenta napadení,
INFIL procento infiltrátu kostní dřeně buňkou leukemie,
LI procento označeného indexu leukemických buněk kostní dřeně,
BLAST absolutní počet napadení v periferní krvi,
TEMP nejvyšší teplota před začátkem léčby.

Otázkou je, které nezávisle proměnné jsou statisticky významné v logistickém regresním modelu.

Řešení: Byl užit program NCSS2000.

1. Odhad regresních koeficientů.

Proměnná	Regresní koeficient	Směrodatná odchylka	χ^2 pro $\beta = 0$	Spočtená hladina *P*	Poslední R^2
Úsek	58.0387	71.23627	0.66	0.415224	0.032124
CELL	24.66053	47.83722	0.27	0.606197	0.013113
SMEAR	19.29247	57.94952	0.11	0.739196	0.005511
INFIL	-19.60012	61.68098	0.10	0.750662	0.005023
LI	3.895928	2.3371	2.78	0.095516	0.121993
BLASTS	0.1510942	2.278567	0.00	0.947130	0.000220
TEMP	-87.43308	67.57322	1.67	0.195699	0.077243

χ^2 udává Pearsonovo testační kritérium χ^2 pro 1 stupeň volnosti k testu H_0: $\beta_i = 0$ vs. H_A: $\beta_i \neq 0$. Vyčíslí se Waldovo kritérium $W^2_{\alpha,i} = [b_i / s(b_i)]^2$.

Test významnosti b_i: je-li spočtená hladina *P* menší než předvolená $\alpha = 0.05$, je parametr b_i statisticky významný. Všechny prediktory se jeví jako statisticky nevýznamné.

Poslední R^2 udává hodnotu, která se přičte k celkové R^2, když se tato nezávisle proměnná přidá do logistického regresního modelu. Vypočte se dle vzorce

$$R^2 = \chi^2(df) / [\chi^2(df) + n - p - 1]$$

2. Nalezený model v transformované formě.

Nalezený logistický regresní model: 58.0387 + 24.66053**CELL* + 19.29247**SMEAR* - 19.60012**INFIL* + 3.895928**LI* + 0.1510942**BLAST* - 87.43308**TEMP*.

3. Přehled modelu.

R^2 modelu	*df*	Odchylka *D*	Spočtená hladina významnosti *P*
0.386900	6	12.62	0.049463

Odchylka *D* testuje, zda všechny regresní koeficienty β_i kromě úseku β_0 jsou rovny nule. Protože je spočtená *P* menší než $\alpha = 0.05$, je regresní model statisticky významný.

4. Klasifikační tabulka.

Dáno závisle proměnnou		Nalezeno predikcí logistickým modelem Ne	Ano	Celkově
Ne	Četnost	15	3	18
	Řádkové procento	83.33	16.67	100.00
	Sloupcové procento	78.95	37.50	66.67
Ano	Četnost	4	5	9
	Řádkové procento	44.44	55.56	100.00
	Sloupcové procento	21.05	62.50	33.33
Celkově	Četnost	19	8	27
	Řádkové procento	70.37	29.63	
	Sloupcové procento	100.00	100.00	
Procento správně klasifikovaných = 74.07				

Tabulka přináší četnosti, řádková procenta a sloupcová procenta predikovaných objektů a nakonec je procento správně klasifikovaných objektů. Jde o procento z celkového počtu, které padne na diagonálu tabulky.

5. Predikovaná klasifikace.

Řádek	Daná třída	Nalezená třída	Logistické skóre	Reziduum
1	Ano (1)	Ano (1)	0.790659	0.709341
2	Ano (1)	Ne (0)	0.435096	0.564904
3	Ne (0)	Ne (0)	0.135573	-0.135573
4	Ne (0)	Ne (0)	0.290662	-0.290662
5	Ano (1)	Ano (1)	0.697019	0.302981
6	Ne (0)	Ne (0)	0.298081	-0.298081
7	Ano (1)	Ne (0)	0.376587	0.623413
8	Ne (0)	Ano (1)	0.600523	-0.600523
9	Ne (0)	Ne (0)	0.165277	-0.165277
10	Ne (0)	Ne (0)	0.000858	-0.000858
11	Ne (0)	Ne (0)	0.027695	-0.027695
12	Ne (0)	Ne (0)	0.169829	-0.169829
13	Ne (0)	Ne (0)	0.000956	-0.000956
14	Ne (0)	Ne (0)	0.000001	-0.000001
15	Ne (0)	Ano (1)	0.575858	-0.575858
16	Ano (1)	Ano (1)	0.725752	0.274248
17	Ne (0)	Ne (0)	0.000055	-0.000055
18	Ne (0)	Ne (0)	0.228173	-0.228173
19	Ne (0)	Ne (0)	0.000064	-0.000064
20	Ano (1)	Ano (1)	0.672843	0.327157
21	Ne (0)	Ne (0)	0.015910	-0.015910
22	Ne (0)	Ne (0)	0.007445	-0.007445
23	Ano (1)	Ne (0)	0.247684	0.752316
24	Ne (0)	Ano (1)	0.851096	-0.851096
25	Ano (1)	Ano (1)	0.934464	0.065536
26	Ano (1)	Ne (0)	0.461177	0.538823
27	Ne (0)	Ne (0)	0.279469	-0.279469

Daná třída určuje zadanou skutečnou třídu. *Nalezená třída* představuje nalezenou třídu na základě logistického regresního modelu. *Logistické skóre* je odhad pravděpodobnosti, že objekt patří do třídy *Ne*. Reziduum představu je to rozdíl mezi Logistickým skóre a indexem skutečné třídy. Index třídy *Ne* je 0 a index třídy *Ano* je 1.

6. Chybně klasifikované objekty.

Řádek	Daná třída	Nalezená třída	Skóre	Reziduum
2	Ano (1)	Ne (0)	0.435096	0.564904
7	Ano (1)	Ne (0)	0.376587	0.623413
8	Ne (0)	Ano (1)	0.600523	-0.600523
15	Ne (0)	Ano (1)	0.575858	-0.575858
23	Ano (1)	Ne (0)	0.247684	0.752316
24	Ne (0)	Ano (1)	0.851096	-0.851096
26	Ano (1)	Ne (0)	0.461177	0.538823

Jsou zde zobrazeny pouze chybně zařazené řádky.