NCSS2007
ve vícerozměrné statistické analýze

Logistická regrese

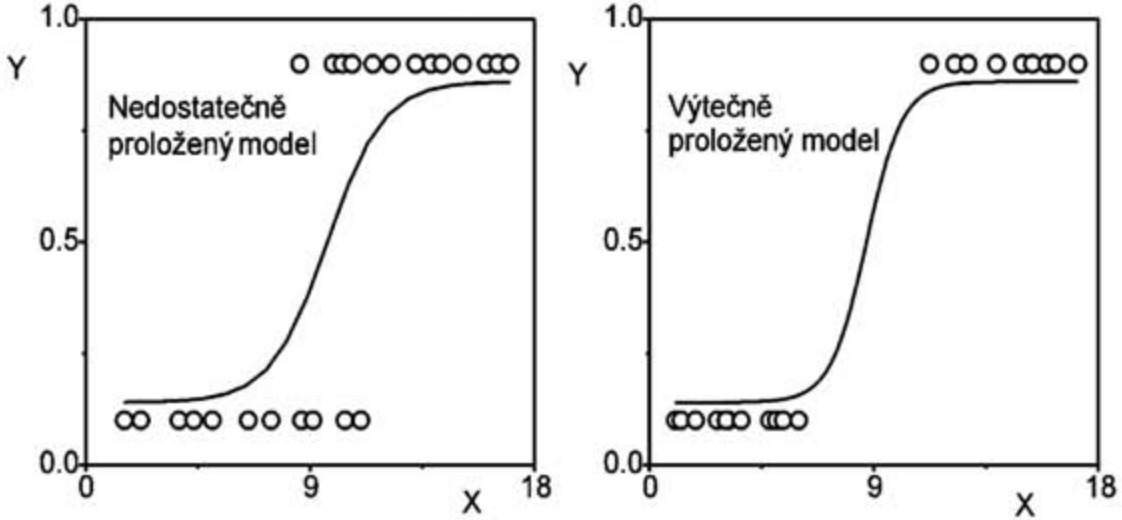
LR

Logistická regrese LR

1. Zaměření metody LR

Navržena v 60tých letech jako alternativní postup k MNČ: závisle proměnná *y* je binární (medicina) značí přítomnost (1) nebo nepřítomnost (0) choroby.

Jde o klasifikaci, když není splněna normalita vícerozměrného modelu.



Krajní případy užití logistického regresního modelu za rozličného proložení.

Rozdíl 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 nezávisle proměnnými $\boldsymbol{x}$.

Při nízkých x se pravděpodobnost proměnné y blíží k nule, při vysokých 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 *y* se rozlišují:

Binární logistická regrese: *binární závisle proměnná* nabývá pouze dvou hodnot, například přítomnost-absence, muž-žena.
Vektor $\boldsymbol{x}$ 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 $\boldsymbol{x}$ 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 $\boldsymbol{x}$ může obsahovat jak ***prediktory***, tak i ***faktory***.

2. Logistický regresní model

Potřebujeme vědět, zda se událost stala (1) nebo nestala (0).

Dichotomická hodnota 0 - 1 závisle proměnné *y* 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).

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

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

$$L_{(1)} = \frac{1}{1 + e^{C - Z}}$$

Pravděpodobnostní poměr (zvaný "*poměr šancí*") je vyjádřen

$$\frac{L(1)}{L(0)} = e^{a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_p x_p}$$

kde odhadované koeficienty $a_0, a_1, a_2, \ldots, a_p$ jsou **míry změny** poměru pravděpodobností $L_{(1)}/L_{(0)}$.

Poměr je lineární funkcí diskriminační funkce o ***p* nezávisle proměnných**

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

Po zlogaritmování a úpravě vyjde $c - z = \ln\left(\frac{L(0)}{L(1)}\right)$, kde C je absolutní člen a_0.

Dle klasifikačního postupu je

$L_{(0)} = P(G = 1 \mid x)$ a

$L_{(1)} = P(G = 1 \mid x) = 1 - P(G = 1 \mid x)$.

a po úpravách bude $\ln\left(\frac{L(1)}{L(0)}\right) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$,

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

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 $\frac{L(1)}{L(0)} = \frac{0.75}{1 - 0.75} = \frac{3}{1}$.

Interpretace regresních koeficientů

Předpoklady o x nejsou a x mohou být **diskrétní** (*faktory*) a **spojité** veličiny (*prediktory*).

$\ln(L_{(1)}/L_{(0)})$ je lineární funkcí nezávisle proměnných a $\ln(L_{(1)}/L_{(0)})$ je nazván **logit** nebo-li **logit transformace pravděpodobnosti.**

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

Logit lze dále 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.

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.

Waldovo testační kritérium $\boldsymbol{W_{a,i} = (b_i/s(b_i))^2}$ vyčísluje statistickou významnost pro odhady regresních koeficientů stejně jako ve vícenásobné regresi.

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

Waldova statistika W_a má nežádoucí vlastnost: pro velikou hodnotu regresního koeficientu b_i a veliký odhad jeho směrodatné odchylky $s(b_i)$ 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.

Kategorické proměnné

Výhodou 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*.

Nejjednodušší situace je **jediný faktor** *x* se dvěma možnými hodnotami.

Pojmu 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 $L(1) = \frac{1}{1 + e^{-a - b x}}$, odhad parametru *a* a odhad *b*.

Odhad *b* představuje přirozený logaritmus pravděpodobnostního poměru žen a mužů.

Odhad *a* je přirozený logaritmus pravděpodobnostního poměru mužů ($x = 0$). 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.

3. Volba proměnných

Například úloha logistické předpovědi infarktu: data jsou z dlouhodobého sledování z počátku zdravých pacientů, u kterých byla dlouhodobě provedena opakovaná měření. Několik jedinců bylo postiženo infarktem. Byl sledován **výběr nezávisle proměnných**, které by mohly odhalit blížící se infarkt.

Výběr **účinných** nezávisle proměnných byl předem lékaři vytypován. Častěji však uživatel předem neví nic o nezávisle proměnných.

Proměnné ***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.

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* = -2*LL, když ***D* představuje** *míru těsnosti proložení dat logistickým 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 G = *D*(model bez proměnné) - *D*(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:

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 $\chi^2_{\exp} = \sum_{i=1}^{\cdot} 2E\left(\ln \frac{E}{V}\right)$, kde suma se provede přes všechny odlišné vzory. Rezidua se sledují právě pro tyto odlišné vzory.

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.

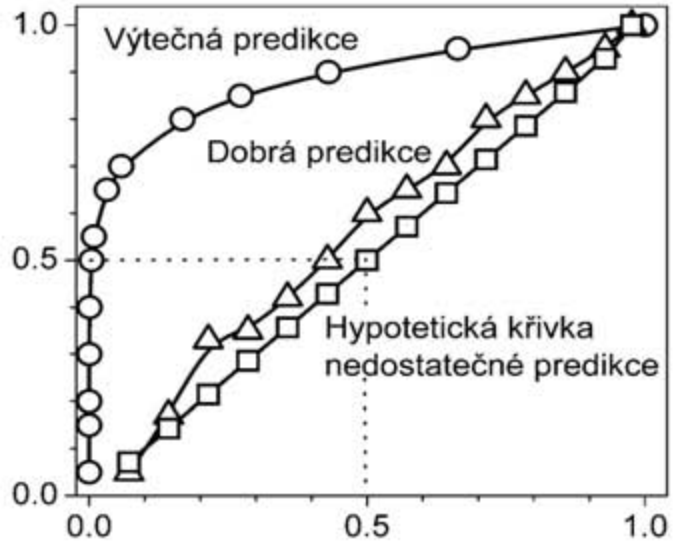
Na ose *y* je procento správně zařazených objektů "v události" nazvané **pozitivní podíl** (a v lékařském výzkumu nazývané **citlivost**).

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 malé podíly nesprávně zařazených objektů se získá vysoké procento správně zařazených objektů, které **skutečně jsou "v události"**.

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

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



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

(a) **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".

(b) **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.

6. Aplikace logistické regrese

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

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.

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.

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 $\boldsymbol{x}$ 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 navrženém logistickém regresním modelu.

Data: $n = 27$, $p = 6$,

Index	*REMISS*	*CELL*	*SMEAR*	*INFIL*	*LI*	*BLAST*	*TEMP*
1	1	0.8	0.83	0.66	1.9	1.1	0.996
2	1	0.9	0.36	0.32	1.4	0.74	0.992
3	0	0.8	0.88	0.7	0.8	0.176	0.982
4	0	1	0.87	0.87	0.7	1.053	0.986
5	1	0.9	0.75	0.68	1.3	0.519	0.98
6	0	1	0.65	0.65	0.6	0.519	0.982
7	1	0.95	0.97	0.92	1	1.23	0.992
8	0	0.95	0.87	0.83	1.9	1.354	1.02
9	0	1	0.45	0.45	0.8	0.322	0.999
10	0	0.95	0.36	0.34	0.5	0	1.038
11	0	0.85	0.39	0.33	0.7	0.279	0.988
12	0	0.7	0.76	0.53	1.2	0.146	0.982
13	0	0.8	0.46	0.37	0.4	0.38	1.006

Index	*REMISS*	*CELL*	*SMEAR*	*INFIL*	*LI*	*BLAST*	*TEMP*
14	0	0.2	0.39	0.08	0.8	0.114	0.99
15	0	1	0.9	0.9	1.1	1.037	0.99
16	1	1	0.84	0.84	1.9	2.064	1.02
17	0	0.65	0.42	0.27	0.5	0.114	1.014
18	0	1	0.75	0.75	1	1.322	1.004
19	0	0.5	0.44	0.22	0.6	0.114	0.99
20	1	1	0.63	0.63	1.1	1.072	0.986
21	0	1	0.33	0.33	0.4	0.176	1.01
22	0	0.9	0.93	0.84	0.6	1.591	1.02
23	1	1	0.58	0.58	1	0.531	1.002
24	0	0.95	0.32	0.3	1.6	0.886	0.988
25	1	1	0.6	0.6	1.7	0.964	0.99
26	1	1	0.69	0.69	0.9	0.398	0.986
27	0	1	0.73	0.73	0.7	0.398	0.986

Řešení: Byl užit program NCSS2000.

1. Odhad regresních koeficientů.

Proměnná	Regresní koefficient	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 $W^2_{a,i} = [b_i / s(b_i)]^2$.

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

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.

		Nalezeno predikcí logistickým modelem		
Dáno závisle proměnnou		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	0.209341
2	Ano (1)	Ne (0)	0.435096	0.564904
3	Ne (0)	Ne (0)	0.155573	-0.155573
4	Ne (0)	Ne (0)	0.290862	-0.290862
5	Ano (1)	Ano (1)	0.697019	0.302981
6	Ne (0)	Ne (0)	0.280081	-0.280081
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)	0.677843	0.322157
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)	0.938464	0.061536
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.

Postup:

Příklad 8.2/str. 235

1. **Otevření souboru dat LEUKEMIA:** Ve **File** menu klikněte na **Open, Data, LEUKEMIA.S0, Open.** Obdržíte zdrojovou matici dat.

NCSS 2007 Data - [C:\Program Files\NCSS2007\Data\Leukemia.s0]

File Edit Data Analysis Graphics PASS GESS Tools Window Help

	REMISS	CELL	SMEAR	INFIL	LI	BLAST	TEMP	RE
1	1	0,8	0,83	0,66	1,9	1,1	0,996	
2	1	0,9	0,36	0,32	1,4	0,74	0,992	
3	0	0,8	0,88	0,7	0,8	0,176	0,982	
4	0	1	0,87	0,87	0,7	1,053	0,986	
5	1	0,9	0,75	0,68	1,3	0,519	0,98	
6	0	1	0,65	0,65	0,6	0,519	0,982	
7	1	0,95	0,97	0,92	1	1,23	0,992	
8	0	0,95	0,87	0,83	1,9	1,354	1,02	
9	0	1	0,45	0,45	0,8	0,322	0,999	
10	0	0,95	0,36	0,34	0,5	0	1,038	
11	0	0,85	0,39	0,33	0,7	0,279	0,988	
12	0	0,7	0,76	0,53	1,2	0,146	0,982	
13	0	0,8	0,46	0,37	0,4	0,38	1,006	
14	0	0,2	0,39	0,08	0,8	0,114	0,99	
15	0	1	0,9	0,9	1,1	1,037	0,99	
16	1	1	0,84	0,84	1,9	2,064	1,02	
17	0	0,65	0,42	0,27	0,5	0,114	1,014	
18	0	1	0,75	0,75	1	1,322	1,004	
19	0	0,5	0,44	0,22	0,6	0,114	0,99	
20	1	1	0,63	0,63	1,1	1,072	0,986	
21	0	1	0,33	0,33	0,4	0,176	1,01	
22	0	0,9	0,93	0,84	0,6	1,591	1,02	
23	1	1	0,58	0,58	1	0,531	1,002	
24	0	0,95	0,32	0,3	1,6	0,886	0,988	
25	1	1	0,6	0,6	1,7	0,964	0,99	
26	1	1	0,69	0,69	0,9	0,398	0,986	
27	0	1	0,73	0,73	0,7	0,398	0,986	
28		1			0,7		0,986	
29		1			0,7		0,986	
30								

2. Volba metody: V menu vyberte **Analysis, Regression/Correlation, Logistic Regression.**

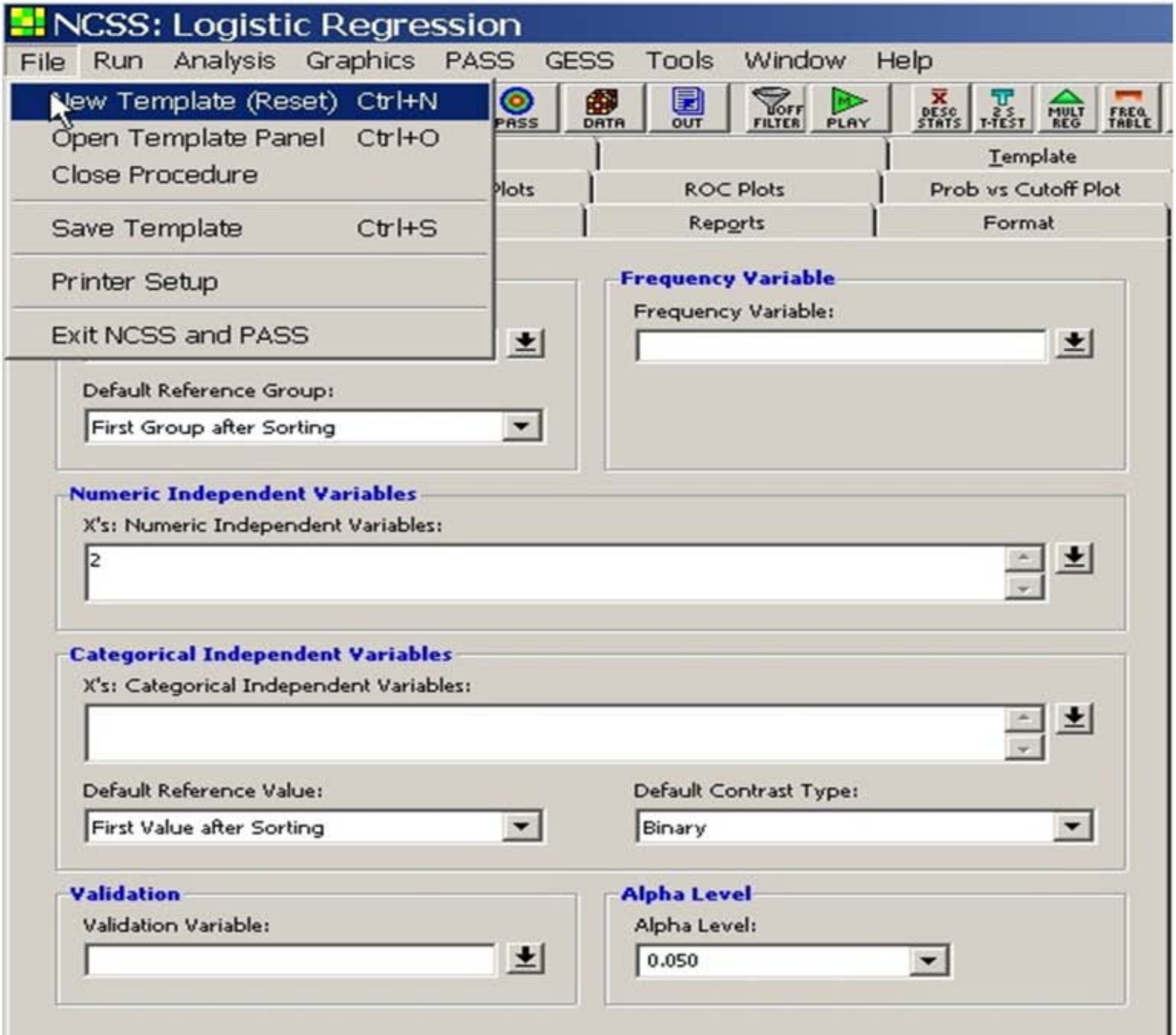
NCSS 2007 Data - [C:\Program Files\NCSS2007\Data\Leukemia.s0]

File Edit Data Analysis Graphics PASS GESS Tools Window Help

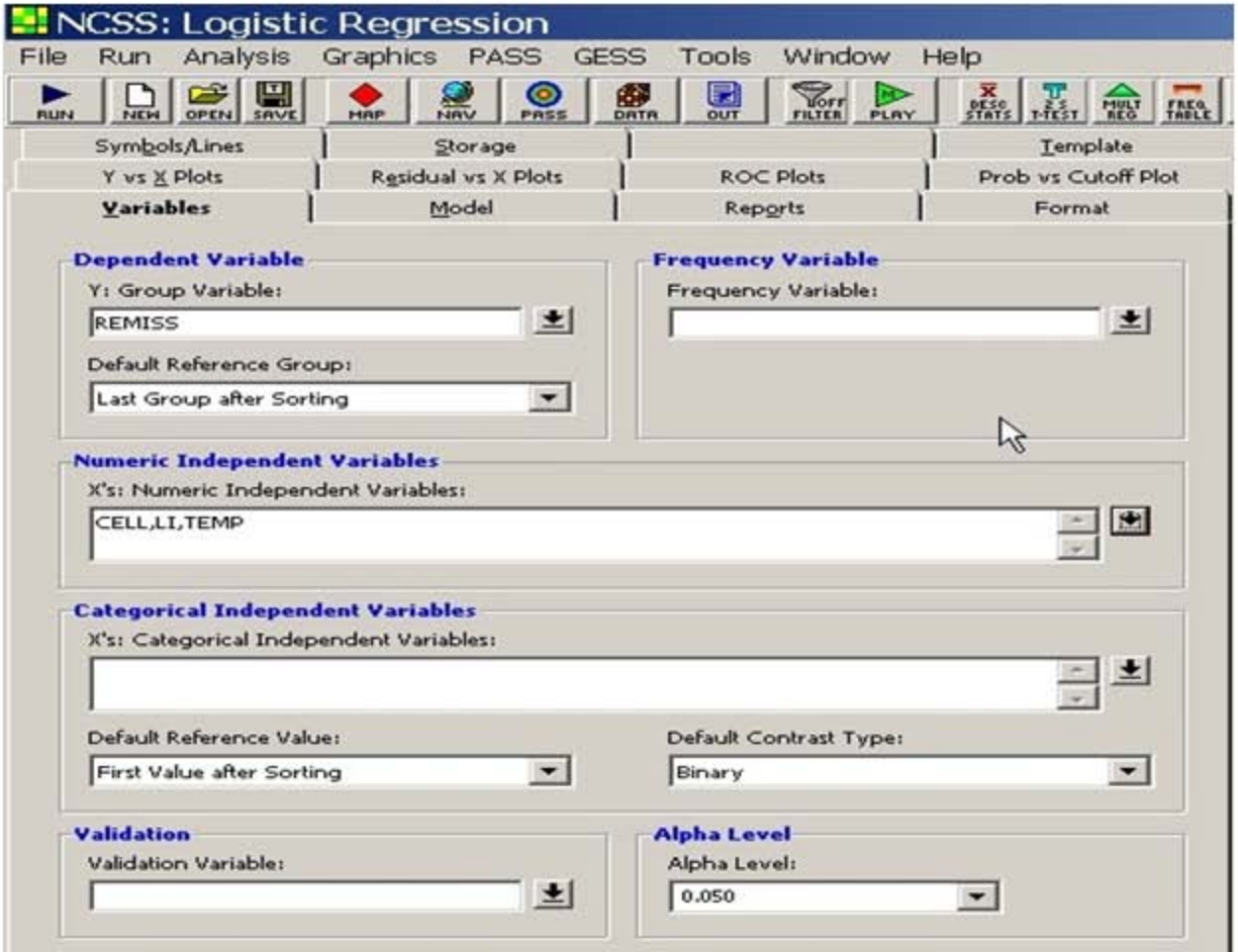
- Analysis of Variance (ANOVA)
- Clustering
- Curve Fitting
- Descriptive Statistics
- Design of Experiments
- Diagnostic Tests
- Forecasting / Time Series
- Mass Appraisal
- Meta-Analysis
- Mixed Models
- Multivariate Analysis
- Nondetects
- Operations Research
- Proportions
- Quality Control
- Regression / Correlation
  - Linear Regression and Correlation
  - Multiple Regression
  - Cox Regression
  - Logistic Regression
  - Correlation Routines
  - Nonlinear Routines
  - Other Regression Routines
  - Variable Selection Routines
  - Legacy Routines
- Survival / Reliability
- T-Tests
- Other

	REMISS			LI	BLAST	TEMP
1	1		0,66	1,9	1,1	0,996
2	1		0,32	1,4	0,74	0,992
3	0		0,7	0,8	0,176	0,982
4	0		0,87	0,7	1,053	0,986
5	1		0,68	1,3	0,519	0,98
6	0		0,65	0,6	0,519	0,982
7	1		0,92	1	1,23	0,992
8	0		0,83	1,9	1,354	1,02
9	0		0,45	0,8	0,322	0,999
10	0		0,34	0,5	0	1,038
11	0		0,33	0,7	0,279	0,988
12	0		0,53	1,2	0,146	0,982
13	0		0,37	0,4	0,38	1,006
14	0		0,08	0,8	0,114	0,99
15	0				,037	0,99
16	1				,064	1,02
17	0				,114	1,014
18	0				,322	1,004
19	0				,114	0,99
20	1	1	0,63		,072	0,986
21	0	1	0,33		,176	1,01
22	0	0,9	0,93		,591	1,02
23	1	1	0,58		,531	1,002
24	0	0,95	0,32		,886	0,988
25	1	1	0,6		,964	0,99
26	1	1	0,69		,398	0,986
27	0	1	0,73	0,73	0,7	0,398	0,986
28		1			0,7		0,986
29		1			0,7		0,986
30						

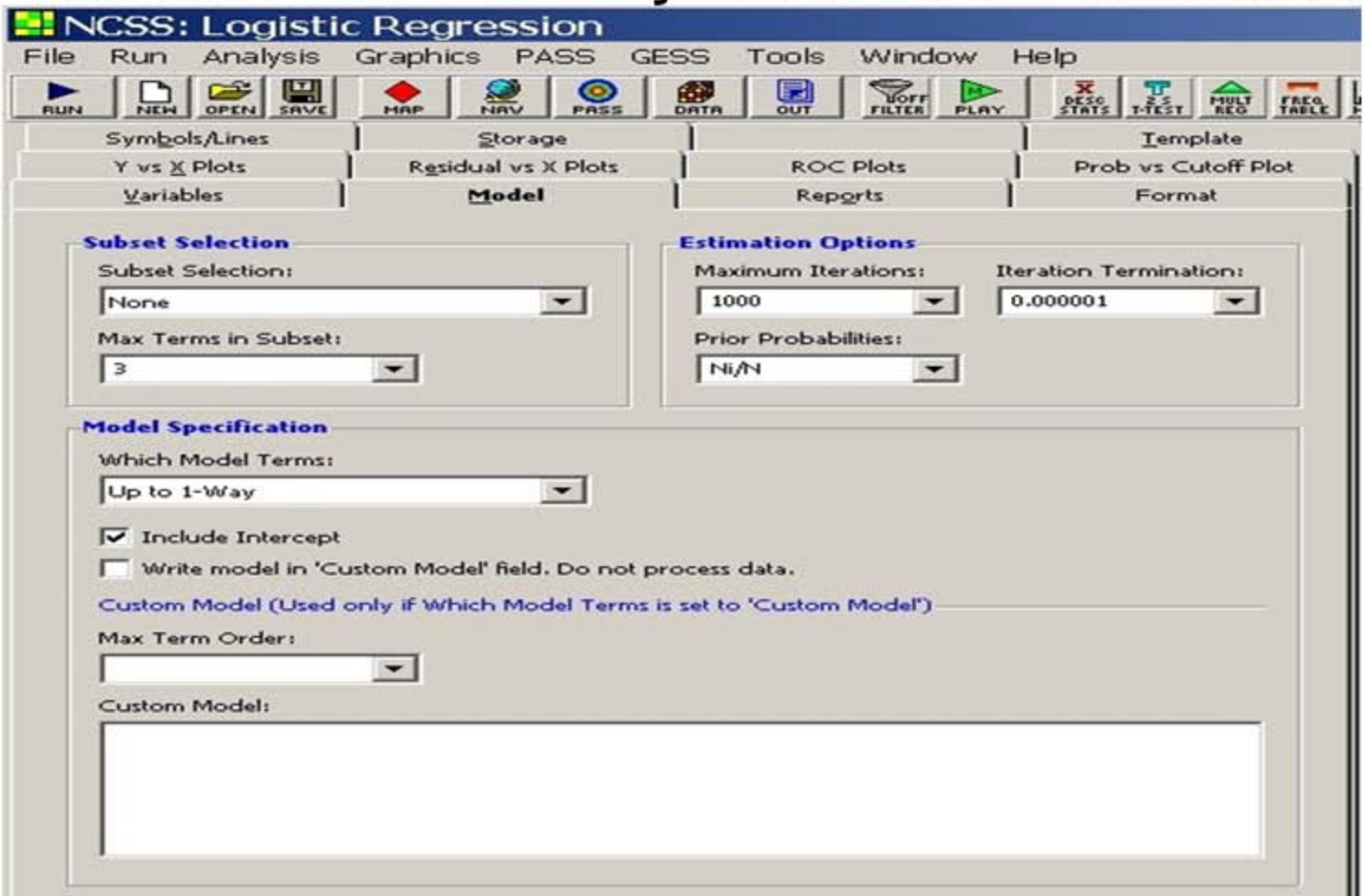
3. Zadání nové šablony inputu dat: Kliknete na **File** a zvolíte **New Template.**



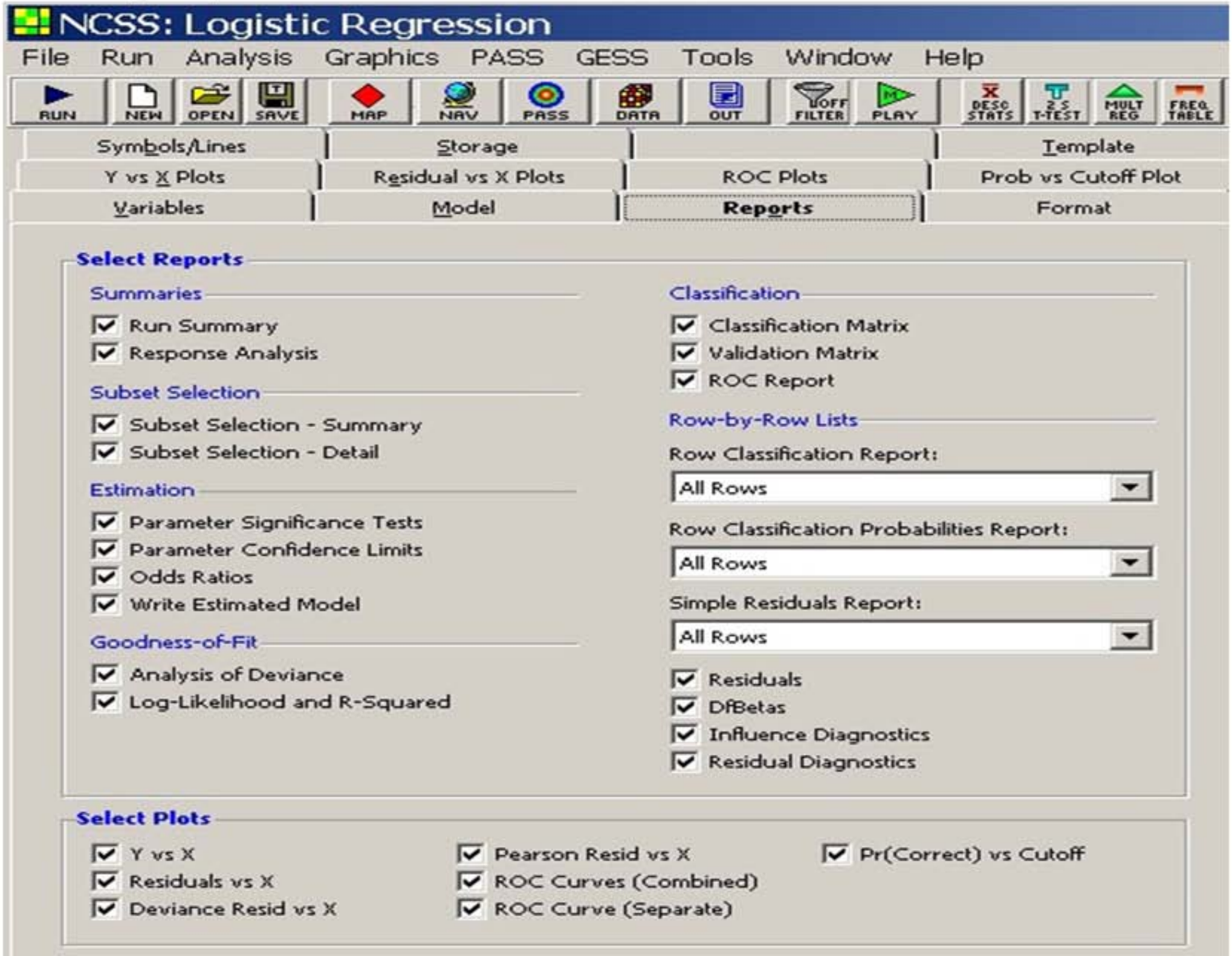
4. Zadání závisle proměnné *y*: Vyberte záložku **Variables**, kde v okénku **Y: Group Variable** zadáte proměnnou **REMISS** a v **Default Reference Group** zadáte **Last Group after Sorting.** V okénku numerické nezávisle proměnné **X´s:Numeric Independent Variables** zadáte jenom vybrané tři nezávisle proměnné znaky **CELL, LI, TEMP**. Při zadávání těchto spojitých proměnných podržíte klávesu Ctrl. Ostatní okénka jsou na předvolených hodnotách nebo prázdná.



5. Vložení modelu: Zadáte logistický regresní model kliknutím na záložku **Model**. V okénku **Subset Selection** dáte volbu **None** a v okénku **Max Terms in Subset** pak číslo 3. Dále v okénku **Which Model Terms** zvolíte **Up to 1-Way.** Zaškrtnete **Include Intercept,** protože požadujete v regresním modelu také absolutní člen. V **Maximum Iteration** dáte doporučenou hodnotu **1000**, v **Iteration Termination** kritérium **1E-7** a konečně v **Prior Probability** zvolíte **Ni/N**. Ostatní okénka necháte volná.



6. Zadání protokolu: Vyberte záložku **Reports**. Zaškrtnete všechny volby pro bohatý output.



7. Proběhne výpočet: Kliknutím na **Run Procedure** se spustí výpočet.

8. Komentář k výstupu: Je nejprve uveden přehled zadaných podmínek výpočtu:

Logistic Regression Report

Page/Date/Time	1 20.7.2009 7:49:55
Database	C:\Program Files\NCSS2007\Data\Leukemia.S0
Response	REMISS

Run Summary Section

Parameter	Value	Parameter	Value
Dependent Variable	REMISS	Rows Processed	29
Reference Group	1	Rows Used	27
Number of Groups	2	Rows for Validation	0
Frequency Variable	None	Rows X's Missing	0
Numeric Ind. Variables	3	Rows Freq Miss. or 0	0
Categorical Ind. Variables	0	Rows Prediction Only	2
Final Log Likelihood	-10,97669	Unique Row Patterns	28
Model R-Squared	0,36130	Sum of Frequencies	27
Actual Convergence	2,94261E-07	Likelihood Iterations	7
Target Convergence	0,000001	Maximum Iterations	1000
Model D.F.	4	Max Like Message	Normal Completion
Model	CELL|LI|TEMP		

Run Summary Section ukazuje zda podmínky výpočty byly zadány vhodně: **Reference Group** je jedna z hodnot (čili kategorie) predikované závisle proměnné. Tato však bývá v outputu vynechávána. Když není **Reference group** u proměnné Y specifikována uživatelem, je zvolena defaultně dle **Default Reference Group.** Je důležitá pro výklad celého zbytku outputu.

Number of Groups značí počet kategorií čili jednotek závisle proměnné Y, obvykle 0 a 1, takže 2.

Final Log Likelihood značí dosaženou hodnotu logaritmu věrohodnosti po terminaci iteračního výpočtu.

Model R-Squared značí hodnotu R^2 po ukončeném regresním výpočtu.

Actual and Target Convergence značí jednak dosaženou a jednak ukončovací hodnotu při terminaci regresního procesu, kdy je dosaženo maximální věrohodnosti. Výpočet by neměl terminovat, pokud není dosažená hodnota nižší než předvolená ukončovací.

Rows Processed, Used, etc. značí počet řádků zdrojové matice (=objektů, pacientů, pozorování), které byly načteny a použity k výpočtu.

Unique Row Patterns značí počet jednotkových předloh nalezených v obou, závisle a nezávisle proměnných.

Likelihood and Maximum Iterations značí počet iterací k dosažení terminace a ukončení a počet předvolených iterací. Počet iterací nižší jak 10 bývá obvyklý při úspěšném výpočtu.

Max Like Message přináší návěští při terminaci výpočtu. Návěští „Normal Competion“ je po úspěšně proběhlém výpočtu.

Model D.F. značí počet stupňů volnosti v G-1 logistických regresních modelech.

Model značí zkratkový zápis regresního modelu formou členů nezávisle proměnných, který byl proložen načtenými daty.

Blok **Response Analysis Section:**

Response Analysis Section

REMISS Categories	Count	Unique Rows	Prior	Act vs Pred R-Squared	% Correctly Classified
0	18	17	0,62069		83,333
1	9	9	0,31034	0,39181	66,667
Total	27	26			77,778

Tento blok statisticky popisuje závisle proměnnou Y a ukazuje, jak dokonale je závisle proměnná Y aproximována svým zde navrženým regresním modelem.

Categories značí jednotkové hodnoty obou kategorií, které může závisle proměnná Y nabývat. U binárních dat jde vždy o dvě hodnoty, a to 0 a 1.

Count značí sumu četností řádků objektů, které může každá z kategorií závisle proměnné Y nabývat.

Unique Rows značí počet jednotkových řádků v každé kategorii, které byly vypočteny z nezávisle proměnných.

Prior značí úvodní pravděpodobnost každé kategorie zadané uživatelem ve volbě v okénku **Prior Probabilities.**

Act vs Před R-Squared značí hodnotu R^2, která je dosažena po regresi, když jsou nezávisle proměnné pro tuto kategorii regresovány na predikovanou pravděpodobnost.

%Correctly Classified značí procento objektů (zde pacientů) v každé kategorii, kteří byli správně klasifikováni, když by to byl vícerozměrný logist. regr. model.

Blok **Parameter Significance Tests Section:**

Parameter Significance Tests Section (Reference Group: REMISS = 1)

Parameter	Regression Coefficient (B or Beta)	Standard Error	Wald Z-Value (Beta=0)	Wald Prob Level	Odds Ratio Exp(B)
B0: Intercept	-67,63381	56,88751	-1,189	0,23448	0,00000
B1: CELL	-9,65213	7,75107	-1,245	0,21303	0,00006
B2: LI	-3,86710	1,77828	-2,175	0,02966	0,02092
B3: TEMP	82,07365	61,71233	1,330	0,18354	10000+

Blok obsahuje odhady logistických regresních rovnic a s nimi spojené statistické testy významnosti. Referenční hodnota je uvedena hned v nadpisu, zde to je REMISS=1. Má-li závisle proměnná více jak dvě možné kategorie, je tato tabulka vyčíslena pro celkem G-1 rovnic.

Parameter je označení odhadovaného parametru. Spojitý parametr je zobrazen přímo. Je-li však diskrétní, je zde uvedena definice binární proměnné generované. Např. diskrétní nezávisle proměnná GRADE nabývá tří hodnot: A, B, C. Název zde uvedený bude mít třeba podobu B2:GRADE=B. To pak odkazuje na binární proměnnou, u které bude jednička pro řádky, ve kterých je GRADE rovna B zatímco nula bude u ostatních. Všimněte si, že umístění jména je ovládáno v záložce **Format** ve volbě **Skip Line After.**

Parameter Significance Tests Section (Reference Group: REMISS = 1)

Parameter	Regression Coefficient (B or Beta)	Standard Error	Wald Z-Value (Beta=0)	Wald Prob Level	Odds Ratio Exp(B)
B0: Intercept	-67,63381	56,88751	-1,189	0,23448	0,00000
B1: CELL	-9,65213	7,75107	-1,245	0,21303	0,00006
B2: LI	-3,86710	1,77828	-2,175	0,02966	0,02092
B3: TEMP	82,07365	61,71233	1,330	0,18354	10000+

Regression Coefficient (B or Beta) značí číselný odhad odpovídajícího regresního koeficientu, někdy značeného B jindy Beta.

Standard Error značí míru přesnosti odhadovaného regresního koeficientu čili jeho odhad směrodatné odchylky. Užije se dále ve jmenovateli Waldova testu.

Wald Z-Value (Beta=0) značí Waldovo kritérium k testu hypotézy H_0: $B_j=0$ ve formě z_j= odhad parametru/směrodatná odchylka.

Wald Prob Level značí spočtenou hladinu významnosti **p** pro Waldovo kritérium. Je-li **p** menší než **alfa**=0.05, je parametr statisticky významný.

Odds Ratio Exp(B) značí **OR** odhad procenta šancí spojeného s tímto regresním koeficientem. Je užitečný pouze pro binární nezávisle proměnné, které mohou nabývat dvou hodnot, 0 a 1. Jde o hodnoty, které jsou vyčísleny pro kategorické nezávisle proměnné dle vztahu **OR**=exp(B).

Blok **Parameter Confidence Limits Section:**

Parameter Confidence Limits Section (Reference Group: REMISS = 1)

Parameter	Regression Coefficient (B or Beta)	Standard Error	Lower 95% Confidence Limit	Upper 95% Confidence Limit	Odds Ratio Exp(B)
B0: Intercept	-67,63381	56,88751	-179,13128	43,86365	0,00000
B1: CELL	-9,65213	7,75107	-24,84394	5,53968	0,00006
B2: LI	-3,86710	1,77828	-7,35245	-0,38174	0,02092
B3: TEMP	82,07365	61,71233	-38,88029	203,02760	10000+

Tento blok obsahuje odhady logistických regresních rovnic a s nimi spojené intervaly spolehlivosti parametrů. Referenční hodnota je uvedena hned v nadpisu, zde REMISS=1. Má-li závisle proměnná více jak dvě možné kategorie, je tato tabulka uvedena pro celkem G-1 rovnic.

Parameter je označení odhadovaného parametru. Spojitý parametr je zobrazen přímo. Je-li však diskrétní, je zde uvedena definice binární proměnné generované. Např. diskrétní nezávisle proměnná GRADE nabývá tří hodnot: A, B, C. Název zde uvedený bude mít třeba podobu B2:GRADE=B. To pak odkazuje na binární proměnnou, u které bude jednička pro řádky, ve kterých je GRADE rovna B zatímco nula bude u ostatních. Všimněte si, že umístění jména je ovládáno v záložce **Format** ve volbě **Skip Line After.**

Parameter Confidence Limits Section (Reference Group: REMISS = 1)

Parameter	Regression Coefficient (B or Beta)	Standard Error	Lower 95% Confidence Limit	Upper 95% Confidence Limit	Odds Ratio Exp(B)
B0: Intercept	-67,63381	56,88751	-179,13128	43,86365	0,00000
B1: CELL	-9,65213	7,75107	-24,84394	5,53968	0,00006
B2: LI	-3,86710	1,77828	-7,35245	-0,38174	0,02092
B3: TEMP	82,07365	61,71233	-38,88029	203,02760	10000+

Regression Coefficient (B or Beta) je číselný odhad regresního koeficientu. **Standard Error** značí míru přesnosti regresního koeficientu čili jeho odhad směrodatné odchylky. Užije se dále ve jmenovateli Waldova testu.

Confidence Limits značí dolní a horní meze intervalového odhadu neznámého koeficientu pro každou nezávisle proměnnou. Jeli dolní mez záporná a horní mez kladná, leží mezi mezemi nula a koeficient je proto nulový čili statisticky nevýznamný. Meze jsou platné stejně jako Waldův test jenom pro velké výběry zdrojové matice dat.

Odds Ratio Exp(B) značí **OR** odhad procenta šancí spojeného s tímto regresním koeficientem. Je užitečný pouze pro binární nezávisle proměnné, které mohou nabývat dvou hodnot, 0 a 1. Jde o hodnoty, které jsou vyčísleny pro kategorické nezávisle proměnné dle vztahu **OR**=exp(B). Kvůli formátovacím problémům tato hodnota se nezobrazí nabyde-li hodnoty větší než 10.000.

Blok **Odds Ratio Estimation Section:**

Odds Ratios Section (Reference Group: REMISS = 1)

Parameter	Regression Coefficient (B or Beta)	Odds Ratio Exp(B)	Lower 95% Confidence Limit	Upper 95% Confidence Limit
B0: Intercept	-67,63381	0,00000	0,00000	10000+
B1: CELL	-9,65213	0,00006	0,00000	254,59635
B2: LI	-3,86710	0,02092	0,00064	0,68267
B3: TEMP	82,07365	10000+	0,00000	10000+

Tento blok obsahuje procento šance a jeho intervaly spolehlivosti.

Parameter je název odhadovaného parametru. Spojitý parametr je zobrazen přímo. Je-li však parametr diskrétní, je zde uvedena definice binární proměnné generované. Např. diskrétní nezávisle proměnná GRADE nabývá tří hodnot: A, B, C. Název zde uvedený bude mít třeba podobu B2:GRADE=B. To pak odkazuje na binární proměnnou, u které bude jednička pro řádky, ve kterých je GRADE rovna B zatímco nula bude u ostatních. Všimněte si, že umístění jména je ovládáno v záložce **Format** ve volbě **Skip Line After.**

Odds Ratio Exp(B) značí **OR** odhad procenta šancí spojeného s tímto regresním koeficientem. Je užitečný pouze pro binární nezávisle proměnné, které mohou nabývat dvou hodnot, 0 a 1. Jde o hodnoty, které jsou vyčísleny pro kategorické nezávisle proměnné dle vztahu **OR**=exp(B).

Confidence Limits značí dolní a horní meze intervalu spolehlivosti procenta šance. Platí podobně jako Waldovo kritérium převážně pro velké výběry. Kvůli formátovacím problémům tato hodnota se nezobrazí nabude-li hodnoty větší než 10.000.

Blok **Estimated Logistic Regression Model(s) Section:**

Estimated Logistic Regression Model(s)

Model For REMISS = 0

-67.6338131249461 -9.65212973758*CELL -3.86709587172718*LI +82.0736535775612*TEMP

Note that each model estimates B for a specific group, where Logit(Y) = XB.
To calculate a probability, transform the logit using Prob(Y<>group) = 1/(1+Exp(-XB))
or Prob(Y=group) = Exp(-XB)/(1+Exp(-XB)).

Transformation Note:
Regular transformations must be less the 255 characters. If this expression is longer the 255 characters, copy this expression and paste it into a text file, then use the transformation FILE(filename.txt) to access the text file.

Rovnice logistického regresního modelu je v normálním textovém formátu a může být proto použita jako transformační vzorec.

Regresní rovnice je zobrazena pro každou ze G-1 kategorií závisle proměnné.

Regresní koeficienty jsou zobrazeny v double precision negarantují však nutou přesnost k výpočtu skore (logitu) a predikovaných pravděpodobností.

Blok **Analysis of Deviance Section:**

Analysis of Deviance Section

Term Omitted	DF	Deviance	Increase From Model Deviance (Chi Square)	Prob Level
All	3	34,37177	12,41839	0,00608
CELL	1	24,64782	2,69445	0,10070
LI	1	30,82856	8,87518	0,00289
TEMP	1	24,34072	2,38734	0,12232
None(Model)	3	21,95337		

The Prob Level is for testing the significance of that term after considering all other terms.

Blok je podobný tabulce ANOVA. Přináší výsledky Chi-kvadrát testu k ověření, zda každý jednotlivý člen v regresní rovnice je statisticky významný po regresním nastavení všech ostatních členů v modelu. Je-li výpočet příliš dlouhý, tento blok se vynechává, protože je časově náročný.

Term Omitted značí člen regresní rovnice, který je právě testován tím, že je zde vynechán. Test je postaven na srovnání odchylky pro model bez tohoto členu vůči odchylce kompletního regresního modelu. Zde řádek „All“ značí, že model obsahuje pouze absolutní člen a tento řádek proto testuje plný model. Řádek „None(Model)“ značí plný model, u kterého zde nebyl odstraněn žádný člen. Všimněte si, že není vhodně do modelu zařadit interakční člen, když jeden z hlavních efektů dosud schází - což je případ také zde. Přesto však věříme, že jde o velice užitečný test.

Blok **Analysis of Deviance Section:**

Analysis of Deviance Section

Term Omitted	DF	Deviance	Increase From Model Deviance (Chi Square)	Prob Level
All	3	34,37177	12,41839	0,00608
CELL	1	24,64782	2,69445	0,10070
LI	1	30,82856	8,87518	0,00289
TEMP	1	24,34072	2,38734	0,12232
None(Model)	3	21,95337		

The Prob Level is for testing the significance of that term after considering all other terms.

DF značí počet stupňů volnosti Chi-square testu v tomto řádku a rovná se (G-1)DFt, kde Dft je počet stupňů volnosti členu.

Deviance D je -2log věrohodnosti modelu v řádku a značí se také **D=-2LL.**

Increase From Model Deviance (Chi Square) značí rozdíl v odchylce mezi modelem na tomto řádku a celkovým plným modelem a značí se obvykle **G.** Je to test redundance (nadbytečnosti) tohoto členu při uvažování všech členů regresního modelu.

Prob Level značí spočtenou hladinu významnosti **p**. Je-li **p** menší než zadaná hladina významnosti **alfa** =0.05, je člen statisticky významný.

Blok **Log Likelihood &R-Squared Section:**

Log Likelihood & R-Squared Section

Term(s) Omitted	DF	Log Likelihood	R-Squared of Remaining Term(s)	Reduction From Model R-Squared	Reduction From Saturated R-Squared
All	1	-17,18588	0,00000		
CELL	1	-12,32391	0,28290	0,07839	0,71710
LI	1	-15,41428	0,10308	0,25821	0,89692
TEMP	1	-12,17036	0,29184	0,06946	0,70816
None(Model)	3	-10,97669	0,36130	0,00000	0,63870
None(Saturated)	28	0,00000	1,00000		0,00000

Blok poskytuje log věrohodnosti a R^2 pro rozličné modely. Výpočet je časově náročnější a bývá proto většinou vypuštěn.

Term Omitted značí člen regresní rovnice, který je právě testován tím, že je vypušten. Test je postaven na srovnání odchylky pro model bez tohoto členu vůči odchylce kompletního regresního modelu. V řádku zápis „All“ značí, že model obsahuje pouze absolutní člen a tento řádek testuje vlastně plný model. V řádku zápis „None(Model)“ značí plný model, u kterého nebyl odstraněn žádný člen. Všimněte si, že není vhodně do modelu zařadit interakční člen, když jeden z hlavních efektů dosud schází - což je případ také zde. Přesto však věříme, že jde o velice užitečný test.

Blok **Analysis of Deviance Section:**

Analysis of Deviance Section

Term Omitted	DF	Deviance	Increase From Model Deviance (Chi Square)	Prob Level
All	3	34,37177	12,41839	0,00608
CELL	1	24,64782	2,69445	0,10070
LI	1	30,82856	8,87518	0,00289
TEMP	1	24,34072	2,38734	0,12232
None(Model)	3	21,95337		

The Prob Level is for testing the significance of that term after considering all other terms.

DF značí počet stupňů volnosti Chi-square testu zobrazeného v tomto řádku. DF se rovná (G-1)DFt, kde Dft je počet stupňů volnosti členu.

Deviance D je -2log věrohodnosti modelu na tomto řádku a značí se **D=-2LL.**

Increase From Model Deviance (Chi Square) značí rozdíl v odchylce mezi modelem na tomto řádku a celkovým plným modelem, značí se obvykle **G.** Jinými slovy jde o test redundance (nadbytečnosti) tohoto členu při uvažování všech členů regresního modelu.

Prob Level značí spočtenou hladinu významnosti **p**. Je-li **p** menší než zadaná hladina významnosti **alfa** =0.05, je člen statisticky významný.

Blok **Classification Table:**

Classification Table

	Estimated		
Actual	**0**	**1**	**Total**
0	15	3	18
1	3	6	9
Total	18	9	27

Percent Correctly classified = 77,8%

Blok ukazuje výsledky klasifikace dat na základě logistických regresních rovnic. Tabulka ukazuje počty pro jednotlivé kategorie závisle proměnné Y.

Je uvedeno také procento správně klasifikovaných objektů, což je vlastně procento z celkového počtu, které padnou na diagonálu této tabulky.

ROC diagram pro kategorii rovnou 0

ROC Section for Value 0

Prob Cutoff	N(1|1) A	N(1|0) B	N(0|1) C	N(0|0) D	Sensitivity A/(A+C)	Specificity D/(B+D)	Sensitivity +Specificity	Proportion Correct
0,05000	18	9	0	0	1,00000	0,00000	1,00000	0,66667
0,10000	18	8	0	1	1,00000	0,11111	1,11111	0,70370
0,15000	17	8	1	1	0,94444	0,11111	1,05556	0,66667
0,20000	17	8	1	1	0,94444	0,11111	1,05556	0,66667
0,25000	17	8	1	1	0,94444	0,11111	1,05556	0,66667
0,30000	17	5	1	4	0,94444	0,44444	1,38889	0,77778
0,35000	17	5	1	4	0,94444	0,44444	1,38889	0,77778
0,40000	16	4	2	5	0,88889	0,55556	1,44444	0,77778
0,45000	15	3	3	6	0,83333	0,66667	1,50000	0,77778
0,50000	15	3	3	6	0,83333	0,66667	1,50000	0,77778
0,55000	15	2	3	7	0,83333	0,77778	1,61111	0,81481
0,60000	15	2	3	7	0,83333	0,77778	1,61111	0,81481
0,65000	15	2	3	7	0,83333	0,77778	1,61111	0,81481
0,70000	15	1	3	8	0,83333	0,88889	1,72222	0,85185
0,75000	12	0	6	9	0,66667	1,00000	1,66667	0,77778
0,80000	11	0	7	9	0,61111	1,00000	1,61111	0,74074
0,85000	9	0	9	9	0,50000	1,00000	1,50000	0,66667
0,90000	8	0	10	9	0,44444	1,00000	1,44444	0,62963
0,95000	7	0	11	9	0,38889	1,00000	1,38889	0,59259

Area Under ROC Curve = 0,89198

ROC diagram je vyčíslen pro každou kategorii závisle proměnné Y, obvykle pro 2, pro 0 a 1.

Křivky grafu prahové operační charakteristiky ROC se užijí k určení vhodného **prahového bodu** (**Cutoff**) ke klasifikaci objektů a umožní porovnat sensitivitu a specificitu pro rozličné prahové body.

Při klasifikaci objektů je kritériem zařazení řádku do kategorie nejvyšší hodnota pravděpodobnosti klasifikovaného.

ROC analýza je nejúčinnější pro dvou-skupinový případ. Pro multi-skupinový případ jde pouze o okrajovou užitečnost, protože prahový bod není specifikován. Zde je potom každý objekt klasifikován dle nejvyšší skupinové hodnoty pravděpodobnosti.

ROC diagram pro kategorii rovnou 0

ROC Section for Value 0

Prob Cutoff	N(1|1) A	N(1|0) B	N(0|1) C	N(0|0) D	Sensitivity A/(A+C)	Specificity D/(B+D)	Sensitivity +Specificity	Proportion Correct
0,05000	18	9	0	0	1,00000	0,00000	1,00000	0,66667
0,10000	18	8	0	1	1,00000	0,11111	1,11111	0,70370
0,15000	17	8	1	1	0,94444	0,11111	1,05556	0,66667
0,20000	17	8	1	1	0,94444	0,11111	1,05556	0,66667
0,25000	17	8	1	1	0,94444	0,11111	1,05556	0,66667
0,30000	17	5	1	4	0,94444	0,44444	1,38889	0,77778
0,35000	17	5	1	4	0,94444	0,44444	1,38889	0,77778
0,40000	16	4	2	5	0,88889	0,55556	1,44444	0,77778
0,45000	15	3	3	6	0,83333	0,66667	1,50000	0,77778
0,50000	15	3	3	6	0,83333	0,66667	1,50000	0,77778
0,55000	15	2	3	7	0,83333	0,77778	1,61111	0,81481
0,60000	15	2	3	7	0,83333	0,77778	1,61111	0,81481
0,65000	15	2	3	7	0,83333	0,77778	1,61111	0,81481
0,70000	15	1	3	8	0,83333	0,88889	1,72222	0,85185
0,75000	12	0	6	9	0,66667	1,00000	1,66667	0,77778
0,80000	11	0	7	9	0,61111	1,00000	1,61111	0,74074
0,85000	9	0	9	9	0,50000	1,00000	1,50000	0,66667
0,90000	8	0	10	9	0,44444	1,00000	1,44444	0,62963
0,95000	7	0	11	9	0,38889	1,00000	1,38889	0,59259

Area Under ROC Curve = 0,89198

Klasifikace objektu může nabývat 4 výsledků:

1) Objekt z dané skupiny je správně zařazen (A) nebo nesprávně, že je z jiné skupiny (C).

2) Objekt z jiné skupiny je nesprávně zařazen, že je z této skupiny (B) nebo správně zařazen, že je z jiné skupiny (D).

Počet objektů v každém stavu je vyčíslen pro každou hodnotu prahového bodu v rozmezí 0 až 1.

Sensitivity je podíl objektů této dané skupiny správně zařazených do této, A/(A+C).

Specificity je podíl objektů jiných skupin než této dané, které jsou správně zařazeny do jiných, D/(B+D).

Optimální prahový bod nabývá maximální hodnoty součtu sensitivity a specificity.

ROC diagram pro kategorii rovnou 1

ROC Section for Value 1

Prob Cutoff	N(1|1) A	N(1|0) B	N(0|1) C	N(0|0) D	Sensitivity A/(A+C)	Specificity D/(B+D)	Sensitivity +Specificity	Proportion Correct
0,05000	9	11	0	7	1,00000	0,38889	1,38889	0,59259
0,10000	9	10	0	8	1,00000	0,44444	1,44444	0,62963
0,15000	9	9	0	9	1,00000	0,50000	1,50000	0,66667
0,20000	9	7	0	11	1,00000	0,61111	1,61111	0,74074
0,25000	9	6	0	12	1,00000	0,66667	1,66667	0,77778
0,30000	8	3	1	15	0,88889	0,83333	1,72222	0,85185
0,35000	7	3	2	15	0,77778	0,83333	1,61111	0,81481
0,40000	7	3	2	15	0,77778	0,83333	1,61111	0,81481
0,45000	7	3	2	15	0,77778	0,83333	1,61111	0,81481
0,50000	6	3	3	15	0,66667	0,83333	1,50000	0,77778
0,55000	6	3	3	15	0,66667	0,83333	1,50000	0,77778
0,60000	5	2	4	16	0,55556	0,88889	1,44444	0,77778
0,65000	4	1	5	17	0,44444	0,94444	1,38889	0,77778
0,70000	4	1	5	17	0,44444	0,94444	1,38889	0,77778
0,75000	1	1	8	17	0,11111	0,94444	1,05556	0,66667
0,80000	1	1	8	17	0,11111	0,94444	1,05556	0,66667
0,85000	1	1	8	17	0,11111	0,94444	1,05556	0,66667
0,90000	1	0	8	18	0,11111	1,00000	1,11111	0,70370
0,95000	0	0	9	18	0,00000	1,00000	1,00000	0,66667

Area Under ROC Curve = 0,89198

Prob Cutoff značí hodnotu pravděpodobnosti prahového bodu ke klasifikaci do této skupiny: je-li predikovaná pravdě-podobnost objektu k zařazení do této skupiny větší než tato hodnota, objekt je zařazen do této skupiny. Jinak je zařazen do jiné skupiny.

A B C D značí počty pro každou ze čtyř stavů.V těchto počtech je zavedeno označení tak, že *i* značí klasifikovanou skupinu a *j* značí aktuální skupinu.

Sensitivity je podíl objektů této dané skupiny správně zařazených do této, A/(A+C).
Specificity je podíl objektů jiných skupin než této dané, které jsou správně zařazeny do jiných, D/(B+D).

Sensitivity+Specificity značí společné pravidlo k hledání nejlepšího prahového bodu a říká, že jde o takový prahový bod, který vykazuje největší hodnotu součtu obou.

Proportion Correct značí pravidlo hledání nejlepšího prahového bodu: je to takový bod, který maximalizuje počet objektů, které jsou správně klasifikovány.

Area Under ROC Curve je kritérium, které také umožňuje nalézt nejlepší klasifikaci všech objektů a které je dosaženo při maximální hodnotě tohoto kritéria. Minimální hodnoty je potom dosaženo, když řádky (čili objekty) jsou zcela nesprávně klasifikovány.

Blok **Row Classification Section:**

Row Classification Section

Row	Actual REMISS	Estimated REMISS	Estimated REMISS Probability	Lower 95% Confidence Limit	Upper 95% Confidence Limit
1	1	1	0,72265	0,16892	0,97093
2	1	1	0,57874	0,26788	0,83762
3	0	0	0,89540	0,36581	0,99219
4	0	0	0,71742	0,34317	0,92502
5	1	1	0,71418	0,25218	0,94876
6	0	0	0,72911	0,31049	0,94148
7*	1	0	0,67844	0,40484	0,86745
8*	0	1	0,60723	0,10572	0,95287
9	0	0	0,83368	0,43877	0,96982
10	0	0	0,99843	0,31038	1,00000
11	0	0	0,92715	0,50018	0,99386
12	0	0	0,82714	0,12794	0,99363
13	0	0	0,99654	0,53470	0,99999
14	0	0	0,99982	0,03518	1,00000
15*	0	1	0,57122	0,25303	0,83973
16	1	1	0,71470	0,15362	0,97189
17	0	0	0,99938	0,37335	1,00000
18	0	0	0,77711	0,36330	0,95517
19	0	0	0,99846	0,20356	1,00000
20	1	1	0,64911	0,26305	0,90555
21	0	0	0,98307	0,49525	0,99971
22	0	0	0,99378	0,43938	0,99997
23*	1	0	0,74739	0,36403	0,93863
24*	0	1	0,87011	0,40910	0,98481
25	1	1	0,93132	0,44114	0,99573
26*	1	0	0,53949	0,21471	0,83388
27	0	0	0,71742	0,34317	0,92502
28		0	0,71742	0,34317	0,92502
29		0	0,71742	0,34317	0,92502

Blok ukazuje na aktuální a predikovanou hodnotu skupiny závisle proměnné Y a její pravděpodobnosti pro každý řádek (=objekt). Přináší také dolní a horní mez intervalu pravděpodobnosti.

Row značí číslo řádku (=objektu). Hvězdička značí chybné zařazení.

Actual Group značí skutečnou skupinu závisle proměnné Y, do které objekt patří.

Estimated Group značí odhadnutou skupinu závisle proměnné Y s největší zařazovací hodnotou pravděpodobnosti.

Estimated Probability značí odhadnutou pravděpodobnost, se kterou byl objekt zařazen do skupiny závisle proměnné Y.
Hodnoty od 0.7 do 1.0 vyjadřují jistotu správného zařazení, zatímco hodnoty okolo 0.5 vyjadřují nejisté až nespolehlivé zařazení.

Lower and Upper Confidence Limits značí intervalový odhad odhadované hodnoty zařazovací pravděpodobnosti.

Blok **Row Classification Probabilities:**

Row Classification Probabilities

Row	Actual REMISS	Estimated Prob. in 0	Estimated Prob. in 1
1	1	0,27735	0,72265
2	1	0,42126	0,57874
3	0	0,89540	0,10460
4	0	0,71742	0,28258
5	1	0,28582	0,71418
6	0	0,72911	0,27089
7*	1	0,67844	0,32156
8*	0	0,39277	0,60723
9	0	0,83368	0,16632
10	0	0,99843	0,00157
11	0	0,92715	0,07285
12	0	0,82714	0,17286
13	0	0,99654	0,00346
14	0	0,99982	0,00018
15*	0	0,42878	0,57122
16	1	0,28530	0,71470
17	0	0,99938	0,00062
18	0	0,77711	0,22289
19	0	0,99846	0,00154
20	1	0,35089	0,64911
21	0	0,98307	0,01693
22	0	0,99378	0,00622
23*	1	0,74739	0,25261
24*	0	0,12989	0,87011
25	1	0,06868	0,93132
26*	1	0,53949	0,46051
27	0	0,71742	0,28258
28		0,71742	0,28258
29		0,71742	0,28258

Blok přináší skutečnou a odhadnutou hodnotu obou skupinových pravděpodobností. To umožňuje posoudit jak spolehlivě byla klasifikace provedena.

Row značí číslo řádku (=objektu). Hvězdička značí chybné zařazení.

Actual Group značí skutečnou skupinu závisle proměnné Y, do které objekt patří.

Estimated Prob. In Group značí odhadnutou pravděpodobnost, se kterou byl objekt zařazen do skupiny závisle proměnné Y. Hodnoty umožňují zařadit objekt do skupiny 0 nebo 1.

Hodnoty od 0.7 do 1.0 vyjadřují jistotu správného zařazení, zatímco hodnoty okolo 0.5 vyjadřují nejisté až nespolehlivé zařazení.

Blok **Simple Residual Report:**

Simple Residual Report

Row	Actual REMISS	Residual for Group 0	Residual for Group 1
1	1	-0,27735	0,27735
2	1	-0,42126	0,42126
3	0	0,10460	-0,10460
4	0	0,28258	-0,28258
5	1	-0,28582	0,28582
6	0	0,27089	-0,27089
7*	1	-0,67844	0,67844
8*	0	0,60723	-0,60723
9	0	0,16632	-0,16632
10	0	0,00157	-0,00157
11	0	0,07285	-0,07285
12	0	0,17286	-0,17286
13	0	0,00346	-0,00346
14	0	0,00018	-0,00018
15*	0	0,57122	-0,57122
16	1	-0,28530	0,28530
17	0	0,00062	-0,00062
18	0	0,22289	-0,22289
19	0	0,00154	-0,00154
20	1	-0,35089	0,35089
21	0	0,01693	-0,01693
22	0	0,00622	-0,00622
23*	1	-0,74739	0,74739
24*	0	0,87011	-0,87011
25	1	-0,06868	0,06868
26*	1	-0,53949	0,53949
27	0	0,28258	-0,28258
28		-0,71742	-0,28258
29		-0,71742	-0,28258

Blok přináší klasická rezidua pro každou skupinu. Každá z G logistických regresních rovnic je zde použita k vyčíslení pravděpodobností, že objekt patří do odpovídající skupiny.

Row značí číslo řádku (=objektu). Hvězdička značí chybné zařazení.

Actual Group značí skutečnou skupinu závisle proměnné Y, do které objekt patří.

Residual for Group ukazuje, že rezidua jsou definována $r = y - p$, kde p je odhadnutá hodnota pravděpodobnosti, y je hodnota závisle proměnné objektu aktuální skupiny G nebo 0.

Když bude logistická regrese prokládat data perfektně, budou rezidua maličká, nikdy však nuly. Protože rezidua vykazují nekonstantní rozptyl, je třeba použít ještě ostatní druhy reziduí.

Blok **Residual Report:**

Residual Report

Row	Actual REMISS	Pearson Residual	Deviance Residual	Maximum Hat Diagonal
1	1	-0,61951	-0,80602	0,33952
2	1	-0,85317	-1,04585	0,11109
3	0	0,34179	0,47007	0,17739
4	0	0,88756	1,15255	0,26378
5	1	-0,63262	-0,82051	0,21315
6	0	0,60953	0,79489	0,16436
7*	1	-1,45254	-1,50638	0,07277
8*	0	1,24340	1,36714	0,41037
9	0	0,44665	0,60316	0,12462
10	0	0,03965	0,05604	0,02146
11	0	0,28032	0,38895	0,11371
12	0	0,45715	0,61608	0,45200
13	0	0,05890	0,08323	0,02738
14	0	0,01360	0,01924	0,00683
15*	0	1,15421	1,30139	0,11956
16	1	-0,63182	-0,81963	0,36570
17	0	0,02495	0,03528	0,01010
18	0	0,53555	0,71017	0,14772
19	0	0,03931	0,05556	0,02462
20	1	-0,73524	-0,92968	0,16050
21	0	0,13123	0,18480	0,07214
22	0	0,07910	0,11169	0,04549
23*	1	-1,72010	-1,65887	0,13262
24*	0	2,58819	2,02042	0,15155
25	1	-0,27157	-0,37724	0,13465
26*	1	-1,08236	-1,24533	0,13693
27	0	0,88756	1,15255	0,26378
28				0,00000
29				0,00000

Blok přináší také grafické čarové zobrazení Pearsonova rezidua, rezidua odchylky, diagonální prvky H-projekční matice pro každý řádek objektu.

Row značí číslo řádku (=objektu). Hvězdička značí chybné zařazení.

Actual Group značí skutečnou skupinu závisle proměnné Y, do které objekt patří.

Pearson residual značí příspěvek každého řádku (objektu) do celkové hodnoty Pearsonova Chi-kvadrát testu dobré shody. Zde existuje vzorec pro výpočet Chi-kvadrát, když hodnoty nezávisle proměnných jsou jednotkové.

Deviance residual značí existuje vzorec pro výpočet rezidua deviance. Suma čtverců reziduí deviance je potom výsledná deviance -2LL.

Maximum Hat Diagonal značí diagonální prvky H-projekční matice, které slouží k odhalení extrémů (leverages). Čím větší hodnota extrému, tím větší vliv na odhad regresního koeficientu.

Blok **DFBetas Report:**

DFBetas Report For REMISS = 0

Row	Actual REMISS	DFBeta Intercept		DFBeta CELL		DFBeta LI	
1	1	-0,11266	|.....	0,23931	||||||.....	-0,25676	|||.....
2	1	-0,10512	|.....	0,06037	|.....	-0,13381	|.....
3	0	-0,01565	|.....	-0,14643	|||.....	-0,09555	|.....
4	0	0,05433	|.....	0,14863	|||.....	-0,24383	|||.....
5	1	-0,30624	|||||.....	-0,04696	|.....	-0,22267	|||.....
6	0	0,06554	|.....	0,07113	|.....	-0,10567	|.....
7*	1	0,06889	|.....	0,07190	|.....	0,17097	||.....
8*	0	-0,81051	|||||||||||||.	-0,10278	||.....	0,18572	||.....
9	0	-0,11622	||.....	-0,01746	|.....	-0,13122	|.....
10	0	-0,00558	|.....	-0,00300	|.....	-0,00493	|.....
11	0	-0,03830	|.....	-0,08440	||.....	-0,08123	|.....
12	0	0,01728	|.....	-0,49468	||||||||||||..	-0,13585	|.....
13	0	-0,00728	|.....	-0,00783	|.....	-0,00881	|.....
14	0	-0,00042	|.....	-0,00112	|.....	-0,00064	|.....
15*	0	0,18159	|||.....	0,28327	||||||.....	0,14545	||.....
16	1	0,29189	|||||.....	-0,12943	|||.....	-0,16895	||.....
17	0	-0,00178	|.....	-0,00220	|.....	-0,00203	|.....
18	0	-0,18067	|||.....	-0,01490	|.....	-0,14301	||.....
19	0	-0,00264	|.....	-0,00615	|.....	-0,00414	|.....
20	1	-0,21936	|||.....	-0,23074	||||||.....	-0,16455	||.....
21	0	-0,03173	|.....	-0,01401	|.....	-0,03408	|.....
22	0	-0,01564	|.....	-0,01076	|.....	-0,01516	|.....
23*	1	0,48975	|||||||.....	-0,00063	|.....	0,39819	|||||.....
24*	0	0,85306	||||||||||||||	0,56896	|||||||||||||	1,02839	|||||||||||||
25	1	-0,07627	|.....	-0,07733	||.....	-0,10248	|.....
26*	1	-0,18615	|||.....	-0,23305	||||||.....	-0,02377	|.....
27	0	0,05433	|.....	0,14863	|||.....	-0,24383	|||.....
28			|.....		|.....		|.....
29			|.....		|.....		|.....

Jedním ze způsobů jak studovat vliv řádkového objektu na odhad regresních koeficientů je sledování změny koeficientu, když je řádkový objekt odstraněn.

DFBETA je standardizovaný rozdíl mezi regresním koeficientem před a po odstranění *j*tého objektu.

Row značí číslo řádku (=objektu). Hvězdička značí chybné zařazení.

Actual Group značí skutečnou skupinu závisle proměnné Y, do které objekt patří.

DFBeta značí standardizovaný rozdíl mezi regresním koeficientem před a po odstranění *j*tého objektu.

Blok **DFBetas Report:** pokračování

DFBetas Report For REMISS = 0

Row	Actual REMISS	DFBeta TEMP	
1	1	0,08266	|.....
2	1	0,09271	|.....
3	0	0,03589	|.....
4	0	-0,05653	|.....
5	1	0,29571	||||.....
6	0	-0,06431	|.....
7*	1	-0,08162	|.....
8*	0	0,76270	|||||||||||...
9	0	0,11523	|.....
10	0	0,00571	|.....
11	0	0,04878	|.....
12	0	0,04931	|.....
13	0	0,00800	|.....
14	0	0,00055	|.....
15*	0	-0,20450	|||.....
16	1	-0,25160	||||.....
17	0	0,00199	|.....
18	0	0,17545	||.....
19	0	0,00333	|.....
20	1	0,23492	|||.....
21	0	0,03236	|.....
22	0	0,01634	|.....
23*	1	-0,47172	|||||||.....
24*	0	-0,89006	||||||||||||||
25	1	0,08316	|.....
26*	1	0,19868	|||.....
27	0	-0,05653	|.....
28			|.....
29			|.....

Blok **Influence Diagnostics Report:**

Influence Diagnostics Report For REMISS = 0

Row	Actual REMISS	Hat Diagonal	Cook's Distance (C)	Cook's Distance (CBar)
1	1	0,33952	0,29870	0,19729
2	1	0,11109	0,10233	0,09097
3	0	0,17739	0,03062	0,02519
4	0	0,26378	0,38338	0,28225
5	1	0,21315	0,13778	0,10841
6	0	0,16436	0,08745	0,07308
7*	1	0,07277	0,17858	0,16558
8*	0	0,41037	1,82494	1,07603
9	0	0,12462	0,03244	0,02840
10	0	0,02146	0,00004	0,00003
11	0	0,11371	0,01137	0,01008
12	0	0,45200	0,31454	0,17237
13	0	0,02738	0,00010	0,00010
14	0	0,00683	0,00000	0,00000
15*	0	0,11956	0,20547	0,18091
16	1	0,36570	0,36285	0,23016
17	0	0,01010	0,00001	0,00001
18	0	0,14772	0,05833	0,04971
19	0	0,02462	0,00004	0,00004
20	1	0,16050	0,12311	0,10335
21	0	0,07214	0,00144	0,00134
22	0	0,04549	0,00031	0,00030
23*	1	0,13262	0,52154	0,45238
24*	0	0,15155	1,41023	1,19651
25	1	0,13465	0,01326	0,01148
26*	1	0,13693	0,21535	0,18586
27	0	0,26378	0,38338	0,28225
28				
29				

Blok přináší dvě vzdálenostní míry podobné Cookově vzdálenosti ve vícenásobné lineární regresi.

Row značí číslo řádku (=objektu). Hvězdička značí chybné zařazení.

Actual Group značí skutečnou skupinu závisle proměnné Y, do které objekt patří.

Hat Diagonal značí diagonální prvky H-projekční matice, které slouží k odhalení extrémů (leverages). Čím větší, tím větší vliv na odhad regr. koeficientu.

Cook´s Distance (C) and (Cbar) značí rozšířené Cookovy vzdálenosti. **Cbar** měří celkovou změnu ve fitovaném logitu při vyloučení /tého objektu pro jinak všechny body kromě právě jediného, vyloučeného.

Blok **Residual Diagnostics Report:**

Residual Diagnostics Report For REMISS = 0

Row	Actual REMISS	Hat Diagonal	Deviance Change (DFDev)	Chi-Square Change (DFChi2)
1	1	0,33952	0,84695	0,58109
2	1	0,11109	1,18477	0,81886
3	0	0,17739	0,24616	0,14201
4	0	0,26378	1,61061	1,07001
5	1	0,21315	0,78165	0,50862
6	0	0,16436	0,70493	0,44461
7*	1	0,07277	2,43475	2,27547
8*	0	0,41037	2,94510	2,62206
9	0	0,12462	0,39220	0,22790
10	0	0,02146	0,00318	0,00161
11	0	0,11371	0,16137	0,08866
12	0	0,45200	0,55193	0,38135
13	0	0,02738	0,00702	0,00357
14	0	0,00683	0,00037	0,00019
15*	0	0,11956	1,87453	1,51311
16	1	0,36570	0,90195	0,62935
17	0	0,01010	0,00125	0,00063
18	0	0,14772	0,55405	0,33653
19	0	0,02462	0,00313	0,00158
20	1	0,16050	0,96766	0,64393
21	0	0,07214	0,03549	0,01856
22	0	0,04549	0,01277	0,00655
23*	1	0,13262	3,20423	3,41111
24*	0	0,15155	5,27863	7,89526
25	1	0,13465	0,15379	0,08522
26*	1	0,13693	1,73671	1,35737
27	0	0,26378	1,61061	1,07001
28				
29				

Blok odhaluje objekty, které nebyly dobře proloženy regresním modelem.

Row značí číslo řádku (=objektu). Ohvězdičkované číslo značí chybné zařazení.

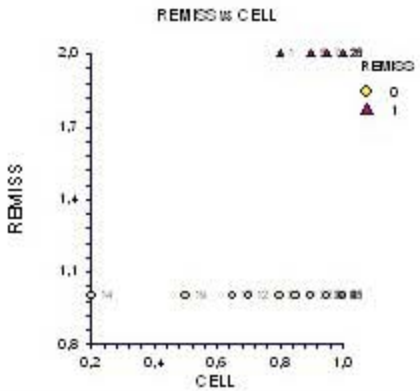
Actual Group značí skutečnou skupinu závisle proměnné Y, do které objekt patří.

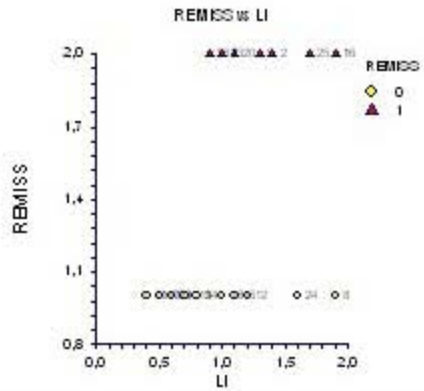
Hat Diagonal značí diagonální prvky H-projekční matice, které slouží k odhalení extrémů (leverages). Čím větší, tím větší vliv na odhad.

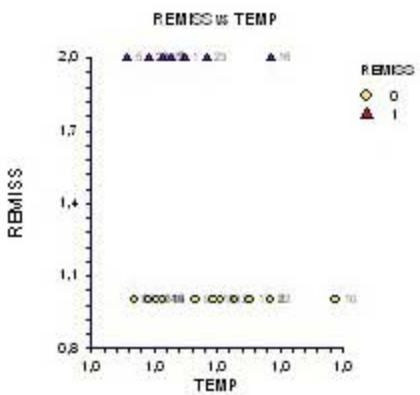
Deviance Change (DFDev) and Chi-Square Change (DFChi2) značí statistiky, které měří změnu v devianci a v Pearsonově Chi-kvadrát hodnotě, jež se objeví, když je jeden objekt vyloučen ze zdrojové matice. Veliké hodnoty obou těchto statistik značí, že objekt nebyl dobře proložen regresním modelem. Tabulka je vylepšena o čarový diagram, který umožní nalézt velké hodnoty špatných objektů.

Graf Y versus X Plots:

Diagnostic Plots





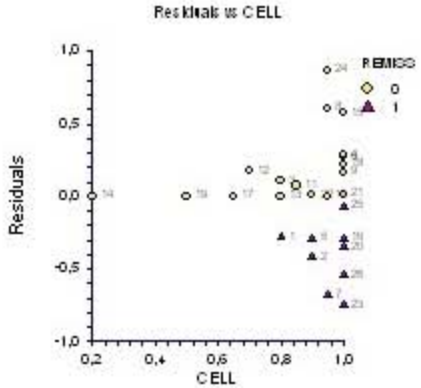


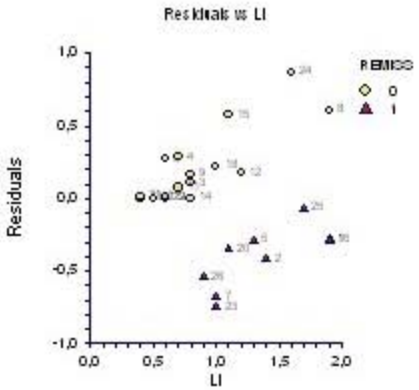
Grafy ukazují závisle proměnnou REMISS na vertikální ose a jednotlivé nezávisle proměnné na horizontální ose. Graf je k odhalení chyb, odleh-lých hodnot a veškerých anomálií.

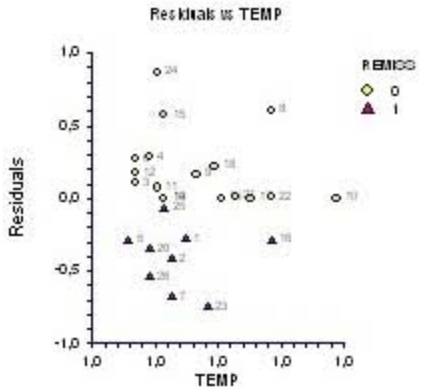
Vertical Axis přináší kategorie závisle proměnné REMISS. Každá kategorie je zobrazena celým číslem, začína-jícím v čísle 1. Body jsou dle pořado-vého čísla. Závisle proměnné o hodnotách A, B, C budou seřazeny v numerickém intervalu od 0.8 do 3.2 a skupiny budou zobrazeny pomocí integerů 1, 2, 3.

Horizontal Axis přináší nezávisle proměnné. Je-li nezávisle proměnná kategorická, binární je obrázek vypočten pro každou kategorii a samostatný rozptylový graf je vypočten pro každou binární proměnnou.

Graf Simple Residuals versus X Plot:





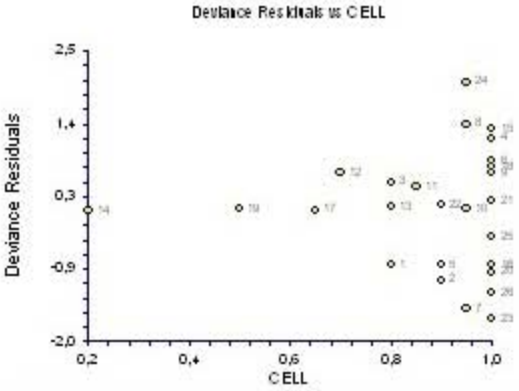


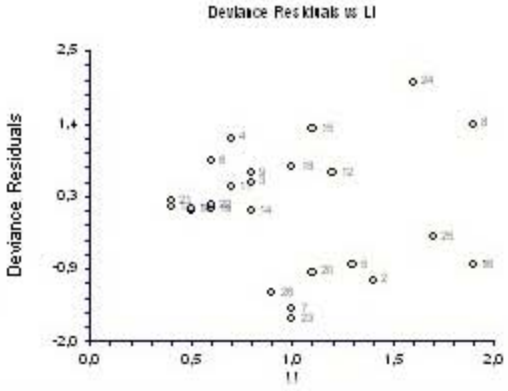
Tyto rozptylové grafy ukazují klasická rezidua na vertikální ose a jednotlivé nezávisle proměnné na horizontální ose. Graf je vhodný k odhalení chyb, odlehlých hodnot a veškerých anomálií.

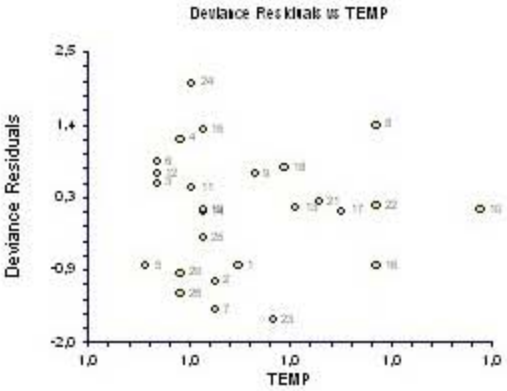
Vertical Axis přináší rezidua. Je zobrazeno G druhů reziduí a pro N bodů to znamená GN zobrazených bodů.

Horizontal Axis přináší nezávisle proměnné. Je-li nezávisle proměnná kategorická, binární je obrázek vypočten pro každou kategorii a samostatný rozptylový graf je vypočten pro každou binární proměnnou.

Graf Deviance Residuals versus X Plot:





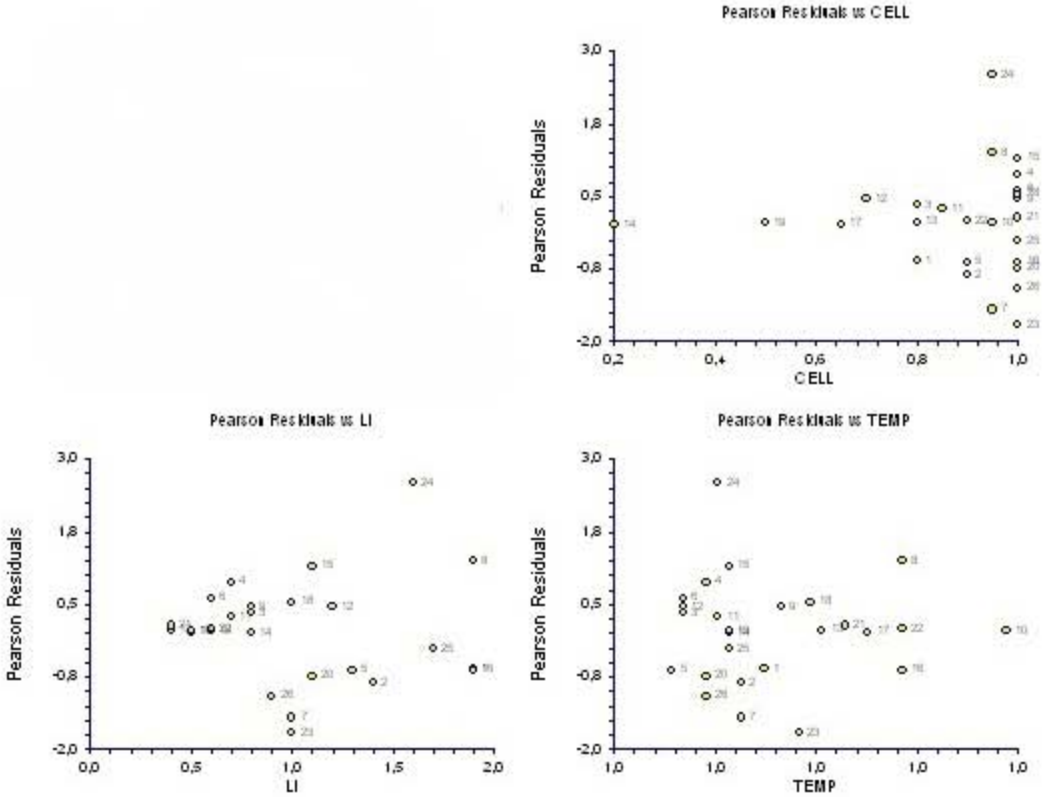


Tyto rozptylové grafy ukazují rezidua deviance na vertikální ose a jednotlivé nezávisle proměnné na horizontální ose. Graf je vhodný k odhalení chyb, odlehlých hodnot a veškerých anomálií.

Vertical Axis přináší rezidua deviance.

Horizontal Axis přináší nezávisle proměnné. Je-li nezávisle proměnná kategorická, binární je obrázek vypočten pro každou kategorii a samostatný rozptylový graf je vypočten pro každou binární proměnnou.

Graf **Pearson Residuals versus X Plot:**

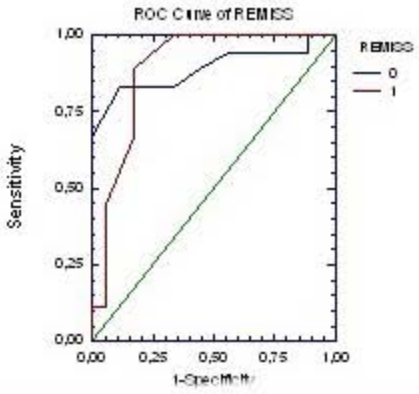


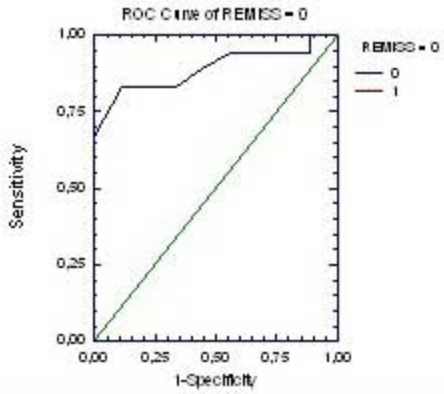
Tyto rozptylové grafy ukazují Pearsonova rezidua na vertikální ose a jednotlivé nezávisle proměnné na horizontální ose. Graf je vhodný k odhalení chyb, odlehlých hodnot a veškerých anomálií.

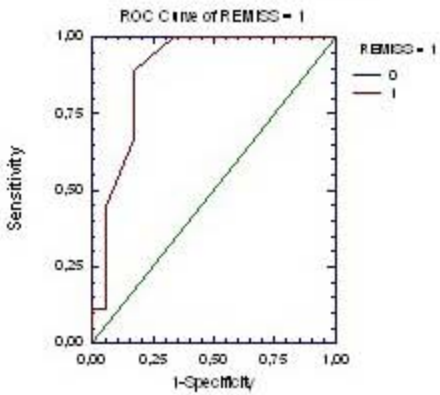
Vertical Axis přináší Pearsonova rezidua.

Horizontal Axis přináší nezávisle proměnné. Je-li nezávisle proměnná kategorická, binární je obrázek vypočten pro každou kategorii a samostatný rozptylový graf je vypočten pro každou binární proměnnou.

Graf ROC Curves – Combined and Separate Plot:





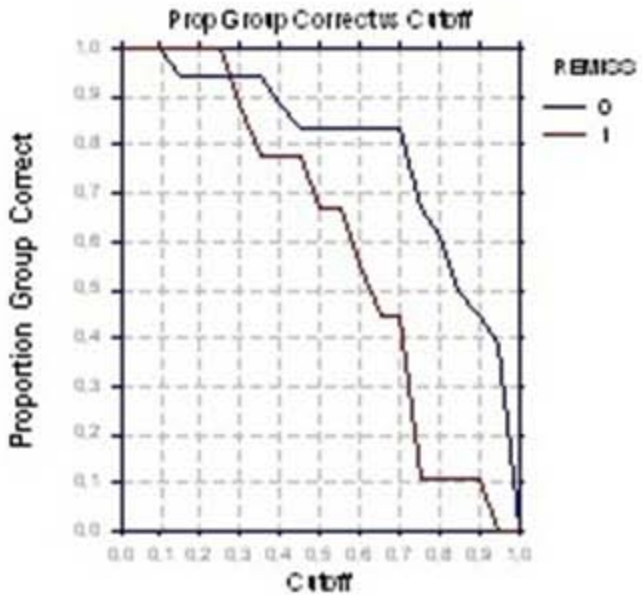


Tyto ROC grafy pomáhají nalézt nejlepší hodnotu prahového bodu ke klasifikaci objektů.

Vertical Axis přináší sensitivitu.

Horizontal Axis přináší 1 minus specificita.

Graf Prob Correct versus Cutoff Plot:



Graf přináší podíl správně zařazených objektů proti prahovému bodu pravděpodobnosti. Je proto užitečný k určení prahového bodu klasifikace. Tento graf je obtížné užít nebo vůbec nepoužitelný pro více jak 3 či více kategorií.

Vertical Axis přináší procentuální podíl správně zařazených objektů.

Horizontal Axis přináší prahový bod pravděpodobnosti jsou ve členech odhadovaných skupin. Proto prahový bod 0.4 znamená, že všechny řádky objektů se skupinovou pravděpodobností 0.4 nebo větší jsou zařazeny do této skupiny.

Konec úlohy

Example 2 – Subset Selection

(Hledání nejlepší sestavy numerických spojitých nezávisle proměnných)

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 navrženém logistickém regresním modelu.

Data: $n = 27, p = 6,$

Index	*REMISS*	*CELL*	*SMEAR*	*INFIL*	*LI*	*BLAST*	*TEMP*
1	1	0.8	0.83	0.66	1.9	1.1	0.996
2	1	0.9	0.36	0.32	1.4	0.74	0.992
3	0	0.8	0.88	0.7	0.8	0.176	0.982
4	0	1	0.87	0.87	0.7	1.053	0.986
5	1	0.9	0.75	0.68	1.3	0.519	0.98
6	0	1	0.65	0.65	0.6	0.519	0.982
7	1	0.95	0.97	0.92	1	1.23	0.992
8	0	0.95	0.87	0.83	1.9	1.354	1.02
9	0	1	0.45	0.45	0.8	0.322	0.999
10	0	0.95	0.36	0.34	0.5	0	1.038
11	0	0.85	0.39	0.33	0.7	0.279	0.988
12	0	0.7	0.76	0.53	1.2	0.146	0.982
13	0	0.8	0.46	0.37	0.4	0.38	1.006

Index	*REMISS*	*CELL*	*SMEAR*	*INFIL*	*LI*	*BLAST*	*TEMP*
14	0	0.2	0.39	0.08	0.8	0.114	0.99
15	0	1	0.9	0.9	1.1	1.037	0.99
16	1	1	0.84	0.84	1.9	2.064	1.02
17	0	0.65	0.42	0.27	0.5	0.114	1.014
18	0	1	0.75	0.75	1	1.322	1.004
19	0	0.5	0.44	0.22	0.6	0.114	0.99
20	1	1	0.63	0.63	1.1	1.072	0.986
21	0	1	0.33	0.33	0.4	0.176	1.01
22	0	0.9	0.93	0.84	0.6	1.591	1.02
23	1	1	0.58	0.58	1	0.531	1.002
24	0	0.95	0.32	0.3	1.6	0.886	0.988
25	1	1	0.6	0.6	1.7	0.964	0.99
26	1	1	0.69	0.69	0.9	0.398	0.986
27	0	1	0.73	0.73	0.7	0.398	0.986

Postup:

1. **Otevření souboru dat LEUKEMIA:** Ve **File** menu klikněte na **Open, Data, LEUKEMIA.S0, Open.** Obdržíte zdrojovou matici dat.

NCSS 2007 Data - [C:\Program Files\NCSS2007\Data\Leukemia.s0]

File Edit Data Analysis Graphics PASS GESS Tools Window Help

	REMISS	CELL	SMEAR	INFIL	LI	BLAST	TEMP	RE
1	1	0,8	0,83	0,66	1,9	1,1	0,996	
2	1	0,9	0,36	0,32	1,4	0,74	0,992	
3	0	0,8	0,88	0,7	0,8	0,176	0,982	
4	0	1	0,87	0,87	0,7	1,053	0,986	
5	1	0,9	0,75	0,68	1,3	0,519	0,98	
6	0	1	0,65	0,65	0,6	0,519	0,982	
7	1	0,95	0,97	0,92	1	1,23	0,992	
8	0	0,95	0,87	0,83	1,9	1,354	1,02	
9	0	1	0,45	0,45	0,8	0,322	0,999	
10	0	0,95	0,36	0,34	0,5	0	1,038	
11	0	0,85	0,39	0,33	0,7	0,279	0,988	
12	0	0,7	0,76	0,53	1,2	0,146	0,982	
13	0	0,8	0,46	0,37	0,4	0,38	1,006	
14	0	0,2	0,39	0,08	0,8	0,114	0,99	
15	0	1	0,9	0,9	1,1	1,037	0,99	
16	1	1	0,84	0,84	1,9	2,064	1,02	
17	0	0,65	0,42	0,27	0,5	0,114	1,014	
18	0	1	0,75	0,75	1	1,322	1,004	
19	0	0,5	0,44	0,22	0,6	0,114	0,99	
20	1	1	0,63	0,63	1,1	1,072	0,986	
21	0	1	0,33	0,33	0,4	0,176	1,01	
22	0	0,9	0,93	0,84	0,6	1,591	1,02	
23	1	1	0,58	0,58	1	0,531	1,002	
24	0	0,95	0,32	0,3	1,6	0,886	0,988	
25	1	1	0,6	0,6	1,7	0,964	0,99	
26	1	1	0,69	0,69	0,9	0,398	0,986	
27	0	1	0,73	0,73	0,7	0,398	0,986	
28		1			0,7		0,986	
29		1			0,7		0,986	
30								

2. Volba metody: V menu vyberte **Analysis, Regression/Correlation, Logistic Regression.**

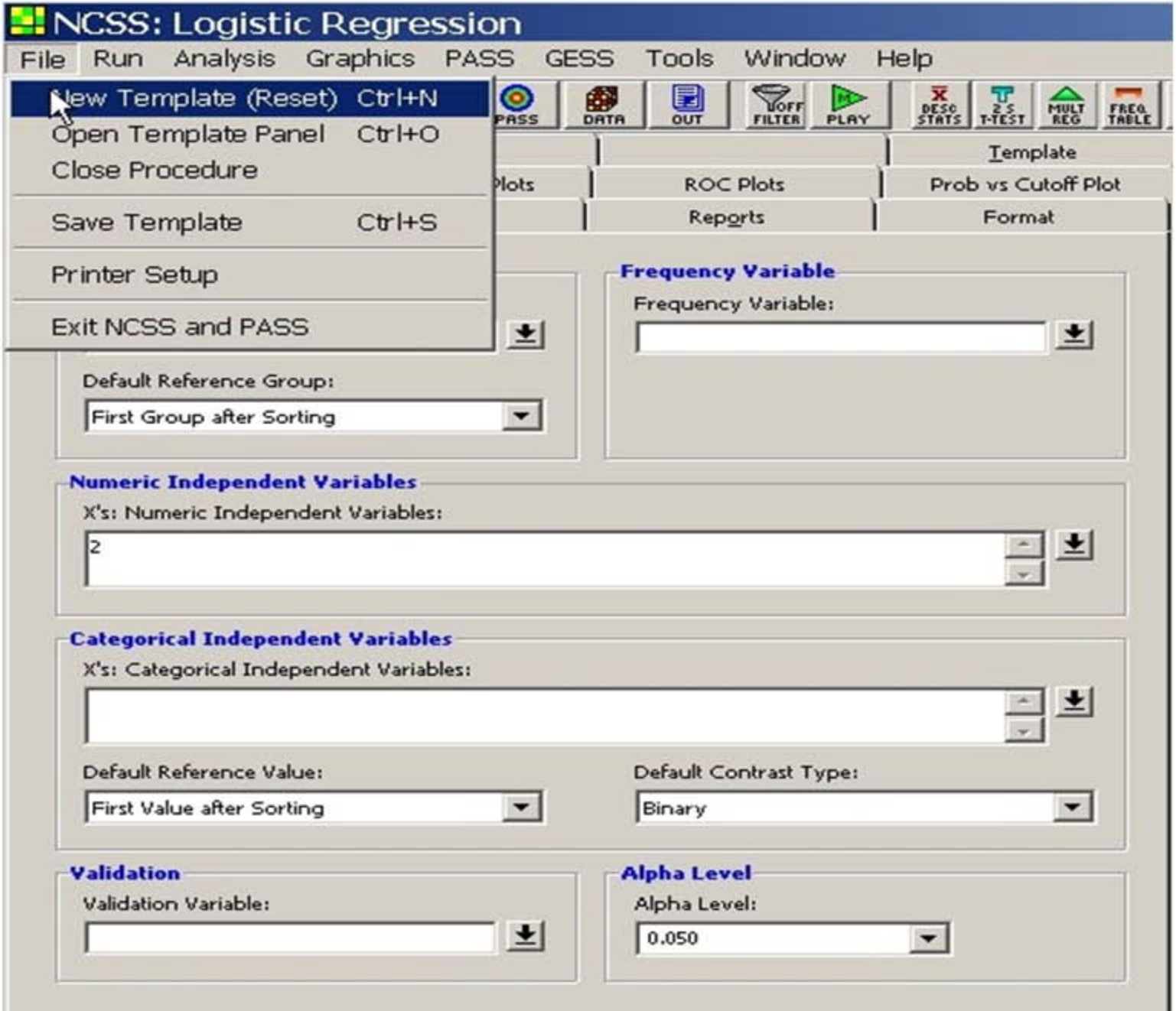
NCSS 2007 Data - [C:\Program Files\NCSS2007\Data\Leukemia.s0]

File Edit Data Analysis Graphics PASS GESS Tools Window Help

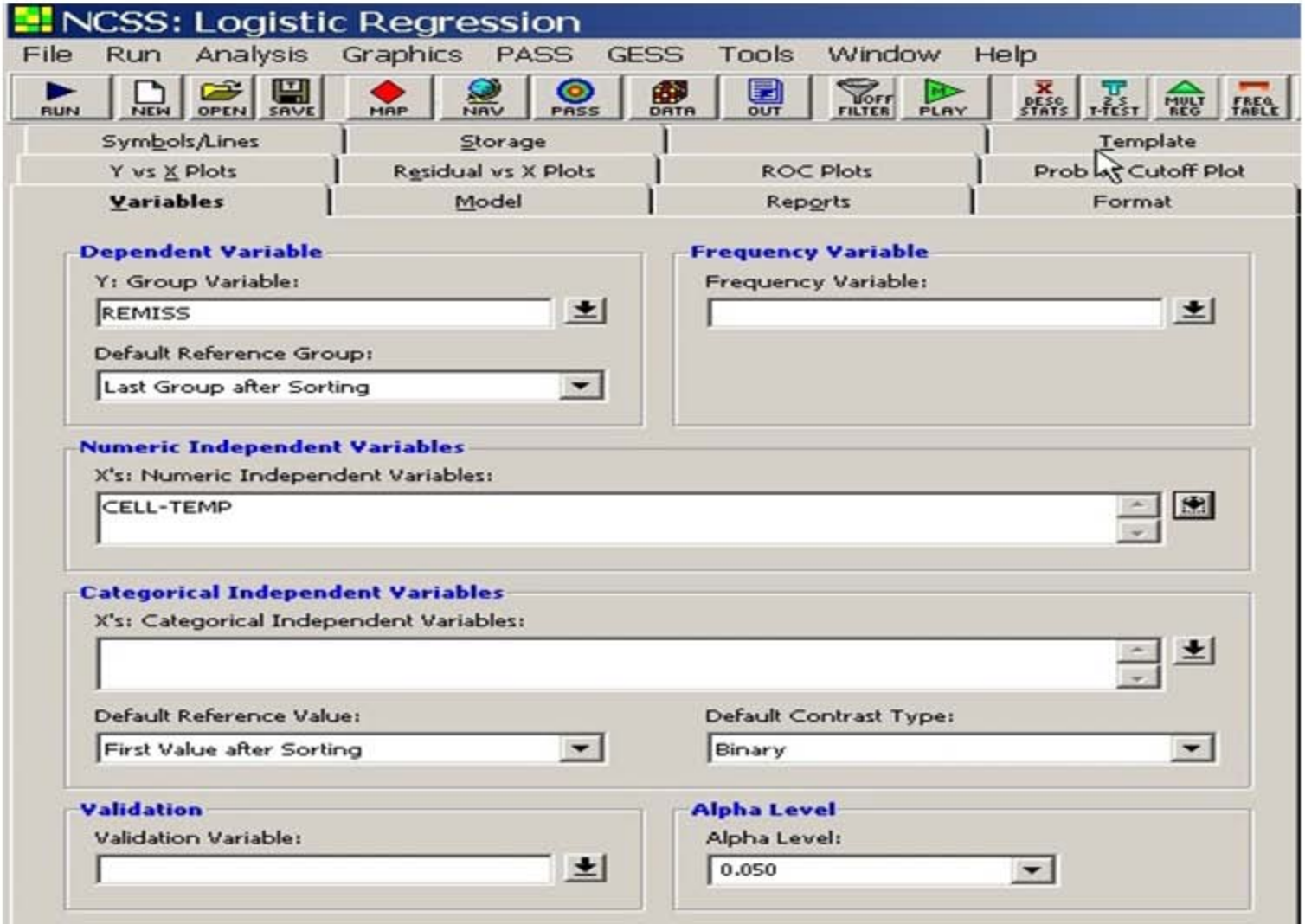
- Analysis of Variance (ANOVA)
- Clustering
- Curve Fitting
- Descriptive Statistics
- Design of Experiments
- Diagnostic Tests
- Forecasting / Time Series
- Mass Appraisal
- Meta-Analysis
- Mixed Models
- Multivariate Analysis
- Nondetects
- Operations Research
- Proportions
- Quality Control
- Regression / Correlation
  - Linear Regression and Correlation
  - Multiple Regression
  - Cox Regression
  - Logistic Regression
  - Correlation Routines
  - Nonlinear Routines
  - Other Regression Routines
  - Variable Selection Routines
  - Legacy Routines
- Survival / Reliability
- T-Tests
- Other

	REMISS				LI	BLAST	TEMP	RE
1	1			0,66	1,9	1,1	0,996	
2	1			0,32	1,4	0,74	0,992	
3	0			0,7	0,8	0,176	0,982	
4	0			0,87	0,7	1,053	0,986	
5	1			0,68	1,3	0,519	0,98	
6	0			0,65	0,6	0,519	0,982	
7	1			0,92	1	1,23	0,992	
8	0			0,83	1,9	1,354	1,02	
9	0			0,45	0,8	0,322	0,999	
10	0			0,34	0,5	0	1,038	
11	0			0,33	0,7	0,279	0,988	
12	0			0,53	1,2	0,146	0,982	
13	0			0,37	0,4	0,38	1,006	
14	0			0,08	0,8	0,114	0,99	
15	0					,037	0,99	
16	1					,064	1,02	
17	0					,114	1,014	
18	0					,322	1,004	
19	0					,114	0,99	
20	1	1	0,63			,072	0,986	
21	0	1	0,33			,176	1,01	
22	0	0,9	0,93			,591	1,02	
23	1	1	0,58			,531	1,002	
24	0	0,95	0,32			,886	0,988	
25	1	1	0,6			,964	0,99	
26	1	1	0,69			,398	0,986	
27	0	1	0,73	0,73	0,7	0,398	0,986	
28		1			0,7		0,986	
29		1			0,7		0,986	
30								

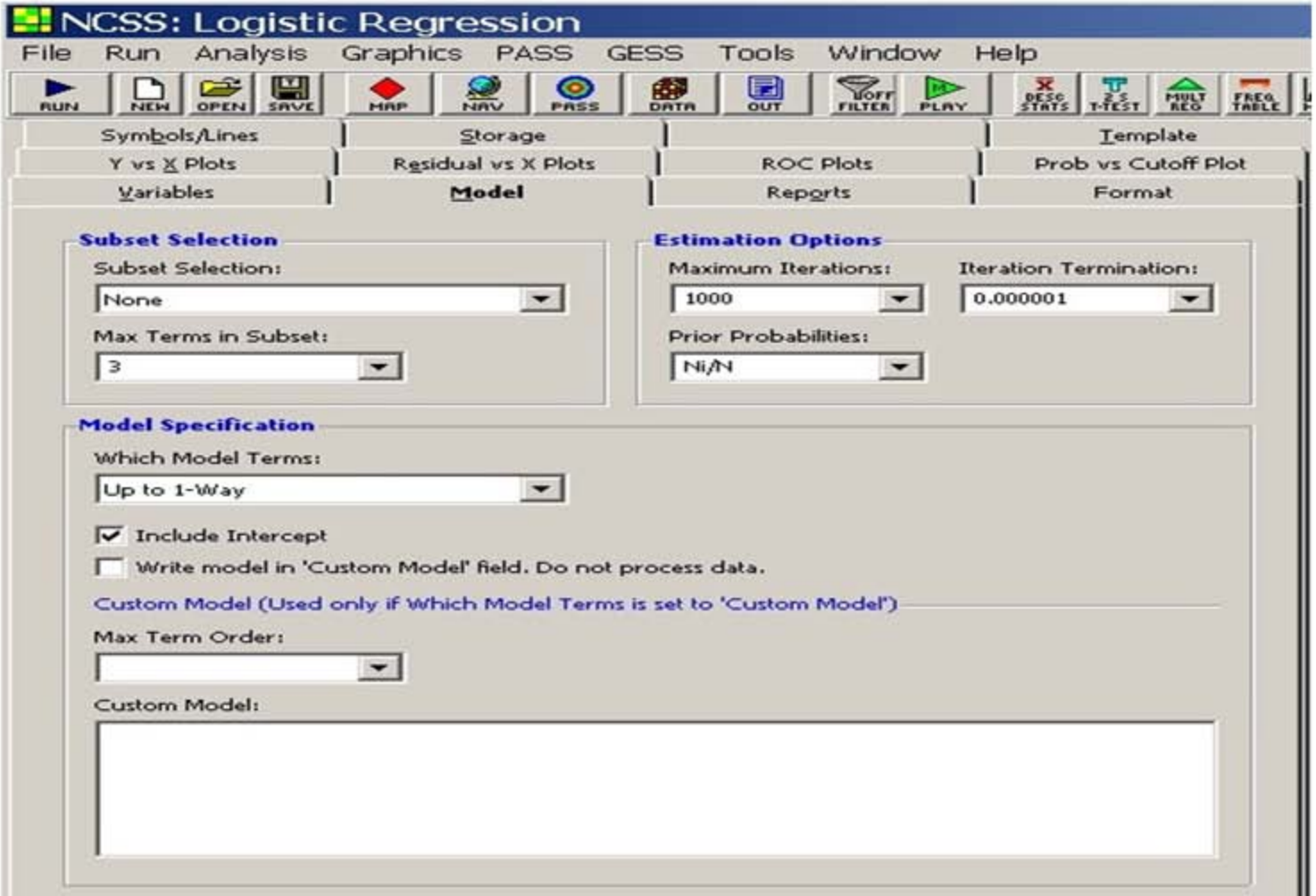
3. Zadání nové šablony inputu dat: Kliknete na **File** a zvolíte **New Template.**



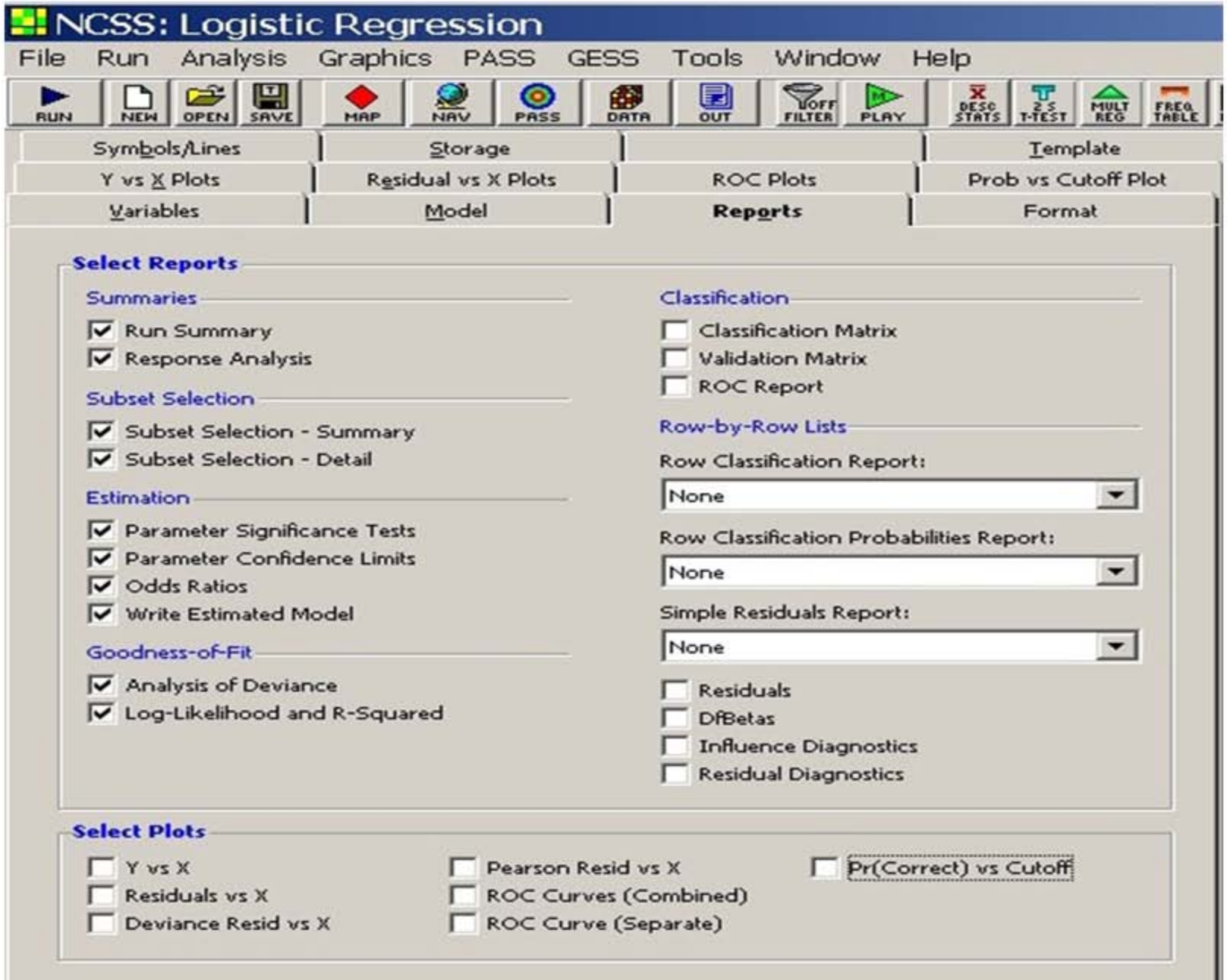
4. Zadání závisle proměnné *y*: Vyberte záložku **Variables**, kde v okénku **Y: Group Variable** zadáte proměnnou **REMISS** a v **Default Reference Group** zadáte **Last Group after Sorting.** V okénku numerické nezávisle proměnné **X´s:Numeric Independent Variables** zadáte jenom vybrané tři nezávisle proměnné znaky **CELL, LI, TEMP**. Při zadávání těchto spojitých proměnných podržíte klávesu Ctrl. Ostatní okénka jsou na předvolených hodnotách nebo prázdná.



5. Vložení modelu: Zadáte logistický regresní model kliknutím na záložku **Model**. V okénku **Subset Selection** dáte volbu **None** a v okénku **Max Terms in Subset** pak číslo 6 pro všechny nezávisle proměnné. Dále v okénku **Which Model Terms** zvolíte **Up to 1-Way.** Zaškrtnete **Include Intercept,** protože požadujete v regresním modelu také absolutní člen. V **Maximum Iteration** dáte doporučenou hodnotu **1000**, v **Iteration Termination** kritérium **1E-7** a konečně v **Prior Probability** zvolíte **Ni/N**. Ostatní okénka necháte volná.



6. Zadání protokolu: Vyberte záložku **Reports**. Zaškrtnete pouze částečný výstup dle vzoru.



7. Proběhne výpočet: Kliknutím na **Run Procedure** se spustí výpočet.

8. Komentář k výstupu:

Logistic Regression Report

Page/Date/Time	1 22.7.2009 7:58:50
Database	C:\PROGRAM FILES\NCSS2007\DATA\LEUKEMIA.S0
Response	REMISS

Run Summary Section

Parameter	Value	Parameter	Value
Dependent Variable	REMISS	Rows Processed	29
Reference Group	1	Rows Used	27
Number of Groups	2	Rows for Validation	0
Frequency Variable	None	Rows X's Missing	2
Numeric Ind. Variables	6	Rows Freq Miss. or 0	0
Categorical Ind. Variables	0	Rows Prediction Only	0
Final Log Likelihood	-10,87752	Unique Row Patterns	27
Model R-Squared	0,36707	Sum of Frequencies	27
Actual Convergence	2,081623E-06	Likelihood Iterations	9
Target Convergence	0,000001	Maximum Iterations	1000
Model D.F.	6	Max Like Message	Quasi-Separation
Model	CELL|SMEAR|INFIL|LI|BLAST|TEMP		

******** WARNING ******** WARNING ******** WARNING ******** WARNING ******** WARNING ********
Your dataset had QUASI-COMPLETE SEPARATION which means that the maximum likelihood routine did NOT converge so the statistical tests are not valid. Although the prediction equations correctly classified much of your data, they may not do so for other observations.
Quasi-Complete Separation often occurs because your sample size is too small.
******** WARNING ******** WARNING ******** WARNING ******** WARNING ******** WARNING ********

Varovné návěští zde v **Max Like Message** říká **quasi-separation**, což znamená, že neexistuje odhady s maximální věrohodností a všechny výsledky jsou proto podezřelé. Proběhlo pouze 9 iterací. Navrhujeme proto opakovat výpočet s menším počtem nezávisle proměnných, místo počtu 6 zadáme nyní pouze 5 v **Max Terms in Subset** v **Modelu.**

Blok **Run Summary Section** ukazuje užitečné informace o proběhlém výpočtu, a to včetně zadání dat a podmínek výpočtu. Je třeba ověřit, zda podmínky byly zadány vhodně pro výpočet. Řada licitovaných nadpisů jsou samobjasňující včetně informace, proto zde budou jenom některé nadpisy vysvětleny:

Reference Group je jedna z nabývaných hodnot (čili kategorie) predikované závisle proměnné. Tato kategorie bývá však vynechávána v celém outputu. Když není **Reference group** u proměnné Y specifikována uživatelem, je zvolena dle defaultního nastavení **Default Reference Group.** Tato hodnota je velice důležitá pro výklad celého zbytku outputu.

Logistic Regression Report

Page/Date/Time	1 22.7.2009 7:58:50
Database	C:\PROGRAM FILES\NCSS2007\DATA\LEUKEMIA.S0
Response	REMISS

Run Summary Section

Parameter	Value	Parameter	Value
Dependent Variable	REMISS	Rows Processed	29
Reference Group	1	Rows Used	27
Number of Groups	2	Rows for Validation	0
Frequency Variable	None	Rows X's Missing	2
Numeric Ind. Variables	6	Rows Freq Miss. or 0	0
Categorical Ind. Variables	0	Rows Prediction Only	0
Final Log Likelihood	-10,87752	Unique Row Patterns	27
Model R-Squared	0,36707	Sum of Frequencies	27
Actual Convergence	2,081623E-06	Likelihood Iterations	9
Target Convergence	0,000001	Maximum Iterations	1000
Model D.F.	6	Max Like Message	Quasi-Separation
Model	CELL|SMEAR|INFIL|LI|BLAST|TEMP		

******** WARNING ******** WARNING ******** WARNING ******** WARNING ******** WARNING ********
Your dataset had QUASI-COMPLETE SEPARATION which means that the maximum likelihood routine did NOT converge so the statistical tests are not valid. Although the prediction equations correctly classified much of your data, they may not do so for other observations.
Quasi-Complete Separation often occurs because your sample size is too small.
******** WARNING ******** WARNING ******** WARNING ******** WARNING ******** WARNING ********

Number of Groups značí počet kategorií čili jednotkových možností závisle proměnné Y, obvykle 0 a 1, takže 2.

Final Log Likelihood značí dosaženou hodnotu logaritmu věrohodnosti po terminaci iteračního výpočtu.

Model R-Squared značí hodnotu R^2 po ukončeném výpočtu.

Actual and Target Convergence značí jednak dosaženou a jednak ukončovací hodnotu při terminaci iteračního regresního procesu, kdy je dosaženo hodnoty maximální věrohodnosti. Výpočet by neměl terminovat, pokud není dosažená hodnota nižší než předvolená ukončovací. Jinak je třeba změnit celé zadání proměnných.

Rows Processed, Used, etc. značí počet řádků zdrojové matice (=objektů, pacientů, pozorování), které byly načteny a použity k výpočtu.

Unique Row Patterns značí počet jednotkových předloh nalezených v obou, závisle a nezávisle proměnných.

Likelihood and Maximum Iterations značí počet iterací k dosažení terminačního kritéria a ukončení výpočtu a také počet předvolených iterací. Nastane-li terminace na nízký počet iterací, je třeba předvolený počet zvýšit. Dosažený počet nižší jak 10 bývá obvyklý při úspěšném výpočtu.

Max Like Message přináší návěští při terminaci výpočtu. Návěští „Normal Competion" je po úspěšně proběhlém výpočtu.

Model D.F. značí počet stupňů volnosti v G-1 logistických regresních modelech.

Model značí zkratkový zápis regresního modelu formou jejich členů nezávisle proměnných, který byl proložen načtenými daty.

Blok **Response Analysis Section:**

Response Analysis Section

REMISS Categories	Count	Unique Rows	Prior	Act vs Pred R-Squared	% Correctly Classified
0	18	18	0,66667	0,38158	83,333
1	9	9	0,33333	0,38158	55,556
Total	27	27			74,074

Tento blok statisticky popisuje závisle proměnnou Y a ukazuje, jak dokonale je závisle proměnná Y aproximována svým zde navrženým regresním modelem.

Categories značí možné jednotkové hodnoty obou kategorií, které může závisle proměnná Y nabývat. U binárních dat jde vždy o dvě hodnoty, a to 0 a 1.

Count značí sumu četností řádků objektů, které může každá z kategorií závisle proměnné Y nabývat.

Unique Rows značí počet jednotkových řádků v každé kategorii, které byly vypočteny z nezávisle proměnných.

Prior značí úvodní pravděpodobnost každé kategorie zadané uživatelem ve volbě v okénku **Prior Probabilities.**

Act vs Před R-Squared značí hodnotu R^2, která je dosažena po regresi, když jsou nezávisle proměnné pro tuto kategorii regresovány na predikovanou pravděpodobnost.

%Correctly Classified značí procento objektů (zde pacientů) v každé kategorii, kteří byli správně klasifikováni jako by to byl vícerozměrný logistický regresní model.

Blok **Subset Selection Summary Section:** Blok ukazuje nejlepší hodnotu logaritmu věrohodnosti pro každý člen. V tomto příkladě se ukazuje, že 4 členy (tj. úsek a tři nezávisle proměnné) vedou na nejlepší model. Všimněte si, že zavedení 5. nezávisle proměnné nezvýší podstatně hodnotu R^2 zde značenou **R-Squared.**

Subset Selection Summary Section

No. Terms	No. X's	Log Likelihood	R-Squared Value	R-Squared Change
1	1	-17,18588	0,00000	0,00000
2	2	-13,03648	0,24144	0,24144
3	3	-12,17036	0,29184	0,05040
4	4	-10,97669	0,36130	0,06946
5	5	-10,92900	0,36407	0,00277

No. Terms značí počet použitých členů (čili nezávisle proměnných). Všimněte si, že tento počet zahrnuje také úsek.

No. X´s značí počet x, které byly zahrnuty do modelu. Tentp počet souhlasí s počtem členů. To ale nebude v případě, když některý ze členů bude kategorická proměnná.

Log Likelihood značí hodnotu maximální věrohodnosti pro nejlepší dosažené odhady parametrů. Našim cílem je nalézt takový soubor proměnných, u kterého přidáním dalších proměnných se tato hodnota zlepší (čili sníží) jenom velmi nepatrně.

R-Squared Value značí hodnotu koeficientu determinace. Našim cílem je nalézt takový soubor proměnných, u kterého přidáním dalších proměnných se tato hodnota zlepší (čili zvýší) jenom velmi nepatrně.

Blok **Subset Selection Detail**

Subset Selection Detail Section

Step	Action	No. of Terms	No. of X's	Log Likelihood	Term Entered	Term Removed
1	Add	1	1	-17,18588	Intercept	
2	Add	2	2	-13,03648	LI	
3	Add	3	3	-12,17036	CELL	
4	Add	4	4	-10,97669	TEMP	
5	Add	5	5	-10,92900	SMEAR	

Blok ukazuje nejlepší hodnotu logaritmu věrohodnosti pro každý člen. V tomto příkladě se ukazuje, že 4 členy (tj. úsek a tři nezávisle proměnné) vedou na nejlepší model. Všimněte si, že zavedení 5. nezávisle proměnné nezvýší podstatně hodnotu R^2 .

Action značí činnost, která se stane v tomto kroku. Člen může být přidán, odebrán nebo přeskočen odebráním.

No. Terms značí počet použitých členů (čili nezávisle proměnných). Všimněte si, že tento počet zahrnuje také úsek.

No. X´s značí počet x, které byly zahrnuty do modelu. Tentp počet souhlasí s počtem členů. To ale nebude v případě, když některý ze členů bude kategorická proměnná.

Log Likelihood značí hodnotu maximální věrohodnosti pro nejlepší dosažené odhady parametrů. Našim cílem je nalézt takový soubor proměnných, u kterého přidáním dalších proměnných se tato hodnota zlepší (čili sníží) jenom velmi nepatrně.

Terms Entered and Removed identifikuje členy přidané, odebrané a přeskočené odebráním.

Závěr: Bylo rozhodnuto, že zařadíme nezávisle proměnné **CELL, LI,** a **TEMP.**

Blok **Parameter Significance Tests Section:**

Parameter Significance Tests Section (Reference Group: REMISS = 1)

Parameter	Regression Coefficient (B or Beta)	Standard Error	Wald Z-Value (Beta=0)	Wald Prob Level	Odds Ratio Exp(B)
B0: Intercept	-70,09918	58,76332	-1,193	0,23291	0,00000
B2: CELL	-9,85060	7,82632	-1,259	0,20816	0,00005
B4: LI	-3,90515	1,81674	-2,150	0,03159	0,02014
B6: TEMP	85,44441	64,21420	1,331	0,18332	10000+
B5: SMEAR	-0,91240	2,95996	-0,308	0,75789	0,40156

Parameter je označení odhadovaného parametru. Spojitý parametr je zobrazen přímo. Je-li diskrétní, je uvedena definice binární proměnné generované. Např. diskrétní nezávisle proměnná GRADE nabývá tří hodnot: A, B, C. Název zde uvedený bude mít třeba podobu B2:GRADE=B. To pak odkazuje na binární proměnnou, u které bude jednička pro řádky, ve kterých je GRADE rovna B zatímco nula bude u ostatních.
Regression Coefficient (B or Beta) značí číselný odhad regresního koeficientu.
Standard Error značí míru přesnosti odhadovaného regresního koeficientu.

Wald Z-Value (Beta=0) značí Waldovo kritérium k testu hypotézy H_0: $B_j=0$ ve formě z_j= odhad parametru/směrodatná odchylka.
Wald Prob Level značí spočtenou hladinu významnosti **p** pro Waldovo kritérium. Je-li **p** menší než **alfa**=0.05, je dotyčný parametr statisticky významný.

Odds Ratio Exp(B) značí **OR** odhad procenta šancí spojeného s tímto regresním koeficientem. Je užitečný pouze pro binární nezávisle proměnné, které mohou nabývat dvou hodnot, 0 a 1. Jde o hodnoty, které jsou vyčísleny pro kategorické nezávisle proměnné dle vztahu **OR**=exp(B).

Blok **Parameter Significance Tests**

Parameter Significance Tests Section (Reference Group: REMISS = 1)

Parameter	Regression Coefficient (B or Beta)	Standard Error	Wald Z-Value (Beta=0)	Wald Prob Level	Odds Ratio Exp(B)
B0: Intercept	-70,09918	58,76332	-1,193	0,23291	0,00000
B2: CELL	-9,85060	7,82632	-1,259	0,20816	0,00005
B4: LI	-3,90515	1,81674	-2,150	0,03159	0,02014
B6: TEMP	85,44441	64,21420	1,331	0,18332	10000+
B5: SMEAR	-0,91240	2,95996	-0,308	0,75789	0,40156

Blok obsahuje odhady logistických regresních rovnic a s nimi spojené statistické testy významnosti. Referenční hodnota je uvedena hned v nadpisu, zde to je REMISS=1. Má-li závisle proměnná více jak dvě možné kategorie, je tato tabulka vyčíslena pro celkem G-1 rovnic.

Blok **Parameter Confidence Limits Section:**

Parameter Confidence Limits Section (Reference Group: REMISS = 1)

Parameter	Regression Coefficient (B or Beta)	Standard Error	Lower 95% Confidence Limit	Upper 95% Confidence Limit	Odds Ratio Exp(B)
B0: Intercept	-70,09918	58,76332	-185,27317	45,07482	0,00000
B2: CELL	-9,85060	7,82632	-25,18991	5,48871	0,00005
B4: LI	-3,90515	1,81674	-7,46588	-0,34441	0,02014
B6: TEMP	85,44441	64,21420	-40,41311	211,30192	10000+
B5: SMEAR	-0,91240	2,95996	-6,71382	4,88902	0,40156

Tento blok obsahuje odhady logistických regresních rovnic a s nimi spojené intervaly spolehlivosti parametrů. Referenční hodnota je uvedena hned v nadpisu, REMISS=1. Má-li závisle proměnná více jak dvě možné kategorie, je tato tabulka uvedena pro celkem G-1 rovnic.

Confidence Limits značí dolní a horní meze intervalového odhadu neznámého koeficientu pro každou nezávisle proměnnou. Jeli dolní mez záporná a horní mez kladná, leží mezi mezemi nula a koeficient je proto nulový čili statisticky nevýznamný. Meze jsou platné stejně jako Waldův test jenom pro velké výběry zdrojové matice dat.

Odds Ratio Exp(B) značí **OR** odhad procenta šancí spojeného s tímto regresním koeficientem. Je užitečný pouze pro binární nezávisle proměnné, které mohou nabývat dvou hodnot, 0 a 1. Jde o hodnoty, které jsou vyčísleny pro kategorické nezávisle proměnné dle vztahu **OR**=exp(B). Kvůli formátovacím problémům tato hodnota se nezobrazí nabyde-li hodnoty větší než 10.000.

Blok **Odds Ratio Estimation Section:**

Odds Ratios Section (Reference Group: REMISS = 1)

Parameter	Regression Coefficient (B or Beta)	Odds Ratio Exp(B)	Lower 95% Confidence Limit	Upper 95% Confidence Limit
B0: Intercept	-70,09918	0,00000	0,00000	10000+
B2: CELL	-9,85060	0,00005	0,00000	241,94443
B4: LI	-3,90515	0,02014	0,00057	0,70864
B6: TEMP	85,44441	10000+	0,00000	10000+
B5: SMEAR	-0,91240	0,40156	0,00121	132,82300

Tento blok obsahuje procento šance a jeho intervaly spolehlivosti.

Parameter je název odhadovaného parametru. Spojitý parametr je zobrazen přímo. Je-li však parametr diskrétní, je zde uvedena definice binární proměnné generované. Např. diskrétní nezávisle proměnná GRADE nabývá tří hodnot: A, B, C. Název zde uvedený bude mít třeba podobu B2:GRADE=B. To pak odkazuje na binární proměnnou, u které bude jednička pro řádky, ve kterých je GRADE rovna B zatímco nula bude u ostatních. Všimněte si, že umístění jména je ovládáno v záložce **Format** ve volbě **Skip Line After.**

Odds Ratio Exp(B) značí **OR** odhad procenta šancí spojeného s tímto regresním koeficientem. Je užitečný pouze pro binární nezávisle proměnné, které mohou nabývat dvou hodnot, 0 a 1. Jde o hodnoty, které jsou vyčísleny pro kategorické nezávisle proměnné dle vztahu **OR**=exp(B).

Confidence Limits značí dolní a horní meze intervalu spolehlivosti procenta šance. Platí podobně jako Waldovo kritérium převážně pro velké výběry. Kvůli formátovacím problémům tato hodnota se nezobrazí nabude-li hodnoty větší než 10.000.

Blok **Estimated Logistic Regression Model(s) Section:**

Estimated Logistic Regression Model(s)
Model For REMISS = 0
-70.0991769634316 -9.85060157750821*CELL -3.90514711780638*LI +85.444405276173*TEMP -.91239944894054*SMEAR

Note that each model estimates B for a specific group, where Logit(Y) = XB.
To calculate a probability, transform the logit using Prob(Y<>group) = 1/(1+Exp(-XB))
or Prob(Y=group) = Exp(-XB)/(1+Exp(-XB)).

Transformation Note:
Regular transformations must be less the 255 characters. If this expression is longer the 255 characters, copy this expression and paste it into a text file, then use the transformation FILE(filename.txt) to access the text file.

Rovnice logistického regresního modelu je v normálním textovém formátu a může být proto použita jako transformační vzorec. Regresní rovnice je zobrazena pro každou ze G-1 kategorií závisle proměnné. Regresní koeficienty jsou zobrazeny v double precision negarantují však nutou přesnost k výpočtu skore (logitu) a predikovaných pravděpodobností.

Konec úlohy