



## Binary Logistic Regression – Binární logistická regrese

### Myšlenka metody

Binární logistická regrese je metoda umožňující odhadnout hodnotu dichotomické proměnné (nabývá pouze dvou hodnot) na základě spojitých i kategorizovaných nezávislých proměnných. V programu *IBM SPSS Statistics* je obsažena v modulu *Regression*. Má široké uplatnění v mnoha oborech i situacích, kdy je třeba předpovídat, zda nastane či nenastane určitá událost (zákazník nebude splácet půjčku, nahlášená pojistná událost je podvod, pacient trpí určitým onemocněním, zákazník si koupí nabízený produkt, student dokončí studium na vysoké škole, u pacienta dojde k návratu onemocnění, poplatník se dopustil daňového úniku apod.). Metoda rovněž umožňuje odpovědět na otázku, které faktory zvyšují, resp. snižují pravděpodobnost sledované události. Na základě nalezeného modelu lze potom provádět predikce pro neznámé případy včetně odhadu pravděpodobnosti výskytu jevu.

Cílem logistické regrese je vytvořit model, který by umožňoval odhadnout na základě lineární kombinace prediktorů kategorii cílové proměnné. Na rozdíl od klasické lineární regrese, která předpokládá spojitou závislou proměnnou, je však závislá proměnná kategorizovaná, což s sebou přináší řadu komplikací. Zatímco levá strana regresní rovnice by v takovém případě mohla nabývat pouze dvou hodnot, pravá strana (lineární kombinace prediktorů) může obecně nabývat libovolných hodnot. Z toho důvodu se ukazuje jako vhodnější předpovídat místo původních kategorií jejich pravděpodobnost, avšak hodnoty pravé strany rovnice je třeba ještě vhodným způsobem transformovat na interval  $<0,1>$ . K tomuto účelu se užívá tzv. logistická funkce:

$$f = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

Logistický regresní model má potom tvar:

$$p(Y=1) = \frac{e^{b_0 + b_1X_1 + \dots + b_nX_n}}{1 + e^{b_0 + b_1X_1 + \dots + b_nX_n}} = \frac{1}{1 + e^{-(b_0 + b_1X_1 + \dots + b_nX_n)}}$$

Při interpretaci regresních koeficientů se velmi často užívá rovněž vyjádření pro tzv. *logit*(Y):

$$\text{logit}(Y) = \ln\left(\frac{p(Y=1)}{1 - p(Y=1)}\right) = b_0 + b_1X_1 + \dots + b_nX_n$$

nebo vztah vyjadřující tzv. *šanci*(Y=1):

$$\text{šance}(Y=1) = \frac{p(Y=1)}{1 - p(Y=1)} = e^{b_0 + b_1X_1 + \dots + b_nX_n} = e^{b_0} * e^{b_1X_1} * \dots * e^{b_nX_n}$$

## Listy procedur IBM SPSS Statistics

Parametry logistického regresního modelu se obvykle odhadují metodou maximální věrohodnosti.

Konstanta ( $b_0$ ) vyjadřuje odhad *logitu*  $Y$  pro situaci, kdy jsou všechny nezávislé proměnné rovny nule. Ostatní regresní koeficienty ( $b_i$ ) vyjadřují, o kolik by se podle modelu změnil *logit*( $Y$ ), jestliže se hodnota dané proměnné zvýší o jednotku a hodnoty ostatních nezávislých proměnných se nezmění. Hodnota *Eulerova čísla* umocněného na tento koeficient ( $e^{b_i}$ ) potom vypovídá o tom, kolikrát se podle odhadu změni *šance*( $Y=1$ ), jestliže se hodnota dané proměnné zvýší o jednotku a ostatní nezávislé proměnné zůstanou konstantní.

Procedura umožňuje přidat do modelu rovněž interakce mezi nezávislými proměnnými a nabízí různé typy transformací kategorizovaných prediktorů na kontrasty.

Model předpokládá nezávislost pozorování a nezávislost skutečných hodnot a chyb. Logistická regrese nevychází z tak striktních předpokladů o rozložení proměnných jako diskriminační analýza, řešení však může být stabilnější, pokud mají nezávislé proměnné mnohorozměrné normální rozložení. Kategorizované nezávislé proměnné je třeba převést na indikátory nebo na jiný typ kontrastů. Pro úspěšnou aplikaci metody je rovněž třeba zajistit dostatečný počet případů. Podobně jako u jiných regresních modelů může zkreslovat odhady i standardní chyby kolinearita prediktorů.

Logistická regrese je značně náročná na přípravu dat. Před jejím užitím je třeba ošetřit extrémy a vynechané hodnoty v datech (metoda nepracuje s vynechanými hodnotami) a zvážit případnou kategorizaci číselných prediktorů do malého počtu tříd. Vzhledem k poměrně komplikovanému algoritmu a jeho numerické nestabilitě, se doporučuje budovat model postupně. Metoda je rovněž velmi citlivá na nevyvážené kategorie závislé proměnné (v případě potřeby lze pro účely modelování kategorie vyvážit).

Procedura nabízí široký výběr nástrojů pro vyhodnocení kvality modelu, testování významnosti jednotlivých koeficientů, diagnostiku problematických případů a další. K dispozici jsou rovněž metody pro automatický výběr prediktorů. Vybrané informace lze uložit do datové matice nebo celý model exportovat do formátu XML.

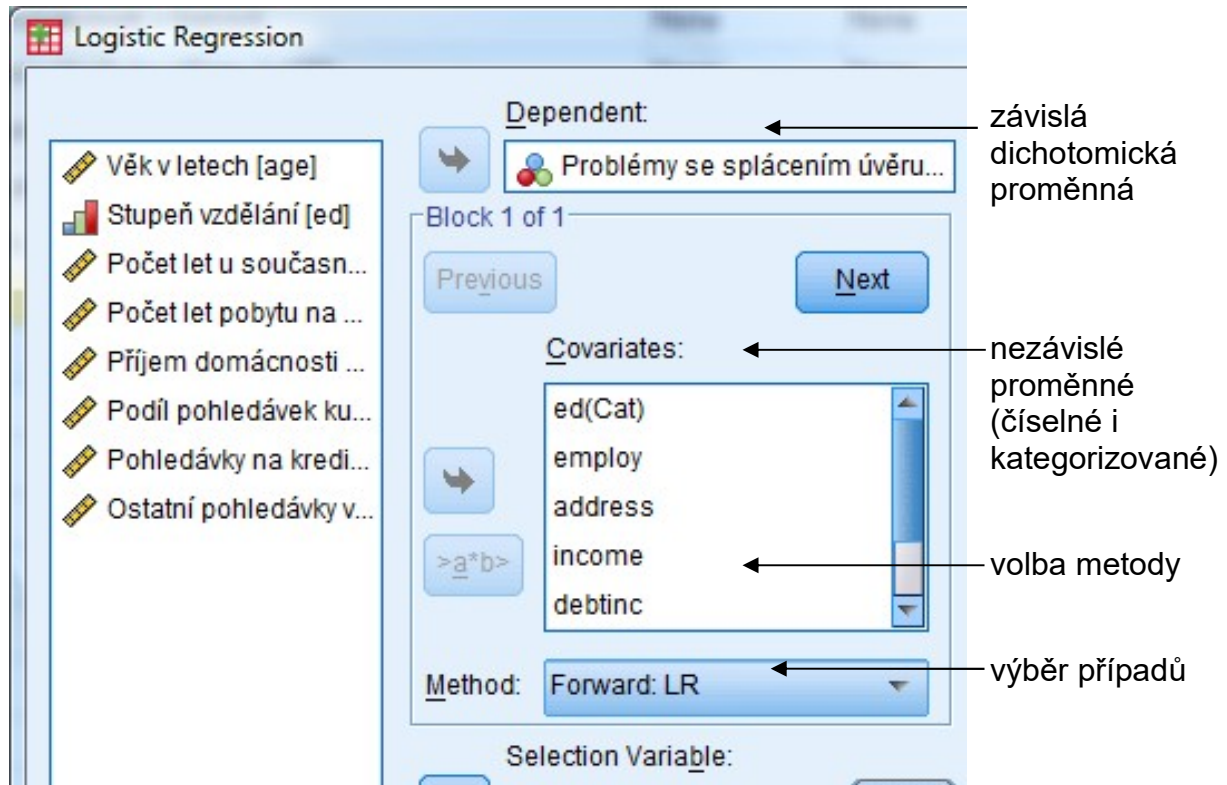
Binární logistická regrese předpokládá dichotomickou závislou proměnnou. Při více kategoriích je třeba užít její zobecnění – multinomickou logistickou regresi (v programu *IBM SPSS Statistics* je obsažena v modulu *Regression* v nabídce *Analyze, Regression, Multinomial Logistic*). Pro případ ordinální závislé proměnné je určena ordinální regrese (modul *Statistics Base*, nabídka *Analyze, Regression, Ordinal*).

Z hlediska řešení úlohy má logistická regrese blízko k diskriminační analýze. Detailnější srovnání těchto metod je k dispozici v příloze (viz *Příloha 1, Srovnání logistické regrese a diskriminační analýzy*).

## Volání procedury v IBM SPSS Statistics

Analyze → Regression → Binary Logistic...

### Nastavení dialogu



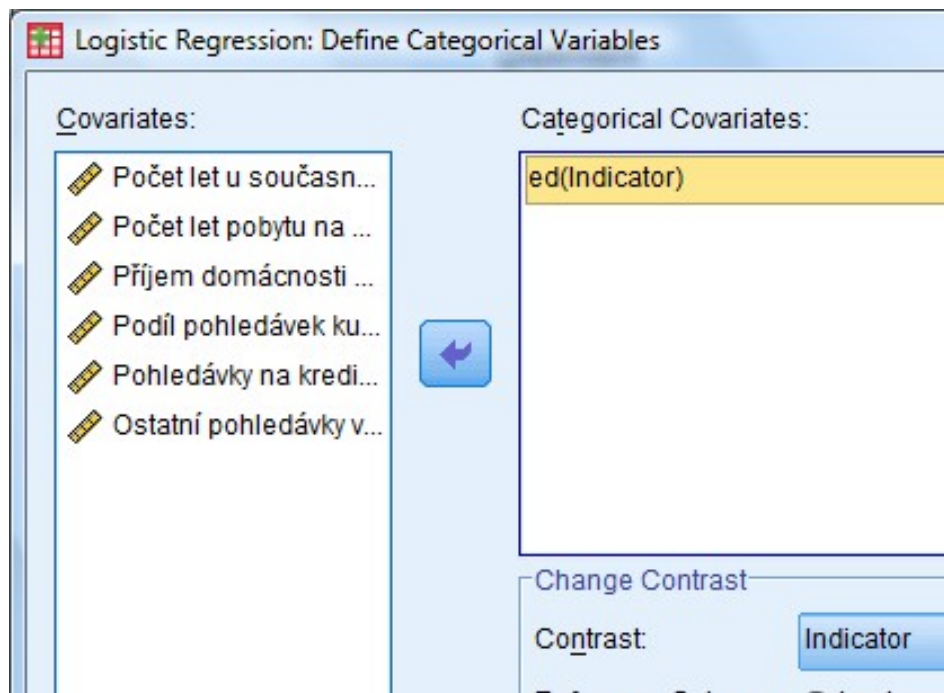
- Do pole *Dependent* přeneseme závislou dichotomickou proměnnou.
- Do pole *Covariates* zadáme nezávislé proměnné, které mohou být číselné i kategorizované. Pro kategorizované proměnné je však třeba určit pomocí tlačítka *Categorical* vhodný způsob transformace. Tlačítko *a\*b* dále umožňuje přidat do modelu interakce mezi dvěma nebo více nezávislými proměnnými (v seznamu proměnných označíme příslušné proměnné a tlačítkem *a\*b* je převedeme do pole *Covariates*). Proměnné lze rovněž rozdělit do několika skupin (bloků) a v každém z nich například zadat jinou metodu automatického výběru prediktorů (viz rozbalovací seznam *Method*). Pro přecházení mezi bloky jsou určena tlačítka *Previous* a *Next*.
- V rozbalovacím seznamu *Method* zvolíme některou z metod pro automatický výběr prediktorů, které pomáhají zajistit, aby model obsahoval všechny důležité prediktory, ale nebyl přeuročen:
  - *Enter* – model se všemi zadanými nezávislými proměnnými.
  - *Forward Conditional* - do modelu se postupně přidávají proměnné a současně se v každém kroku kontroluje, zda není možné odebrat některou z již zařazených proměnných jako nadbytečnou. Kritériem pro přidání proměnné do modulu je dosažená hladina významnosti

## Listy procedur IBM SPSS Statistics

statistiky *Score*, pro odebrání proměnné z modelu test poměrem věrohodností založený na podmíněném odhadu parametrů.

- *Forward LR* – obdoba předchozí metody, avšak jako kritérium pro odebrání proměnné z modelu se užívá test poměrem věrohodností.
- *Forward Wald* - obdoba předchozí metody, avšak jako kritérium pro odebrání proměnné z modelu se užívá Waldova statistika.
- *Backward Conditional* - z modelu se všemi zadanými prediktory se v každém kroku ubírá nejvíce nadbytečná proměnná tak dlouho, dokud zhoršení modelu nepřekročí stanovenou hranici. Současně se v každém kroku kontroluje, zda není třeba do modelu znovu přidat některou z dříve vyřazených proměnných. Kritériem pro odebrání proměnné z modelu je test poměrem věrohodností založený na podmíněném odhadu parametrů, pro přidání proměnné do modelu dosažená hladina významnosti statistiky *Score*.
- *Backward LR* - obdoba předchozí metody, avšak jako kritérium pro odebrání proměnné z modelu se užívá test poměrem věrohodností.
- *Backward Wald* - obdoba předchozí metody, avšak jako kritérium pro odebrání proměnné z modelu se užívá Waldova statistika.
- Do pole *Selection Variable* lze zadat proměnnou, která určuje případy vstupující do modelu. To je užitečné především tehdy, když je třeba data rozdělit na dvě skupiny – na první z nich model vytvoříme a na druhé testujeme jeho kvalitu. Do příslušného pole zadáme proměnnou definující výběr a tlačítkem *Rule* specifikujeme podmínku (model bude vytvořen pouze na případech splňujících podmínku, avšak při klasifikaci se úspěšnost obou skupin porovnává).

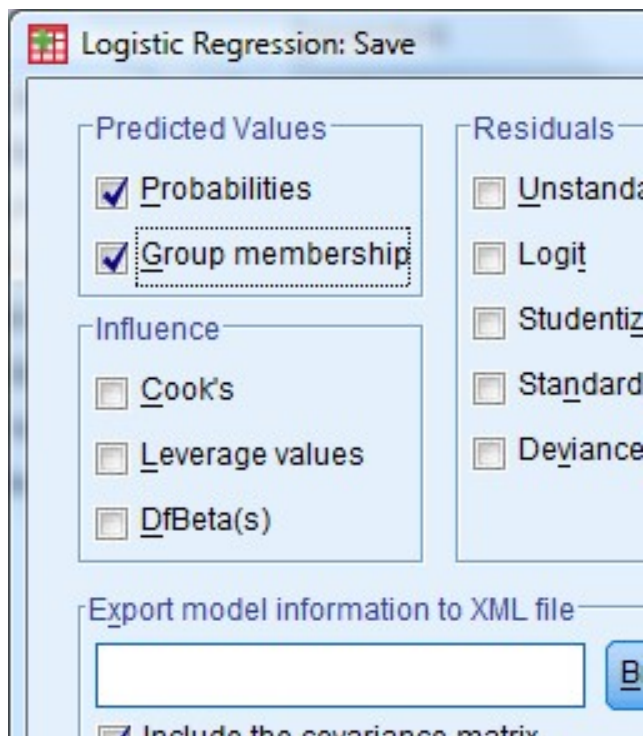
## Tlačítko *Categorical*



Tlačítko *Categorical* umožňuje nastavit vhodné typy transformací pro kategorizované nezávislé proměnné.

- Pole *Covariates* obsahuje seznam všech nezávislých proměnných definovaných v hlavním dialogu. Proměnné, se kterými má být zacházeno jako s kategorizovanými, přeneseme do pole *Categorical Covariates*.
- V části *Change Contrast* definujeme pro označenou proměnnou (proměnné) z pole *Categorical Covariates* typ kontrastu a v případě potřeby specifikujeme také referenční kategorii: indikační proměnné (*Indicator*), jednoduché kontrasty (*Simple*), diferenční kontrasty (*Difference*), Helmertovy kontrasty (*Helmert*), porovnání sousedních kategorií (*Repeated*), polynomické kontrasty (*Polynomial*) nebo odchylkové kontrasty (*Deviation*). V části *Reference Category* určíme, zda referenční kategorie bude poslední (*Last*), nebo první (*First*) kategorie. Tlačítkem *Change* potvrdíme celé zadání. Podrobnější informace o jednotlivých typech kontrastů viz příloha (*Příloha 2, Schémata kódování kategorizovaných proměnných*).

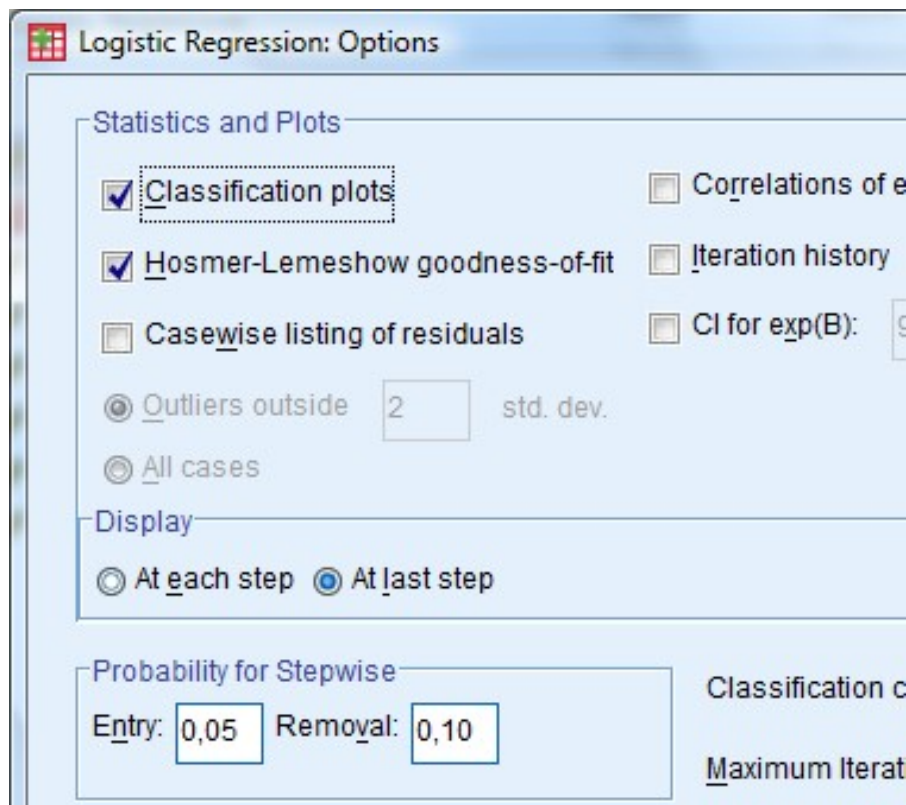
## Tlačítko Save



Tlačítkem **Save** specifikujeme, které informace budou uloženy do datového souboru jako nové proměnné, a případně rozhodujeme o exportu celého modelu do jazyka XML.

- Předpovídané hodnoty (*Predicted Values*): odhad pravděpodobnosti s jakou nastane sledovaná událost pro daný případ (*Probabilities*) a předpovídaná kategorie (*Group membership*).
- Statistiky vlivu (*Influence*): *Cookova vzdálenost* (*Cook's*) – vyjadřuje, jak se změnil rezidua všech případů při vyloučení daného případu, *Leverage values* – charakterizují vliv případu na předpovídané hodnoty, *DfBeta(s)* – míry charakterizující vliv vyloučení případu na odhad jednotlivých koeficientů
- Rezidua (*Residuals*): nestandardizovaná (*Unstandardized*), logitová (*Logit*), studentizovaná (*Studentized*), standardizovaná (*Standardized*), deviance (*Deviance*).
- Export modelu do XML (*Export model information to XML file*): tlačítkem *Browse* definujeme název a umístění souboru a pomocí zaškrtnutí políčka *Include the covariance matrix* rozhodneme, zda má být rovněž exportována kovarianční matice.

## Tlačítko Options



Tlačítkem *Options* nastavíme parametry modelu a požadované výstupy:

- V části *Statistics and Plots* volíme tabulkové a grafické výstupy modelu:
  - klasifikační graf (*Classification plots*)
  - test dobré shody Hosmera a Lemenshowa (*Hosmer-Lemenshow goodness-of-fit*)
  - výpis reziduí *Casewise listing of residuals* všech případů (*All cases*) nebo jen případů, kde hodnota rezidua překročí zvolený násobek směrodatné odchylky (*Outlier outside ... std. dev.*)
  - korelační matice regresních koeficientů (*Correlations of estimates*)
  - historie iterací (*Iteration history*)
  - intervaly spolehlivosti pro  $\exp(B)$  při zvolené hladině spolehlivosti (*CI for  $\exp(B)$* ).
- V poli *Display* určíme, zda v případě užití některé z metod automatického výběru prediktorů mají tabulky a grafy ve výstupu dokumentovat všechny postupné kroky metody (*At each step*) nebo pouze konečný model (*At last step*).
- V části *Probability for Stepwise* lze nastavit pro metody automatického výběru prediktorů dosaženou hladinu významnosti příslušného kritéria pro vstup (*Entry*) a výstup (*Removal*) proměnné do/z modelu.



## *Listy procedur IBM SPSS Statistics*

- Pole *Classification cutoff* definuje hranici (dělicí bod) pravděpodobnosti výskytu jevu pro klasifikaci případů (hodnota z intervalu  $<0,1>$ ). Při defaultním nastavení  $0,5$  jsou jako pozitivní klasifikovány případy, pro něž je odhadnutá pravděpodobnost, že jev nastane, větší než  $0,5$ . V mnoha praktických aplikacích je však nutné tuto hranici posunout (například při rozhodování o tom, kteří pacienti podstoupí další vyšetření, může být vzhledem k závažnosti možného onemocnění stanovena nižší hranice pravděpodobnosti, tj. vyšetření podstoupí i pacienti, u nichž je odhadované riziko onemocnění menší).
- Pole *Maximum Iterations* specifikuje maximální počet iterací, které proběhnou při výpočtu.
- Zaškrtnuté políčko *Include constant in model* určuje, zda bude model zahrnovat konstantní člen.



### Výstupy

Zaměstnanci banky, kteří rozhodují o schválení resp. zamítnutí žádostí o úvěr, potřebují znát charakteristiky, které indikují vyšší pravděpodobnost problémů při splácení úvěru, aby mohli odhadnout úvěrové riziko.

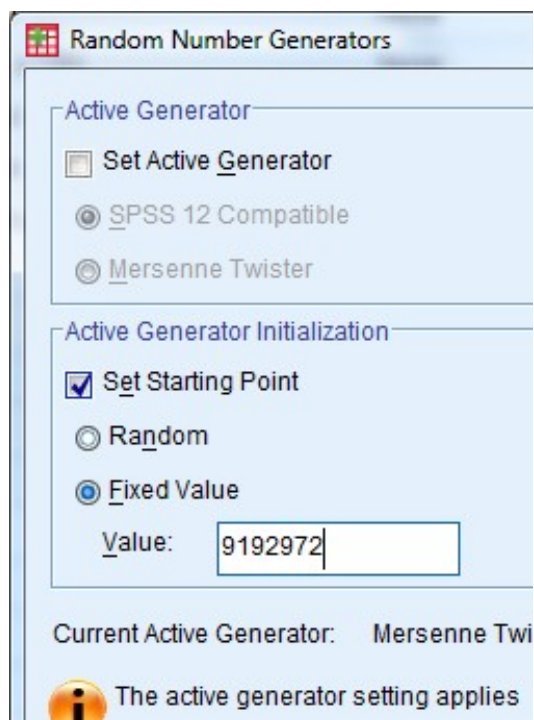
Datový soubor obsahuje informace o 850 minulých i potenciálních zákaznících. Prvních 700 případů představují zákazníci, kteří již v minulosti měli půjčku a o kterých víme, zda došlo při splácení k problémům. Tyto případy budou základem pro vybudování logistického regresního modelu a pro ověření jeho kvality. Následně model využijeme pro odhad úvěrového rizika nových žadatelů.

O každém zákazníkovi jsou k dispozici tyto informace: *věk v letech (age)*, *stupeň vzdělání (ed)*, *počet let u současného zaměstnavatele (employ)*, *délka pobytu na současné adrese v letech (address)*, *příjem domácnosti v tisících (income)*, *podíl pohledávek vzhledem k příjmu (x100) (debtinc)*, *pohledávky na kreditní kartě v tisících (creddebt)*, *ostatní pohledávky v tisících (othdebt)* a dále pro minulé zákazníky informace o tom, zda nastaly *problémy se splácením úvěru (default)*.

Z množiny zákazníků, kteří již v minulosti měli úvěr, náhodně vybereme přibližně 70% případů, na základě kterých model vytvoříme. Zbýlých 30% případů potom využijeme pro ověření jeho kvality.

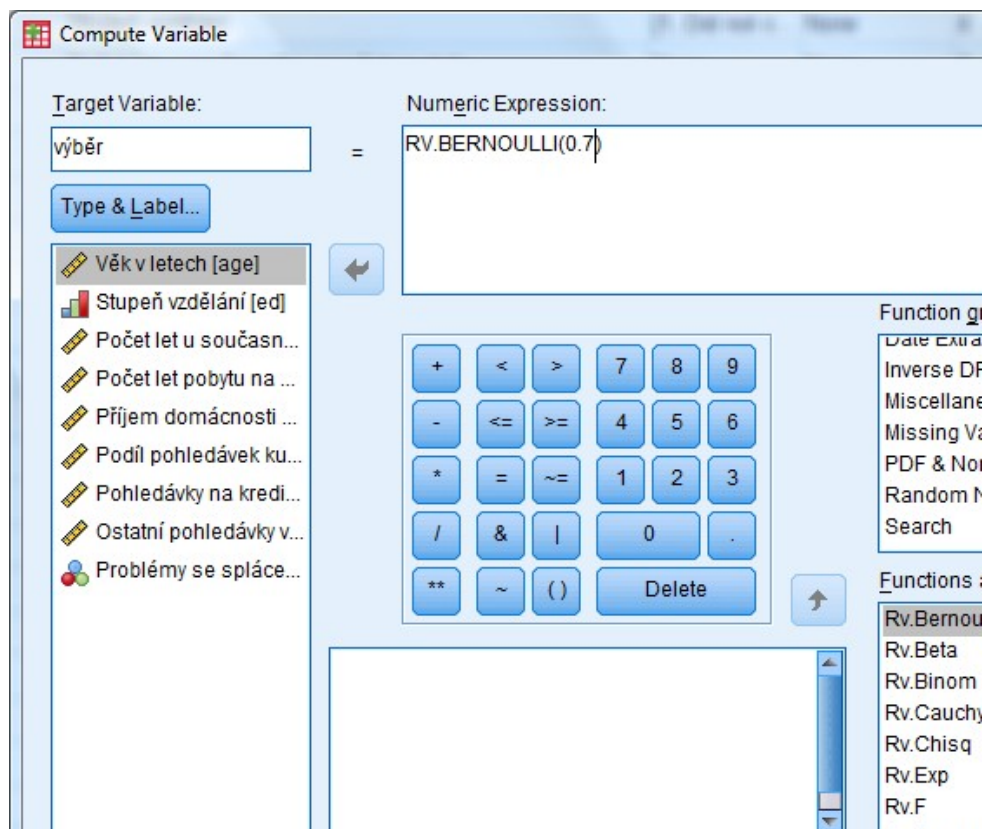
### Nastavení generátoru náhodných čísel

Aby bylo možné v případě potřeby znovu zopakovat stejné výsledky, nastavíme nejprve počáteční hodnotu generátoru náhodných čísel. V hlavním menu v nabídce *Transform, Random Number Generators* zaškrtneme *Set Starting Point*, označíme *Fixed Value* a do pole *Value* zadáme hodnotu 9191972 (viz následující obrázek).



### Rozdělení případů na trénovací a testovací množinu

Dále odvodíme pomocí nabídky *Transform, Compute Variable* proměnnou s názvem *výběr*, která představuje náhodný výběr přibližně 70% zákazníků, kteří již v minulosti úvěr měli. Do pole *Target Variable* zadáme název proměnné a do pole *Numeric Expression* výraz *RV.BERNOULLI(0.7)*. Tímto způsobem program náhodně generuje hodnoty z alternativního rozložení s parametrem  $p=0,7$ . Pomocí tlačítka *IF* dále zadáme podmínku *MISSING(default)=0*, která doplňuje, že budeme pracovat pouze s minulými zákazníky, u kterých víme, zda došlo k problémům se splácením úvěru.



Následující výstupy již byly získány pomocí nabídky *Analyze, Regression, Binary Logistic*.

### Model logistické regrese

Pomocí logistické regrese se pokusíme vytvořit model, který bude odhadovat pravděpodobnost, že nastanou problémy se splácením úvěru. Cílová proměnná *default* je tedy binární s hodnotami  $0=ne$  a  $1=ano$ .

Jako nezávislé zadáme všechny ostatní proměnné, které jsou k dispozici, avšak označíme metodu *Forward LR* pro automatický výběr prediktorů. Z důvodu přehlednosti nastavíme zobrazování výstupních tabulek a grafů pouze pro poslední krok metody.

### Přehled o počtu zahrnutých a vynechaných případů

Case Processing Summary

Unweighted Cases <sup>a</sup>		N	Percent
Selected Cases	Included in Analysis	487	57,3
	Missing Cases	0	,0
	Total	487	57,3
Unselected Cases		363	42,7
Total		850	100,0

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

Tabulka *Case Processing Summary* shrnuje informace o počtu zahrnutých a vynechaných případů. Model bude vytvořen na základě 487 případů (*Selected Cases, Included in Analysis*), z nichž žádný neobsahuje chybějící hodnoty (*Missing Cases*). Mezi nevybranými případy (*Unselected Cases*) jsou jednak ty, které budou užity pro ověření kvality modelu, jednak ty, které reprezentují nové žadatele o úvěr.

### Kódování závislé proměnné

Dependent Variable Encoding

Original Value	Internal Value
ne	0
ano	1

Tabulka *Dependent Variable Encoding* poskytuje přehled hodnot závislé proměnné. V tomto případě se tedy jedná o binární proměnnou s kategoriemi  $0=ne$  a  $1=ano$ .

## Kódování kategorizovaných proměnných

Categorical Variables Codings

		Frequency	Parameter coding			
			(1)	(2)	(3)	(4)
Stupeň vzdělání	Did not complete high school	255	1,000	,000	,000	,000
	High school degree	146	,000	1,000	,000	,000
	Some college	61	,000	,000	1,000	,000
	College degree	21	,000	,000	,000	1,000
	Post-undergraduate degree	4	,000	,000	,000	,000

Tabulka *Categorical Variables Codings* informuje o tom, jakým způsobem vstupují do modelu kategorizované nezávislé proměnné (tj. jaký typ transformace na kontrasty byl zvolen). Jedná se o sloupce designové matice (mimo její první sloupec), viz příloha (*Příloha 2, Schémata kódování kategorizovaných proměnných*).

V tomto případě byla užita transformace na indikační proměnné a poslední kategorie byla zvolena jako referenční. Celkem jsou tedy odvozeny čtyři nové proměnné. První transformační proměnná je rovna jedné pro první kategorii a nule pro ostatní kategorie. Obdobně druhá resp. třetí proměnná jsou rovny jedné pro druhou resp. třetí kategorii a nule pro ostatní kategorie atd.

## Model bez nezávislých proměnných (Block 0: Beginning Block)

Část *Block 0: Beginning Block* se vztahuje k modelu, který neobsahuje žádné nezávislé proměnné pouze absolutní člen. Konečný model se potom vzhledem k tomuto porovnává (jeho předpovědi by měly být výrazně lepší).

## Klasifikační tabulka

Classification Table<sup>d, e</sup>

Observed			Predicted				
			Selected Cases <sup>a</sup>			Unselected Cases <sup>b, c</sup>	
			Problémy se splácením úvěru		Percentage Correct	Problémy se splácením úvěru	
			ne	ano		ne	ano
Step 0	Problémy se splácením úvěru	ne	363	0	100,0	154	0
		ano	124	0	,0	59	0
Overall Percentage					74,5		72,3

a. Selected cases výběr EQ 1

b. Unselected cases výběr NE 1

c. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

d. Constant is included in the model.

e. The cut value is .500

Klasifikační tabulka je rozdělena na dvě části, z nichž první se vztahuje k trénovací množině (*Selected Cases*) a druhá k testovací množině (*Unselected Cases*). Jedná se o dvě kontingenční tabulky umístěné vedle sebe, kde řádky reprezentují skutečné hodnoty v datech, sloupce odhadovanou kategorií na základě logistického regresního modelu.

## Listy procedur IBM SPSS Statistics

Zde se tabulka vztahuje k modelu obsahujícímu pouze absolutní člen, proto předpověď vychází jenom z proporcionálního zastoupení skupin na celé testovací množině a vždy předpovídá čtenější kategorii. Protože zákazníkům, kteří mají problémy se splácením úvěru je méně, model pro všechny případy předpovídá, že problémy se splácením nenastanou. Kategorie „ne“ je tedy předpovídána správně vždy (100%), zatímco kategorie „ano“ nikdy (0%). Celková úspěšnost na trénovací množině je 74,5%, na testovací množině 72,3%.

### Proměnné zahrnuté do modelu

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-1,074	,104	106,637	1	,000	,342

Tabulka *Variables in the Equation* v tomto případě obsahuje pouze jeden řádek, který odpovídá absolutnímu členu. Sloupce postupně reprezentují: odhadnutou hodnotu koeficientu (*B*) a jeho standardní chybu (*S.E.*), Waldovu statistiku (*Wald*) pro testování nulovosti koeficientu, její stupně volnosti (*df*) a dosaženou hladinu významnosti (*Sig.*), a Eulerovo číslo umocněné na odhad koeficientu (*Exp(B)*).

Zde je odhadovaná hodnota konstanty -1,074. Záporná hodnota koeficientu (a rovněž *Exp(B)* menší než jedna) vypovídají o tom, že pravděpodobnost kategorie 1 (problémy se splácením úvěru) je menší než pravděpodobnost kategorie 0.

### Proměnné nezahrnuté do modelu

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	ed	10,780	4	,029
		ed(1)	8,414	1	,004
		ed(2)	1,200	1	,273
		ed(3)	2,953	1	,086
		ed(4)	3,499	1	,061
		employ	37,303	1	,000
		address	11,683	1	,001
		income	,745	1	,388
		debtinc	78,959	1	,000
		creddebt	35,332	1	,000
		othdebt	11,570	1	,001
	Overall Statistics		150,471	10	,000

Tabulka *Variables not in the Equation* nabízí přehled proměnných, které do modelu nebyly zahrnuty jako prediktory. Sloupce obsahují postupně: hodnotu statistiky *Score*, která se rovněž užívá pro testování nulovosti koeficientu, její stupně volnosti (*df*) a dosaženou hladinu významnosti (*Sig.*)

## Konečný model získaný pomocí metody *Forward LR* (Block 1: Method = *Forward Stepwise* (Likelihood Ratio))

Následující výstupy se již vztahují ke konečnému modelu navrženému na základě metody *Forward LR*.

### Souhrnné testy koeficientů modelu

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 4	Step	13,299	1	,000
	Block	184,033	4	,000
	Model	184,033	4	,000

Tabulka *Omnibus Tests of Model Coefficients* obsahuje výsledky testu chí-kvadrát, který je obdobou *F* testu v lineární regresi. Umožňuje testovat, zda takto definovaný model celkově má smysl. Nulová hypotéza je formulovaná tak, že všechny koeficienty  $b_1, \dots, b_k$  v regresní rovnici (mimo absolutní člen) jsou nulové. Testujeme ji proti alternativní hypotéze, že alespoň jeden z těchto koeficientů je nenulový.

Tabulka obsahuje celkem tři testy: pro aktuální krok metody automatického výběru prediktorů (*Step*), pro blok proměnných (*Block*) a pro celý model (*Model*). První dva řádky však mají smysl pouze v případě, že byla zadána některá z metod pro automatický výběr proměnných, resp. bloky proměnných. Testová statistika *chí-kvadrát* (*Chi-square*) je založena na rozdílu statistiky *-2Log likelihood* daného modelu a modelu vzhledem ke kterému je prováděno srovnání (pro celý model se provádí srovnání s modelem obsahujícím pouze absolutní člen, pro aktuální krok vzhledem k předchozímu kroku, pro blok proměnných vzhledem k modelu bez této skupiny proměnných). Stupně volnosti potom odpovídají rozdílu v počtu parametrů srovnávaných modelů. Pro rozhodnutí o zamítnutí nebo nezamítnutí nulové hypotézy je podstatný poslední sloupec tabulky (*Sig.*), který udává dosaženou hladinu významnosti testu. V tomto případě tedy nulovou hypotézu zamítáme jak pro celý model, tak pro čtvrtý krok metody *Forward LR* (bloky nebyly definovány).

### Shrnutí informací o modelu

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
4	368,570 <sup>a</sup>	,315	,464

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Tabulka *Model Summary* shrnuje základní informace o kvalitě konečného modelu: statistiku *-2Log likelihood* a dva různé koeficienty navržené jako analogie ke koeficientu determinace z lineární regrese:  $R^2$  Coxové a Snella (*Cox&Snell R Square*) a  $R^2$  Nagelkerka (*Nagelkerke R Square*). Oba posledně zmíněné koeficienty mají za cíl vyjádřit podíl vysvětlené variability v logistickém regresním modelu, avšak

variabilita v logistické regresi musí být definována jiným způsobem než v lineární regresi. Určitou nevýhodou prvního z nich je, že nemůže dosáhnout maximální hodnoty 1, což se pokusil vyřešit Nagelkerke modifikací tohoto koeficientu.

### Test dobré shody Hosmera a Lemenshowa

Test dobré shody Hosmera a Lemenshowa představuje další možnost, jak posoudit vhodnost logistického regresního modelu. Testuje hypotézu shody modelu s daty. Případy jsou rozděleny do deseti přibližně stejně velkých skupin, které vycházejí z decilů založených na odhadované pravděpodobnosti sledovaného jevu. V těchto skupinách jsou dále porovnávány pozorované a očekávané četnosti<sup>1</sup>. Testová statistika (*Chi-square*) se získá jako součet druhých mocnin reziduí ve skupinách dělených očekávanými četnostmi, počet stupňů volnosti (*df*) odpovídá počtu skupin mínus dva. Pro užití tohoto testu se předpokládá, že očekávané četnosti ve skupinách jsou větší než jedna a většina z nich (nejméně 80%) je větších než 5. Test dobré shody Hosmera a Lemenshowa poskytuje zajímavou informaci o modelu, k jeho interpretaci je však třeba přistupovat velmi opatrně, neboť hodnota testové statistiky je úměrná velikosti výběru a její hodnota tedy může být velká i v případě dobrého modelu.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
4	5,486	8	,705

Na základě dosažené hladiny významnosti tohoto testu (*Sig.*) v tomto případě nezamítáme nulovou hypotézu o shodě logistického regresního modelu s daty.

---

<sup>1</sup> Vzhledem k tomu, že hodnota testové statistiky závisí na tom, jak přesně jsou případy rozděleny do skupin, mohou v některých případech různé softwarové nástroje poskytovat mírně odlišné výsledky.



**Contingency Table for Hosmer and Lemeshow Test**

		Problémy se splácením úvěru = ne		Problémy se splácením úvěru = ano		Total
		Observed	Expected	Observed	Expected	
Step 4	1	48	48,013	1	,987	49
	2	48	46,551	1	2,449	49
	3	46	44,831	3	4,169	49
	4	43	43,299	6	5,701	49
	5	40	41,224	9	7,776	49
	6	39	38,514	10	10,486	49
	7	35	34,661	14	14,339	49
	8	30	30,652	19	18,348	49
	9	19	24,116	30	24,884	49
	10	15	11,140	31	34,860	46

Tabulka *Contingency Table for Hosmer and Lemeshow Test* představuje kontingenční tabulku, ze které vychází test dobré shody Hosmera a Lemeshowa. Řádky reprezentují jednotlivé skupiny (decily podle předpovídané pravděpodobnosti), sloupce jsou rozděleny podle kategorií závislé proměnné a dále podle zobrazované statistiky: pozorované četnosti (*Observed*) a očekávané četnosti (*Expected*). Sloupec *Total* vyjadřuje celkový počet případů ve skupině.

## Klasifikační tabulka

**Classification Table<sup>d</sup>**

Observed			Predicted				
			Selected Cases <sup>a</sup>			Unselected Cases <sup>b, c</sup>	
			Problémy se splácením úvěru		Percentage Correct	Problémy se splácením úvěru	Percentage Correct
			ne	ano		ne	ano
Step 4	Problémy se splácením úvěru	ne	336	27	92,6	144	10
		ano	58	66	53,2	33	26
Overall Percentage					82,5		

a. Selected cases výběr EQ 1

b. Unselected cases výběr NE 1

c. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

d. The cut value is .500

Klasifikační tabulka pro konečný model má obdobnou strukturu jako pro model obsahující pouze absolutní člen. Je rozdělena na dvě části, z nichž první se vztahuje k trénovací množině (*Selected Cases*) a druhá k testovací množině (*Unselected Cases*). Jedná se o dvě kontingenční tabulky umístěné vedle sebe, kde řádky reprezentují skutečné hodnoty v datech, sloupce odhadovanou kategorii na základě logistického regresního modelu.

Z tabulky vyplývá, že na trénovací množině bylo správně předpovězeno 92,6% případů, u nichž ve skutečnosti problém se splácením nenastal a 53,2% případů, u

## Listy procedur IBM SPSS Statistics

nichž problém nastal. Celková úspěšnost předpovědi je zde 82,5%. Na testovací množině je nepatrně lepší výsledek pro případy, u nichž problém se splácením nenastal (93,5%), avšak o něco horší předpověď problematických případů (44,1%) i celkový počet správných předpovědí (79,8%).

### Proměnné zahrnuté do modelu

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 4 <sup>a</sup>	employ	-,261	,036	52,185	1	,000	,771
	address	-,088	,025	12,104	1	,001	,916
	debtinc	,091	,022	16,680	1	,000	1,095
	creddebt	,637	,111	33,222	1	,000	1,891
	Constant	-,867	,304	8,115	1	,004	,420

a. Variable(s) entered on step 4: address.

Tabulka *Variables in the Equation* obsahuje odhady koeficientů konečného logistického regresního modelu a další doplňující statistiky. Sloupce postupně reprezentují: odhady koeficientů (*B*) a jejich standardní chyby (*S.E.*), Waldovu statistiku pro testování nulovosti koeficientů (*Wald*), její stupně volnosti (*df*) a dosaženou hladinu významnosti (*Sig.*), a Eulerovo číslo umocněné na odhad koeficientu (*Exp(B)*).

V tomto případě má tedy rovnice tvar:

$$\text{logit}(Y) = \ln(p(Y=1)/p(Y=0)) = -0,867 - 0,261 \cdot \text{employ} - 0,088 \cdot \text{address} + 0,091 \cdot \text{debtinc} + 0,637 \cdot \text{creddebt}$$

nebo

$$\begin{aligned} \text{šance}(Y=1) &= p(Y=1)/p(Y=0) = e^{-0,867 - 0,261 \cdot \text{employ} - 0,088 \cdot \text{address} + 0,091 \cdot \text{debtinc} + 0,637 \cdot \text{creddebt}} = \\ &= e^{-0,867} \cdot e^{-0,261 \cdot \text{employ}} \cdot e^{-0,088 \cdot \text{address}} \cdot e^{0,091 \cdot \text{debtinc}} \cdot e^{0,637 \cdot \text{creddebt}} \end{aligned}$$

nebo

$$p(Y=1) = 1/(1 + e^{-( -0,867 - 0,261 \cdot \text{employ} - 0,088 \cdot \text{address} + 0,091 \cdot \text{debtinc} + 0,637 \cdot \text{creddebt} )})$$

Přitom konstanta  $-0,867$  vyjadřuje odhad *logitu* *Y* pro situaci, kdy jsou všechny nezávislé proměnné rovny nule. Ostatní regresní koeficienty vyjadřují, o kolik by se podle modelu změnil *logit(Y)*, jestliže se hodnota dané proměnné zvýší o jednotku a ostatní proměnné zůstanou konstantní.

Z druhé rovnice rovněž vyplývá, že koeficienty *Exp(B)* vyjadřují, kolikrát se podle odhadu změní *šance(Y=1)*, jestliže se hodnota dané proměnné zvýší o jednotku a ostatní proměnné zůstanou konstantní. V tomto případě model naznačuje, že šance problémů se splácením úvěru se zvyšuje s vyššími hodnotami u proměnných *debtinc* a *creddebt* (*Exp(B)* je větší než jedna) a naopak s nižšími hodnotami u proměnných *employ* a *address* (*Exp(B)* je menší než jedna). Nejvíce rizikovní zákazníci jsou tedy ti, kteří pracují krátkou dobu pro současného zaměstnavatele, bydlí krátce na současné adrese a mají vysoké pohledávky (absolutně i vzhledem ke svému příjmu).

## Listy procedur IBM SPSS Statistics

*Waldova* statistika se užívá k testování nulovosti jednotlivých koeficientů v modelu logistické regrese (za platnosti nulové hypotézy má rozdělení chí-kvadrát). Její užití však předpokládá dostatečně velký datový soubor. Pro kategorizované proměnné lze spočítat *Waldovu* statistiku pro jednotlivé kontrastní proměnné samostatně i pro kategorizovanou proměnnou jako celek. *Waldova* statistika má však rovněž některé nežádoucí vlastnosti. Pro regresní koeficienty s velkou absolutní hodnotou je standardní chyba odhadu příliš velká, což vede k tomu, že snadno zamítneme hypotézu o nulovosti koeficientu, ačkoliv ve skutečnosti nemusí mít velký význam. V takových situacích se tedy spíše doporučuje vycházet ze změny statistiky *-2Log likelihood* při zahrnutí proměnné do modelu.

### Proměnné nezahrnuté do modelu

Variables not in the Equation			Score	df	Sig.
Step 4	Variables	ed	4,046	4	,400
		ed(1)	,614	1	,433
		ed(2)	,187	1	,666
		ed(3)	1,155	1	,283
		ed(4)	1,485	1	,223
		income	,502	1	,479
		othdebt	1,500	1	,221
	Overall Statistics		5,341	6	,501

Tabulka *Variables not in the Equation* nabízí přehled proměnných, které do konečného modelu nebyly zahrnuty jako prediktory. Sloupce obsahují postupně: hodnotu statistiky *Score*, která se rovněž užívá pro testování nulovosti koeficientu, její stupně volnosti (*df*) a dosaženou hladinu významnosti (*Sig.*). Statistika *Score* je alternativou k *Waldově* statistice pro testování nulovosti koeficientu, na rozdíl od *Waldovy* statistiky však pro její výpočet není třeba explicitně počítat hodnoty parametrů. Proto je v některých situacích preferována z důvodu menší výpočetní náročnosti. Za předpokladu nulové hypotézy jsou však obě statistiky pro velké výběry ekvivalentní.

V tomto případě tedy nebyly do modelu zahrnuty proměnné: *stupeň vzdělání (ed)*, *příjem domácnosti v tisících (income)* a *ostatní pohledávky v tisících (othdebt)*.

## Shrnutí základních statistik modelu po krocích metody *Forward LR*

Step Summary<sup>a, b</sup>

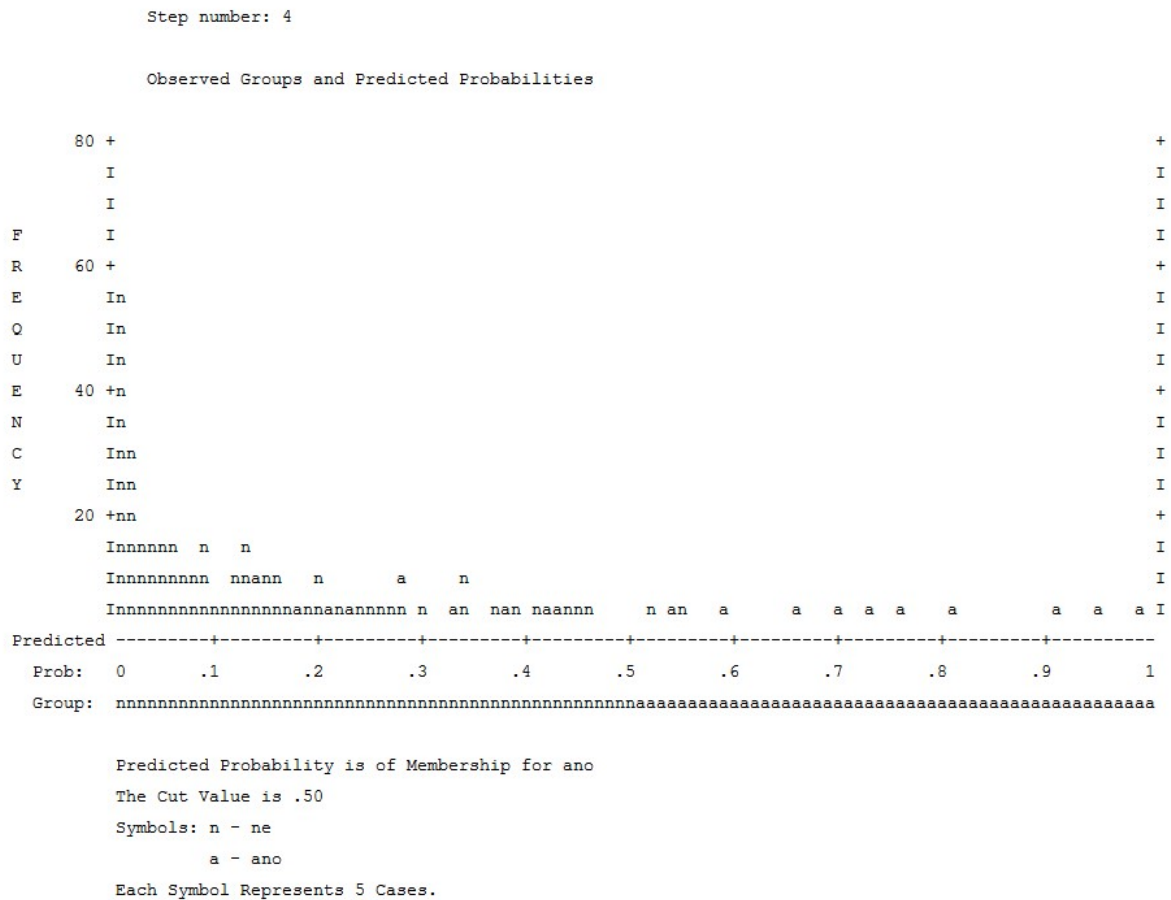
Step	Improvement			Model			Correct Class %	Variable
	Chi-square	df	Sig.	Chi-square	df	Sig.		
1	76,103	1	,000	76,103	1	,000	78,4%	IN: debtinc
2	46,464	1	,000	122,567	2	,000	79,7%	IN: employ
3	48,167	1	,000	170,734	3	,000	80,9%	IN: creddebt
4	13,299	1	,000	184,033	4	,000	82,5%	IN: address

a. No more variables can be deleted from or added to the current model.

b. End block: 1

Tabulka *Step Summary* shrnuje celkové informace o modelu po krocích metody automatického výběru prediktorů *Forward LR*. Řádky odpovídají jednotlivým krokům, sloupce popisují nejprve zlepšení vzhledem k předcházejícímu kroku (*Improvement*) a dále celkové vlastnosti modelu (*Model*). V obou částech jsou uvedeny hodnoty statistiky chí-kvadrát (*Chi-square*), její stupně volnosti (*df*) a dosažená hladina významnosti (*Sig.*) - podrobněji viz popis tabulky *Omnibus Tests of Model Coefficients*. Tabulka dále informuje o celkovém počtu správně klasifikovaných případů na trénovací množině (*Correct Class %*) a o proměnných vstupujících/vystupujících v daném kroku do/z modelu.

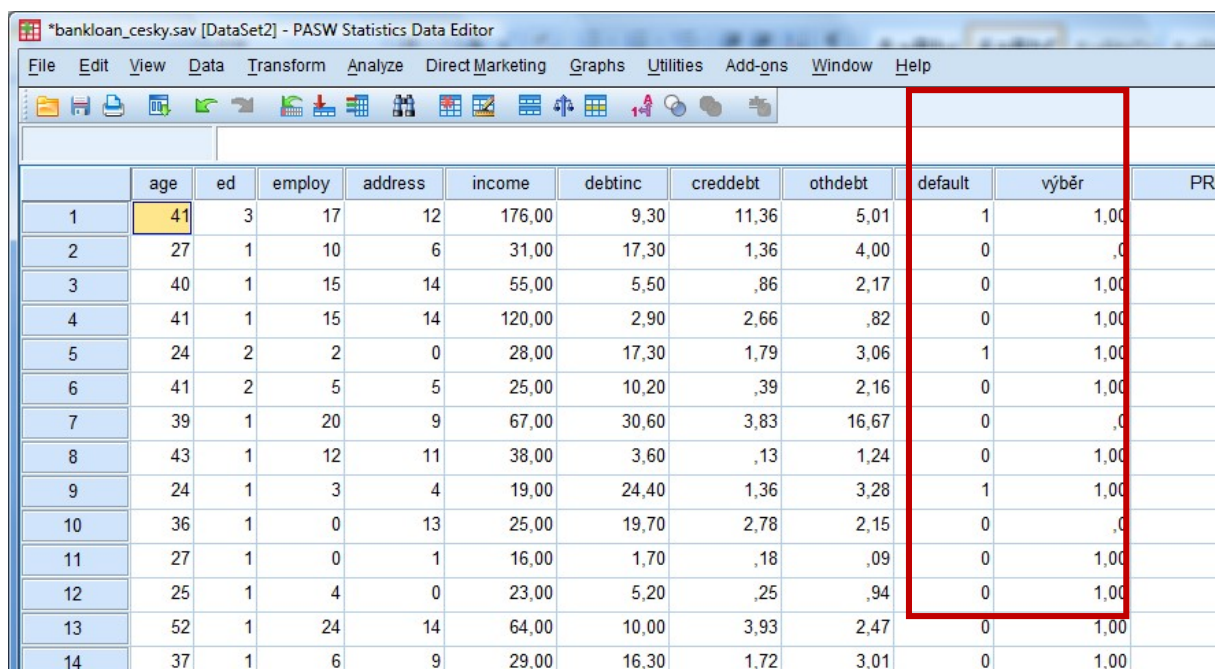
## Histogram předpovídané pravděpodobnosti podle kategorií závislé proměnné



Graf *Observed Groups and Predicted Probabilities* zobrazuje textovou podobu histogramu předpovídané pravděpodobnosti podle kategorií závislé proměnné. Poznámka pod tabulkou informuje, že písmena „n“ představují kategorii „ne“, písmena „a“ kategorii „ano“ a každé z písmen reprezentuje 5 případů.

Z obrázku vyplývá, že nízké hodnoty pravděpodobnosti odpovídají téměř výhradně kategorii „ne“, zatímco vysoké hodnoty kategori „ano“. Ve střední části grafu se však skupiny do určité míry překrývají, což způsobuje chybnou klasifikaci některých případů. Při defaultním nastavení je skupina „ano“ předpovídána pro případy, kde je odhadovaná pravděpodobnost zařazení do této skupiny větší než 0,5. Tuto hranici lze však posunout. To je důležité především v situacích, kde důsledky (náklady) vyplývající ze špatné klasifikace kategorie „ano“ jsou jiné než u kategorie „ne“. Graf rovněž naznačuje, jak by se posunutí této hranice projevilo při klasifikaci.

### Uložení nových proměnných do datové matice



\*bankloan\_cesky.sav [DataSet2] - PASW Statistics Data Editor

	age	ed	employ	address	income	debtinc	creddebt	othdebt	default	výběr	PRI
1	41	3	17	12	176,00	9,30	11,36	5,01	1	1,00	
2	27	1	10	6	31,00	17,30	1,36	4,00	0	,0	
3	40	1	15	14	55,00	5,50	,86	2,17	0	1,00	
4	41	1	15	14	120,00	2,90	2,66	,82	0	1,00	
5	24	2	2	0	28,00	17,30	1,79	3,06	1	1,00	
6	41	2	5	5	25,00	10,20	,39	2,16	0	1,00	
7	39	1	20	9	67,00	30,60	3,83	16,67	0	,0	
8	43	1	12	11	38,00	3,60	,13	1,24	0	1,00	
9	24	1	3	4	19,00	24,40	1,36	3,28	1	1,00	
10	36	1	0	13	25,00	19,70	2,78	2,15	0	,0	
11	27	1	0	1	16,00	1,70	,18	,09	0	1,00	
12	25	1	4	0	23,00	5,20	,25	,94	0	1,00	
13	52	1	24	14	64,00	10,00	3,93	2,47	0	1,00	
14	37	1	6	9	29,00	16,30	1,72	3,01	0	1,00	

Procedura rovněž umožňuje uložit do datové matice jako nové proměnné: pravděpodobnost, že u daného případu nastane sledovaná událost, předpovídanou kategorií, různé typy reziduí, statistiky vlivu a další.

V tomto případě jsme uložili předpovídanou pravděpodobnost a zařazení do skupiny (viz obrázek). Tyto odhady získáme nejen pro případy, na kterých byl model budován, ale také pro všechny ostatní případy, které mají platné hodnoty u nezávislých proměnných. Takto tedy oklasifikujeme také nové žadatele o úvěr a získáme odhad rizika, že u nich při splácení nastanou problémy.

## Příloha 1

### Srovnání logistické regrese a diskriminační analýzy

Z hlediska řešené úlohy má logistická regrese blízko k diskriminační analýze. Následujících několik bodů nabízí určité srovnání těchto metod:

- Binární logistická regrese předpokládá dichotomickou závislou proměnnou. Při více kategoriích je třeba užít její zobecnění – multinomickou logistickou regresi (v programu *IBM SPSS Statistics* je obsažena v modulu *Regression* v nabídce *Analyze, Regression, Multinomial Logistic*). Pro případ ordinální závislé proměnné je určena ordinální regrese (modul *Statistics Base*, nabídka *Analyze, Regression, Ordinal*). Diskriminační analýza umožňuje pracovat s multinomickou závislou proměnnou s více kategoriemi (uspořádání kategorií neuvažuje).
- Diskriminační analýza předpokládá spojitě nezávislé proměnné s normálním rozložením, zatímco logistická regrese pracuje s normálně rozloženými spojitými proměnnými i s kategorizovanými proměnnými.
- Diskriminační analýza má formálně více předpokladů: mnohorozměrné normální rozložení nezávislých proměnných a shodné kovarianční matice ve skupinách. Simulace *Monte Carlo* však ukazují, že splnění předpokladu mnohorozměrné normality není zásadní a metoda pracuje dobře i v situacích, kdy proměnné mají přibližně normální rozdělení a velikost souboru je alespoň 100 případů.
- V praxi bývají výsledky obou metod obvykle srovnatelné. Simulace *Monte Carlo* rovněž ukazují, že za velmi obecných podmínek není důvod upřednostňovat jednu z metod před druhou. Při splnění předpokladu mnohorozměrné normality však diskriminační analýza překonává logistickou regresi.
- Při rozhodování, kterou z metod užít, je třeba zvážit počet kategorizovaných proměnných vstupujících do modelu. Vzhledem k přísnějším předpokladům diskriminační analýzy, je při větším počtu kategorizovaných proměnných vhodnější přiklonit se spíše k logistické regresi.
- Určitou roli při volbě metody hrají rovněž zvyklosti v oboru (například v marketingových výzkumech, kde se nejčastěji pracuje s kategorizovanými daty, se obvykle upřednostňuje logistická regrese).



## Příloha 2

### Schémata kódování kategorizovaných proměnných

Nezávislé kategorizované proměnné nemohou vstupovat do logistické regrese přímo (podobně jako v případě lineární regrese) a je třeba je nejprve vhodným způsobem transformovat. K tomuto účelu se užívají umělé proměnné, kterých je vždy o jednu méně než kategorií původní proměnné, a které s těmito kategoriemi určitým způsobem korespondují.

Pro kategorizované proměnné je jedinou možností, jak vyjádřit efekt určité kategorie, její srovnání vzhledem k jiným kategoriím. Z tohoto důvodu se užívají různé typy transformací, které v konečném důsledku vedou ke stejným odhadům modelu i dalších statistik, avšak umožňují provádět odlišná srovnání efektu kategorií (tzv. kontrasty). Různé typy transformací se liší v odhadech regresních koeficientů, které mají odlišný význam i interpretaci. Vždy se přitom snažíme volit takový typ kontrastu, který zachycuje zajímavé vztahy z hlediska řešené úlohy.

Každý typ kontrastu charakterizují dvě základní matice: designová matice  $X$  (někdy též matice báze) a matice kontrastů  $C$ , mezi kterými je následující vztah:

$$C = (X'X)^{-1}X'$$

Sloupce **designové matice** (mimo první sloupec) vyjadřují, jakým způsobem je z hodnot původní kategorizované proměnné odvozeno  $n-1$  transformovaných proměnných, které vstupují do modelu. První sloupec se vztahuje ke konstantě. Koeficienty této matice (mimo první sloupec) se zobrazují ve výstupu procedury v tabulce s názvem *Categorical Variables Codings*.

Například pro indikační proměnné kódující kategorizovanou proměnnou se čtyřmi kategoriemi, kde poslední kategorie je referenční má designová matice tvar:

1	1	0	0
1	0	1	0
1	0	0	1
1	0	0	0

Z druhého sloupce odvodíme, že první transformační proměnná je rovna jedné pro první kategorii a nule pro ostatní kategorie. Obdobně druhá resp. třetí proměnná jsou rovny jedné pro druhou resp. třetí kategorii a nule pro ostatní kategorie.

Každý řádek designové matice se tedy vztahuje k jedné z kategorií původní proměnné. Jednička v prvním sloupci vyjadřuje, že model obsahuje konstantu, ostatní koeficienty potom odpovídají hodnotám transformovaných proměnných pro tuto kategorii.

Jestliže předpokládáme model, který neobsahuje žádné další prediktory, má regresní rovnice tvar:

## Listy procedur IBM SPSS Statistics

$$\text{logit}(Y) = b_0 + b_1.\text{Ind}_1 + b_2.\text{Ind}_2 + b_3.\text{Ind}_3.$$

Pro případy z první kategorie tedy dostáváme:

$$\text{logit}(Y)_{(\text{kategorie}=1)} = b_0.1 + b_1.1 + b_2.0 + b_3.0 = b_0 + b_1,$$

$$\text{pro druhou kategorii: } \text{logit}(Y)_{(\text{kategorie}=2)} = b_0 + b_2,$$

$$\text{pro třetí kategorii: } \text{logit}(Y)_{(\text{kategorie}=3)} = b_0 + b_3,$$

$$\text{pro čtvrtou (referenční) kategorií: } \text{logit}(Y)_{(\text{kategorie}=4)} = b_0.$$

**Matice kontrastů** umožňuje porozumět významu jednotlivých koeficientů regresního modelu ( $b_0, b_1, \dots$ ). Každý řádek odpovídá jednomu regresnímu koeficientu. Regresní koeficienty vyjadřují určitý vztah mezi logity původních kategorií (jejich lineární kombinací). Vždy se přitom snažíme volit takový typ kontrastu, který zachycuje zajímavé vztahy z hlediska interpretace modelu. Prvky matice kontrastů potom určují koeficienty této lineární kombinace.

Například pro indikační proměnné kódující kategorizovanou proměnnou se čtyřmi kategoriemi, kde poslední kategorie je referenční, má matice kontrastů tvar:

0	0	0	1
1	0	0	-1
0	1	0	-1
0	0	1	-1

Pro regresní koeficienty tedy můžeme odvodit následující vztahy:

$$b_0 = 0.\text{logit}(Y)_{(\text{kategorie}=1)} + 0.\text{logit}(Y)_{(\text{kategorie}=2)} + 0.\text{logit}(Y)_{(\text{kategorie}=3)} + 1.\text{logit}(Y)_{(\text{kategorie}=4)} = \text{logit}(Y)_{(\text{kategorie}=4)},$$

$$b_1 = \text{logit}(Y)_{(\text{kategorie}=1)} - \text{logit}(Y)_{(\text{kategorie}=4)},$$

$$b_2 = \text{logit}(Y)_{(\text{kategorie}=2)} - \text{logit}(Y)_{(\text{kategorie}=4)},$$

$$b_3 = \text{logit}(Y)_{(\text{kategorie}=3)} - \text{logit}(Y)_{(\text{kategorie}=4)}.$$

Užití indikačních proměnných má tedy význam v situaci, kdy se zajímáme o porovnání efektu jednotlivých kategorií vzhledem k určité referenční kategorii. Konstanta vyjadřuje odhad logitu referenční kategorie.

## Příklady designových matic a matic kontrastů

Níže jsou uvedeny designové matice a matice kontrastů pro jednotlivé typy kontrastů pro případ kategorizované proměnné se čtyřmi kategoriemi. Obecné vyjádření pro  $n$  kategorií je k dispozici v manuálu příslušného modulu *IBM SPSS Statistics*.

### 1) Indikační proměnné (Indicator)

Každá kategorie mimo jednu (tzv. referenční kategorie) vytváří 0-1 proměnnou. Jako referenční se obvykle volí první nebo poslední kategorie, v ostatních případech je třeba zadat příkaz pomocí syntaxe. Konstanta vyjadřuje odhad pro referenční kategorii, ostatní koeficienty efekt příslušné kategorie vzhledem k referenční kategorii.

Designová matice pro případ 4 kategorií (poslední kategorie je referenční):

1	1	0	0
1	0	1	0
1	0	0	1
1	0	0	0

Matice kontrastů pro případ 4 kategorií (poslední kategorie je referenční):

0	0	0	1
1	0	0	-1
0	1	0	-1
0	0	1	-1

### 2) Jednoduché kontrasty (Simple)

Každá kategorie je porovnávána vzhledem k referenční kategorii. Jako referenční se obvykle volí první nebo poslední kategorie, v ostatních případech je třeba zadat příkaz pomocí syntaxe. Konstanta vyjadřuje průměr všech kategorií, ostatní koeficienty efekt příslušné kategorie vzhledem k referenční. (Ve srovnání s indikačními proměnnými se liší pouze odhad a význam konstanty, další koeficienty nabývají stejných hodnot).

Designová matice pro případ 4 kategorií (poslední kategorie je referenční):

1	3/4	-1/4	-1/4
1	-1/4	3/4	-1/4
1	-1/4	-1/4	3/4
1	-1/4	-1/4	-1/4

## Listy procedur IBM SPSS Statistics

Matice kontrastů pro případ 4 kategorií (poslední kategorie je referenční) :

1/4	1/4	1/4	1/4
1	0	0	-1
0	1	0	-1
0	0	1	-1

### 3) Diferenční kontrasty (Difference)

Konstanta vyjadřuje průměr všech kategorií, ostatní koeficienty porovnávají efekt příslušné kategorie (kromě první) s průměrným efektem předcházejících kategorií.

Designová matice pro případ 4 kategorií:

1	-1/2	-1/3	-1/4
1	1/2	-1/3	-1/4
1	0	2/3	-1/4
1	0	0	3/4

Matice kontrastů pro případ 4 kategorií:

1/4	1/4	1/4	1/4
-1	1	0	0
-1/2	-1/2	1	0
-1/3	-1/3	-1/3	1

### 4) Helmertovy kontrasty (Helmert)

Konstanta vyjadřuje průměr všech kategorií, ostatní koeficienty porovnávají efekt příslušné kategorie (kromě poslední) s průměrným efektem následujících kategorií.

Designová matice pro případ 4 kategorií:

1	3/4	0	0
1	-1/4	2/3	0
1	-1/4	-1/3	1/2
1	-1/4	-1/3	-1/2

## Listy procedur IBM SPSS Statistics

Matice kontrastů pro případ 4 kategorií:

1/4	1/4	1/4	1/4
1	-1/3	-1/3	-1/3
0	1	-1/2	-1/2
0	0	1	-1

### 5) Porovnání sousedních kategorií (Repeated)

Konstanta vyjadřuje průměr všech kategorií, ostatní koeficienty porovnávají efekt příslušné kategorie (kromě poslední) vzhledem k následující kategorii.

Designová matice pro případ 4 kategorií:

1	3/4	1/2	1/4
1	-1/4	1/2	1/4
1	-1/4	-1/2	1/4
1	-1/4	-1/2	-3/4

Matice kontrastů pro případ 4 kategorií:

1/4	1/4	1/4	1/4
1	-1	0	0
0	1	-1	0
0	0	1	-1

### 6) Odchylkové kontrasty (Deviation)

Konstanta vyjadřuje průměr všech kategorií, ostatní koeficienty porovnávají efekt příslušné kategorie (kromě referenční) vzhledem k průměrnému efektu všech kategorií. Jako referenční se obvykle volí první nebo poslední kategorie, v ostatních případech je třeba zadat příkaz pomocí syntaxe.

Designová matice pro případ 4 kategorií (poslední kategorie je referenční):

1	1	0	0
1	0	1	0
1	0	0	1
1	-1	-1	-1

## Listy procedur IBM SPSS Statistics

Matice kontrastů pro případ 4 kategorií (poslední kategorie je referenční):

1/4	1/4	1/4	1/4
3/4	-1/4	-1/4	-1/4
-1/4	3/4	-1/4	-1/4
-1/4	-1/4	3/4	-1/4

### 7) Polynomické kontrasty (Polynomial)

Konstanta vyjadřuje průměr všech kategorií, ostatní koeficienty postupně lineární efekt všech kategorií, kvadratický efekt, kubický efekt atd. Předpokládá se, že kategorie jsou uspořádané a od sebe stejně vzdálené (případně lze specifikovat vzdálenosti kategorií pomocí syntaxe).

Designová matice pro případ 4 stejně vzdálených kategorií:

1	$-3/\sqrt{20}$	1/2	$-1/\sqrt{20}$
1	$-1/\sqrt{20}$	-1/2	$3/\sqrt{20}$
1	$1/\sqrt{20}$	-1/2	$-3/\sqrt{20}$
1	$3/\sqrt{20}$	1/2	$1/\sqrt{20}$

Matice kontrastů pro případ 4 stejně vzdálených kategorií:

1/4	1/4	1/4	1/4
$-3/\sqrt{20}$	$-1/\sqrt{20}$	$1/\sqrt{20}$	$3/\sqrt{20}$
1/2	-1/2	-1/2	1/2
$-1/\sqrt{20}$	$3/\sqrt{20}$	$-3/\sqrt{20}$	$1/\sqrt{20}$