Chosen Procedures within Multiple Logistic Regression Analysis[1]

Vybrané postupy při aplikaci vícenásobné logistické regrese

Petr Gurný [2]

Abstract

The paper is devoted to the practical issues within credit-scoring models' estimation (based on the multiple logistic regression analysis), particularly in case of the utilization of financial ratios as the independent variables. There are shown the differences in the resulting models and their predictive abilities, taking into account the economic interpretation of the particular ratios on their range of values and the different possibilities of sampling to estimate model, in the illustrative example in the paper.

Key words

Multiple Logistic regression; Financial ratios; ROC curve; BIF analysis.

JEL Classification: C01; C58; C81.

1. Úvod

Využití regresní analýzy k definování vztahu mezi závislou proměnnou a jednou nebo více vysvětlujícími proměnnými je jedním ze základních nástrojů u většiny statistických analýz. Na poli financí se tohoto instrumentu využívá hlavně při vyšetřování finančního zdraví jednotlivých subjektů, které je pak často převáděno na pravděpodobnost úpadku (PD) a vznikají tzv. credit-scoring modely. Tyto modely jsou pak v široké míře využívány při určování bonity dlužníků a také ve spoustě dalších oblastí risk managementu firem, ve kterých hraje klíčovou roli právě zmíněný parametr PD. Credit-scoring modely se obecně člení do několika kategorií (blíže viz Green (2008) nebo Engelmann and Rauhmeier (2006)). Statistická metodologie pro tyto procesy byla představena již autory jako jsou Fischer (1936) a Durand (1941), o rozvoj a aplikaci těchto modelů se dále zasloužil zejména Beaver (1966) a Altman (1968). V rámci České republiky patří mezi nejznámější model tzv. IN model, viz Neumaierová a Neumaier (2002).

V rámci credit-scoring modelů je pak obvykle závislá proměnná definována jako binární veličina, což následně vede ke skupině modelů známých jako zobecněné lineární modely. Z této slupiny jsou pak nejčastěji pro účely odhadu kreditních modelů využívány logit modely, probit modely a diskriminační analýzy. Charakteristickým rysem je dále využívání poměrových ukazatelů finanční analýzy jako vysvětlujících veličin. Zde dochází často k chybám při nerozlišování toho, ve kterém místě definičního oboru se daný ukazatel pohybuje, z čehož mohou plynout značná zkreslení při odhadu výsledného modelu a jeho predikční schopnosti.

[1] Tento článek vznikl za finanční podpory Studentské grantové soutěže EkF, VŠB-TU Ostrava v rámci projektu SP2012/19.
[2] VŠB – Technická univerzita Ostrava, Ekonomická fakulta, katedra financí, Sokolská 33, 701 21 Ostrava 1, e-mail: petr.gurny@vsb.cz.

Cílem příspěvku je poukázat na praktické problémy a úskalí při odhadu scoringových modelů na bázi logistické regrese, a to při použití různých typů poměrových ukazatelů finanční analýzy jako vysvětlujících veličin.

Příspěvek bude strukturován následovně. Nejprve bude v metodologické části stručně představena vícenásobná logistická regrese, včetně nastínění způsobu odhadu parametrů a možnosti hodnocení predikční schopnosti. V aplikační části pak budou popsána ilustrativní empirická data (bez jejich ekonomické interpretace, pouze se zaměřením na typ) a následně budou odhadnuty a porovnány tři postupně se zpřesňující modely logistické regrese, s poukázáním na hlavní chyby, které jsou při odhadu obvyklé. Závěrem bude provedena komparace všech tří modelů.

2. Vícenásobná logistická regrese

Jak již bylo uvedeno výše, vícenásobná logistická regrese je vícerozměrný statistický model sloužící k předpovědi pravděpodobnosti defaultu, přičemž jako vstupy se využívají hlavní ekonomické a finanční ukazatele. Tento model zachycuje vztah mezi závislou proměnnou *Y* (dichotomická proměnná) a jednou nebo více nezávislými proměnnými *X*.

Vysvětlovaná proměnná, y_i, je dána

$$y_i = \begin{cases} 1 & \text{jestliže default nastane} \\ 0 & \text{jestliže default nenastane} \end{cases},$$

a dále předpokládejme, že pravděpodobnost $y_i = 1$ je dána P_i a tedy že $y_i = 0$ je dána pravděpodobností $1 - P_i$:

$$y_i = \begin{cases} 1 & \text{s pravděpodobností } P_i \\ 0 & \text{s pravděpodobností } 1 - P_i \end{cases}.$$

Cílem je tedy modelovat pravděpodobnost P_i, že default nastane, specifikováním následujícího modelu

$$P_i = f(z) = f(\alpha + \beta x_i),$$

kde x_i jsou jednotlivé finanční indikátory a α a β jsou odhadované parametry.

Existuje řada možností, jak specifikovat P_i, v tomto článku se ale zaměříme na logistickou transformaci, tedy na logit model:

$$P_i = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} = \frac{1}{1 + \exp(-\alpha - \beta x_i)} = \frac{1}{1 + exp(-z)}.$$

Vzhledem k nelineárním vlastnostem tohoto modelu není možné při odhadu parametrů využít klasickou OLS (ordinary least squares) metodu, nýbrž je nutné maximalizovat funkci věrohodnosti. Při dané pravděpodobnosti P_i, můžeme formulovat pravděpodobnostní funkci jako

$$L = \prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i}.$$

Z čistě výpočetních důvodů je vhodnější maximalizovat logaritmus této funkce, tedy

$$\ln L = \sum_{i=1}^{n} y_i \ln P_i + \sum_{i=1}^{n} (1 - y_i) \ln(1 - P_i).$$

Konkrétně tedy pro logit model:

$$\ln L = \sum_{i=1}^{n} y_i \ln\left(\frac{1}{1 + \exp(-\alpha - \beta x_i)}\right) + \sum_{i=1}^{n} (1 - y_i) \ln\left(1 - \frac{1}{1 + \exp(-\alpha - \beta x_i)}\right).$$

Maximalizací těchto funkcí získáme odhady parametrů α a β.

Pro výběr jednotlivých proměnných se pak používá stepwise metoda a pro testování významnosti modelu Hosmer-Lemeshow test (Goodness of Fit Test), log-likelihood ratio test nebo Wald test (viz jakákoliv ekonometrická učebnice, např.: Hair, Anderson, Tatham and Black (2005)). Velmi silným nástrojem pro měření kvality predikce odhadnutého modelu je pak tzv. ROC analýza (viz např.: Hanley and McNeil (1982)).

3. Ilustrativní příklad

V této části bude výše uvedený metodologický přístup aplikován na ilustrativní vzorek 700 subjektů (firem) a postupně budou odhadnuty a porovnány tři modely s různými výchozími předpoklady.

3.1 Vstupní data

Jako vstupní vzorek jsou použity části databáze skutečných dat, které byly spojeny čistě pro ilustrativní účely a nemají za cíl být podkladem pro odhad reálně použitého modelu. Z tohoto důvodu nebudou jednotlivé nezávislé proměnné přesně specifikovány, pouze rozděleny do tří typů, viz níže. Vzorek obsahuje 700 subjektů, z nichž 50 % bylo klasifikováno jako defaultní a 50 % jako zdravé společnosti. Důležitým předpokladem je, že data jsou ekonomicky upravena, tedy vylučujeme např. možnost záporných veličin jako je vlastní kapitál, celková aktiva, oběžná aktiva apod., které ovšem reálně v účetních výkazech nastat mohou. Soubor rovněž neobsahuje chybějící hodnoty, nebude tedy nutná jejich analýza, zejména tedy vyšetření, zda chybějící hodnoty mají či nemají vliv na závislou proměnnou. Tabulka 1 zobrazuje základní popisnou statistiku jednotlivých ukazatelů, kde jsou kromě obvyklých parametrů zachyceny také jednotlivé kvartily, ze kterých můžeme následně pozorovat i výskyt odlehlých hodnot.[3]

[3] Tím jsou myšleny hodnoty překračující trojnásobek kvartilového rozpětí.

Tabulka 1: Popisná statistika

		default	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
N	Valid	700,00	700,00	700,00	700,00	700,00	700,00	700,00	700,00	700,00	700,00	700,00
	Missing	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Mean		0,50	2,96	5,79	0,07	3,44	0,19	6,87	2,41	0,43	14,48	12,60
Std. Deviation		0,50	9,47	23,88	0,28	23,24	1,68	15,10	7,64	0,83	24,63	17,81
Variance		0,25	89,66	570,44	0,08	539,88	2,82	228,10	58,44	0,68	606,53	317,04
Minimum		0,00	-60,15	-123,45	-2,00	-360,27	-29,36	0,00	0,00	0,00	0,00	0,00
Maximum		1,00	64,80	207,29	2,69	147,55	19,00	173,50	108,26	19,16	290,00	205,44
Percentiles	25	0,00	0,26	1,83	0,01	0,74	0,01	1,37	0,14	0,15	1,44	4,17
	50	0,50	1,50	4,80	0,04	2,46	0,10	2,38	0,43	0,38	3,92	8,01
	75	1,00	3,58	10,68	0,10	6,00	0,32	6,06	1,36	0,61	16,52	15,05

Vysvětlující proměnné *x1-x10* jsou poměrové finanční ukazatele, které jsou rozděleny do následujících tří typů:

I. typ: proměnné *x1-x2*, kde čitatel může nabývat jakýchkoliv kladných čísel, jmenovatel pak může být jakékoliv reálné číslo s výjimkou nuly. Příkladem ukazatelů prvního typu může být ukazatel úrokového zatížení nebo podíl bankovních úvěrů na nějaké formě zisku apod.

$$I.typ = \frac{\langle 0;\infty)}{(-\infty;\infty)\setminus\{0\}}.$$

II. typ: proměnné *x3-x5*, kde čitatel může nabývat všech reálných hodnot, jmenovatel pak pouze kladných nenulových hodnot. Příkladem ukazatelů druhého typu může být většina typů rentability, úrokové krytí apod.

$$II.typ = \frac{(-\infty;\infty)}{(0;\infty)}.$$

III. typ: proměnné *x6-x10*, kde čitatel i jmenovatel mohou nabývat pouze kladných hodnot. Příkladem ukazatelů třetího typu jsou pak ukazatele likvidity, aktivity, většina ukazatelů zadluženosti apod.

$$III.typ = \frac{\langle 0;\infty)}{(0;\infty)}.$$

3.2 I. model

V rámci odhadu I. modelu budeme postupovat následovně. Nejprve náhodně rozdělíme celkový vzorek na dvě části v poměru 70:30, z nichž prvních bude sloužit pro odhad modelu, druhá pak pro jeho verifikaci. Pro výběr ukazatelů bude použita metoda stepwise forward. Tabulka 2 ukazuje výsledné hodnoty koeficientů těch ukazatelů, které vyšly jako statisticky významné na 5% hladině významnosti, tabulka 3 potom hodnotu Hosmer - Lemeshow testu, který by měl pro dobré fitování dat vyjít na 5% hladině nevýznamný. Je zřejmé, že výsledný model fituje data adekvátně.

Tabulka 2: Proměnné v rovnici (I. model)

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 4	x1	-,054	,015	12,820	1	,000	,948
	x4	,098	,016	38,726	1	,000	1,103
	x5	,841	,227	13,698	1	,000	2,320
	x8	,259	,106	5,965	1	,015	1,295
	Constant	-,459	,136	11,331	1	,001	,632

Tabulka 3: Hosmer-Lemeshow test (I. model)

Step	Chi-square	df	Sig.
4	10,981	8	,117

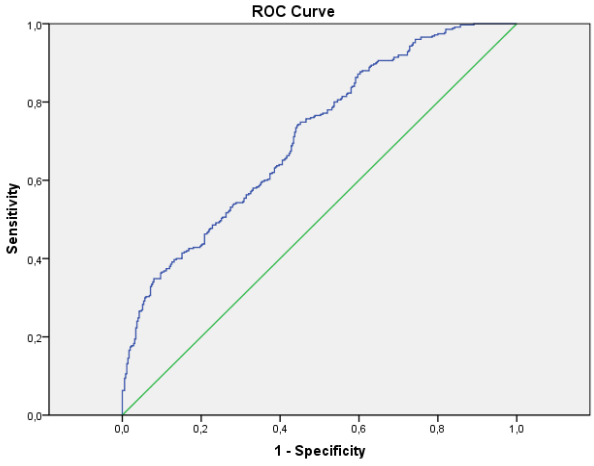
V tabulce 4 pak můžeme pozorovat klasifikační schopnost modelu jak na původním, tak na kontrolním vzorku. Ta se v obou případech pohybuje zhruba na úrovni 64 %.

Tabulka 4: Klasifikační tabulka (I. model)

Observed			Predicted					
			Selected Cases a			Unselected Cases b		
			default		Percentage Correct	default		Percentage Correct
			0	1		0	1	
Step 4	default	0	163	75	66,5	81	31	70,3
		1	93	148	61,4	43	66	60,6
	Overall Percentage				***63,9***			***64,5***

Kvalitnějším nástrojem pro hodnocení efektivity modelu je pak ROC křivka, jelikož na rozdíl od klasifikační tabulky pracuje s různými hodnotami klasifikačního řezu. Tvar ROC křivky pro I. model a zejména hodnotu velikosti plochy pod ROC křivkou lze vidět z obrázku 1 a tabulky 5. Plocha pod ROC křivkou je 0,691, což značí relativně dobrou predikční schopnost.

Obrázek 1: ROC křivka (I. model)



Tabulka 5: Plocha pod ROC křivkou (I. model)

Area	Std. Errora	Asymptotic Sig.b	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,691	,019	,000	,673	,748

3.3 II. model

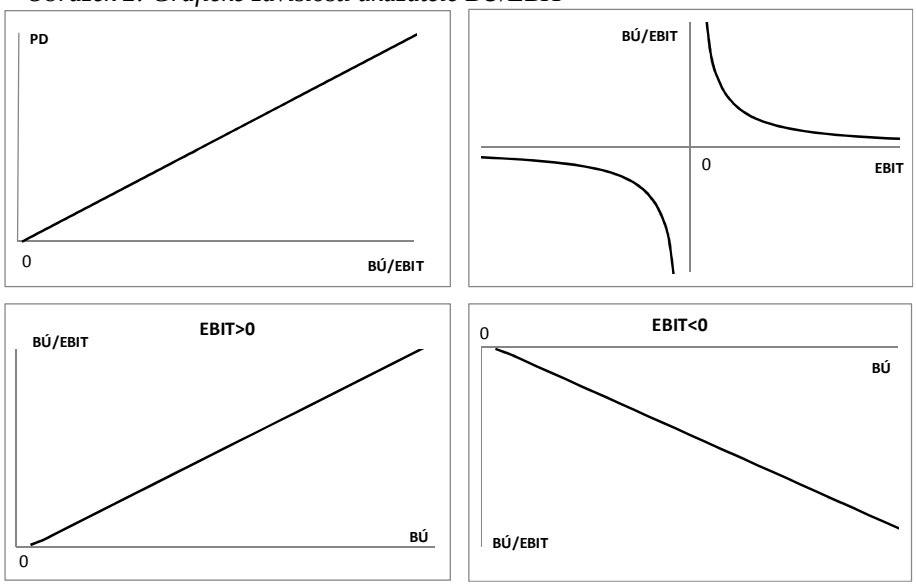
V rámci II. modelu opět rozdělíme celkový vzorek na vzorek pro odhad modelu a kontrolní vzorek ve stejném poměru jako v I. modelu, navíc ještě ale provedeme analýzu a úpravu nezávislých proměnných a to v následujících bodech:

a) vyšetření chování jednotlivých typů proměnných na definičním oboru dílčích veličin,
b) zjištění ekonomické a logické interpretovatelnosti,
c) transformace veličin (logaritmická).

3.3.1 Úprava I. typu proměnných

Jako ukázku I. typu proměnných budeme uvažovat s ukazatelem podílu bankovních úvěrů (BÚ) na EBITu (dále jen ukazatel). Obrázek 2 zachycuje závislost PD na zmíněném ukazateli a rovněž závislost ukazatele na jeho dílčích veličinách (v tom případě tedy na EBITu a bankovních úvěrech).

Obrázek 2: Grafické závislosti ukazatele BÚ/EBIT



V levé horní části obrázku 2 je zachycena závislost PD na ukazateli, tedy vidíme, že s růstem ukazatele PD roste, což je v rámci jednorozměrné analýzy ekonomicky zřejmě správně. Graf v pravé horní části zachycuje závislost ukazatele na velikosti EBITu, což z důvodu nezávislé veličiny ve jmenovateli vede k funkci ve tvaru hyperboly. Pokud bude firma v zisku (kladná část osy x) s růstem zisku bude ukazatel klesat, což ve spojení s prvním grafem značí nižší PD, což je opět logicky správně. Pokud bude firma ve ztrátě, povede růst EBITu (zmenšování ztráty) opět k poklesu ukazatele a tedy správné interpretaci, nicméně je zřejmé, že není možné porovnávat firmy s kladným a záporným EBITem mezi sebou, jelikož firma s nekonečně velkým kladným EBITem by se velikostí analyzovaného ukazatele blížila firmě s nekonečnou ztrátou. Naopak firmy s nekonečně malým ziskem a nekonečně malou

ztrátou by měly velikost ukazatele matematicky nekonečně odlišnou. Jako řešení se jeví rozdělit ukazatel I. typu na kladné a záporné hodnoty a uvažovat s nimi jako se dvěma odlišnými proměnnými. Je ovšem třeba se ještě podívat na závislost ukazatele na BÚ. Pro kladný EBIT (levý spodní graf) vede růst BÚ k růstu ukazatele, což opět vede k vyšší PD, tedy správné ekonomické interpretaci. Ovšem pro zápornou větev EBITu vede nárůst BÚ k poklesu ukazatele, což by znamenalo pokles PD, což je ekonomicky zřejmý nesmysl. Důsledky lze názorně vidět v tabulce 6.

Tabulka 6: Důsledky změny BÚ na ukazatel BÚ/EBIT při ztrátě

	BÚ	EBIT	BÚ/EBIT	BÚ	EBIT	BÚ/EBIT
I. varianta	100	-100	***-1***	200	-100	***-2***
II. varianta	100	-100	***-1***	100	-50	***-2***

V I. variantě jsou dvě firmy se stejnou ztrátou, přičemž druhá má větší BÚ, tedy je na tom hůře. Ve II. variantě mají firmy stejnou velikost BÚ, ale první má větší ztrátu, tedy je na tom za jinak nezměněných okolností hůře. V I. variantě je tedy firma s velikostí ukazatele -1 na tom ekonomicky lépe než firma s velikostí ukazatele -2, ve II. variantě je tomu naopak. Pro záporné hodnoty ukazatele tedy není možné rozhodnout a porovnat firmy mezi sebou navzájem. Řešením je zavést umělou (dummy) proměnnou, D, která by rozlišovala mezi kladnou a zápornou hodnotou ukazatele a hodnoty ukazatele by pak byly brány pouze pro kladná čísla.

Dalším možným a vhodným postupem je transformace značně variovaných dat (viz Tabulka 1), což může být jednoduše provedeno pomocí logaritmické transformace:

$$x^T = ln(x+1), x \geq 0,$$

kde x je hodnota ukazatele I. typu, jedničku přičítáme proto, aby bylo možno provést logaritmickou transformaci i v případě nulové hodnoty, tedy nulového BÚ. Výsledné z vstupující do logistické transformace pak pro ukazatel I. typu vypadá následovně:

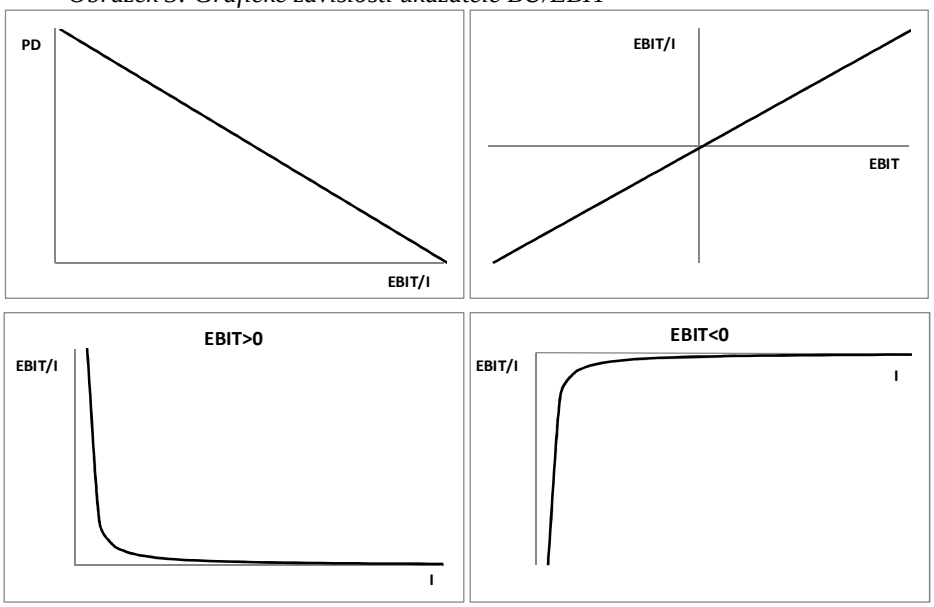
$$z = \beta_1 \cdot D + \beta_2 \cdot ln(x+1) \cdot (1-D), \qquad D = \begin{cases} 0, x \geq 0, \\ 1, x < 1. \end{cases}$$

3.3.2 Úprava II. typu proměnných

Jako ukázku II. typu proměnných budeme uvažovat s ukazatelem úrokového krytí (dále jen ukazatel). Obrázek 3 zachycuje závislost PD na zmíněném ukazateli a rovněž závislost ukazatele na jeho dílčích veličinách (v tom případě tedy na EBITu a nákladových úrocích (*I*)).

Interpretace je obdobná jako u předcházející podsekce. Ekonomicky neinterpretovatelný je pravý spodní graf, kdy při růstu *I* roste i ukazatel, což by znamenalo, že v případě ztráty vede vyšší hodnota nákladových úroků k nižší PD. Podobně jako v předcházejícím případě lze důsledky demonstrovat na jednoduchém příkladě, viz tabulka 7, kdy je na tom v I. variantě ekonomicky lépe firma s hodnotou ukazatele -5, v II. variantě firma s hodnotou ukazatele -2.

Obrázek 3: Grafické závislosti ukazatele BÚ/EBIT



Tabulka 7: Důsledky změny I na ukazatel BÚ/EBIT při ztrátě

	EBIT	I	EBIT/I	EBIT	I	EBIT/I
I. varianta	-100	20	***-5***	-100	50	***-2***
II. varianta	-100	20	***-5***	-40	20	***-2***

Řešením je stejně jako v případě I. typu proměnných zavedení dummy proměnné pro kladnou a zápornou část ukazatele. Společně s logaritmickou transformací pak dojdeme ke stejnému formálnímu zápisu jako v předcházejícím případě.

$$x^T = ln(x+1), x \geq 0,$$

kde x je hodnota ukazatele II. typu. Výsledné z vstupující do logistické transformace pak pro ukazatel II. typu vypadá následovně:

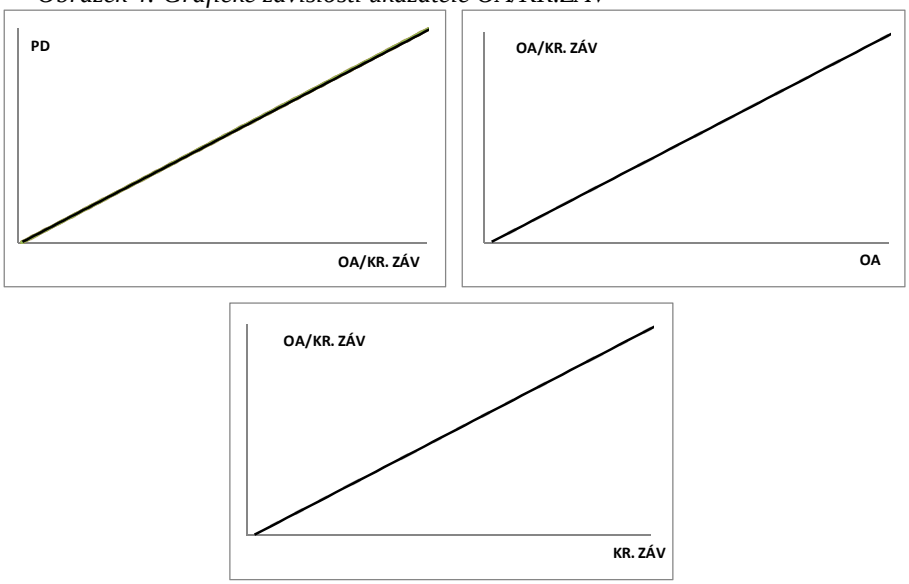
$$z = \beta_1 \cdot D + \beta_2 \cdot ln(x+1) \cdot (1-D), \qquad D = \begin{cases} 0, x \geq 0, \\ 1, x < 1. \end{cases}$$

3.3.3 Úprava III. typu proměnných

Jako ukázku III. typu proměnných budeme uvažovat s ukazatelem celkové likvidity (dále jen ukazatel). Obrázek 4 zachycuje závislost PD na zmíněném ukazateli a rovněž závislost ukazatele na jeho dílčích veličinách (v tom případě tedy na oběžných aktivech (OA) a krátkodobých závazcích (*KR.ZÁV*)).

Zde odpadá problém se zápornými hodnotami ukazatele, ukazatel tedy má vždy správnou a porovnatelnou ekonomickou hodnotu a interpretaci. Grafické znázornění závislostí je patrné z obrázku 4.

Obrázek 4: Grafické závislosti ukazatele OA/KR.ZÁV



Výsledné *z* vstupující do logistické transformace pak po logaritmické transformaci ukazatele vypadá pro ukazatel II. typu následovně:

$$z = \beta_1 \cdot ln(x+1).$$

3.3.4 II. model - výsledky

Po zavedení dummy proměnných a transformaci ukazatelů jsou opět odhadnuty koeficienty logistické regrese a provedena verifikace na kontrolním vzorku. Výsledky shrnují tabulky 8 - 11 a obrázek 5. Interpretace je stejná jako u I. modelu.

Tabulka 8: Proměnné v rovnici (II. model)

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 4	lnx2	-,742	,193	14,773	1	,000	,476
	lnx5	1,073	,524	4,184	1	,041	2,923
	lnx6	,399	,165	5,881	1	,015	1,491
	lnx7	-,677	,209	10,518	1	,021	,508
	Constant	1,247	,518	5,790	1	,016	3,481

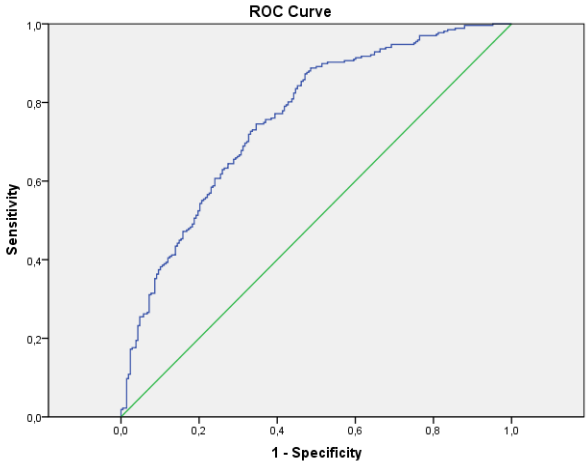
Tabulka 9: Hosmer-Lemeshow test (II. model)

Step	Chi-square	df	Sig.
4	5,686	8	,682

Tabulka 10: Klasifikační tabulka (II. model)

Observed			Predicted					
			Selected Casesa			Unselected Cases		
			default		Percentage Correct	default		Percentage Correct
			0	1		0	1	
Step 4	default	0	75	50	72,0	32	28	73,0
		1	35	124	78,0	8	55	81,3
	Overall Percentage				***73,1***			***69,1***

Obrázek 5: ROC křivka (II. model)



Tabulka 11: Plocha pod ROC křivkou (II. model)

Area	Std. Errora	Asymptotic Sig.b	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,759	,022	,000	,716	,803

V II. modelu zůstalo významných pět transformovaných ukazatelů, z toho žádná dummy proměnná (to znamená, že záporné hodnoty ukazatelů nemají významný vliv na PD, což může být způsobeno právě jejich nejednoznačnou interpretací). Hosmer-Lemeshow test je v pořádku a klasifikační schopnost se u kontrolního vzorku zvedla, stejně tak jako plocha pod ROC křivkou.

3.4 III. model

Při odhadu II. modelu byl po úpravě a transformaci proměnných náhodně vybrán vzorek pro odhad modelu, zbytek pak byl použit jako kontrolní vzorek pro jeho verifikaci. Pokud však provedeme náhodný výběr znovu (při zachování poměru 70:30), vyjdou odlišně jak vybrané významné ukazatele, tak jejich koeficienty, viz nový náhodný výběr s výsledky v tabulce 12 ve srovnání s původními ukazateli a koeficienty v tabulce 8.

Tabulka 12: Proměnné v rovnici (II. model), jiný výběr

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 3	lnx2	-,855	,168	25,863	1	,000	,425
	lnx6	,400	,139	8,321	1	,004	1,492
	lnx7	-,629	,182	11,978	1	,001	,533
	Constant	1,639	,422	15,068	1	,000	5,150

Z tabulky 8 je zřejmé, že při daném výběru se oproti výsledkům v tabulce 8 ukázal jako statisticky nevýznamný transformovaný ukazatel *x5* (konkrétně jeho kladná část), koeficienty u zbylých ukazatelů se pak drobně změnily. Z tohoto důvodu byla následně provedena BIF analýza (Bootstrap Inclusion Fraction), kdy byla provedena stepwise metoda pro odhad statisticky významných parametrů při 100 náhodných výběrech. Tabulka 13 pak shrnuje procentuální výskyt jednotlivých významných ukazatelů ve všech 100 pokusech.

Tabulka 13: BIF analýza

ln(x1)	22%	**ln(x6)**	**91%**	D1(-)	1%
ln(x2)	**85%**	**ln(x7)**	**78%**	D2(-)	3%
ln(x3)	13%	**ln(x8)**	**69%**	D3(-)	0%
ln(x4)	31%	ln(x9)	32%	D4(-)	0%
ln(x5)	6%	ln(x10)	8%	D5(-)	2%

Z výsledků je patrné, že stabilními ukazateli, které se jevily jako statisticky významné v minimálně 70 % případů, jsou transformované ukazatele *x2*, *x6*, *x7* a *x8*. Tyto byly následně použity pro odhad modelu při opětovných 100 pokusech, přičemž výsledná hodnota koeficientu byla určena jako průměr hodnot koeficientů daného ukazatele, viz tabulka 14.

Tabulka 14: Výsledné hodnoty koeficientů III. modelu

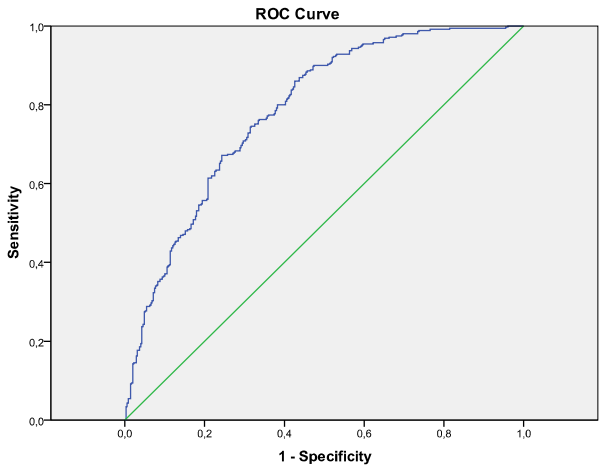
	B
lnx2	-,597
lnx6	,124
lnx7	-,363
lnx8	2,732
Constant	,457

Predikční schopnost výsledného modelu byla následně verifikována na kontrolním vzorku opět pomocí klasifikační tabulky a ROC analýzy, viz tabulky 15 a 16 a obrázek 6.

Tabulka 15: Klasifikační tabulka (III. model)

Observed			Predicted					
			Selected Cases a			Unselected Cases b		
			default		Percentage Correct	default		Percentage Correct
			0	1		0	1	
Step 5	default	0	177	70	71,7	68	35	66,0
		1	77	162	67,8	25	86	77,5
	Overall Percentage				***69,8***			***73,0***

Obrázek 6: ROC křivka (III. model)



Tabulka 16: Plocha pod ROC křivkou (III. model)

Area	Std. Errora	Asymptotic Sig.b	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,786	,017	,000	,752	,819

3.5 Komparace modelů

Tabulka 17 přehledně zobrazuje charakteristiky všech tří odhadnutých modelů. V prvním řádku je výčet proměnných v modelu, druhý řádek uvádí pseudo R2, které lze použít ke srovnání jednotlivých modelů, třetí řádek zobrazuje hodnotu Hosmer-Lemeshow testu, čtvrtý řádek procentuální úspěšnost klasifikace daného modelu na kontrolním vzorku a pátý velikost plochy pod ROC křivkou a poslední Giniho koeficient.[4]

Tabulka 17: Charakteristiky odhadnutých modelů

	I. model	***II. model***	***III. model***
proměnné	x1, x4, x5, x8	ln(x2), ln(x5), ln(x6), ln(x7)	ln(x2), ln (x5), ln(x7), ln(x8)
Nagelkerke R2	0,214	0,276	0,324
Hosmer - Lemeshow test	0,117	0,682	0,805
overall percentage	64,50%	69,10%	73,00%
AUC ROC	0,691	0,759	0,786
Gini. coef.	0,382	0,518	0,572

Z výsledků je vidět zřejmé zlepšování modelů při jejich postupném zpřesňování, tedy při zahrnutí analýzy vstupních dat, transformaci a zohlednění stability. Plocha pod ROC křivkou se u III. modelu dostala až na hodnotu 0,786, což je značný posun oproti původnímu I. modelu, což je typ odhadu, který se používá nejčastěji. Do výsledného modelu se nedostala žádná dummy proměnná indikující záporné hodnoty u ukazatelů I. a II. typu, což je sice překvapivé, ale může to být způsobeno právě tím, že v záporných hodnotách mají tyto ukazatele nejednoznačnou interpretaci.

4. Závěr

Příspěvek byl věnován vybraným úskalím a problémům při odhadu scoringových modelů na základě logistické regrese, a to konkrétně při využití poměrových ukazatelů jako nezávislých proměnných. Na ilustrativním příkladu byly demonstrovány rozdíly ve výsledných modelech a jejich predikční schopnosti při zohlednění chování jednotlivých ukazatelů na definičním oboru, stability ukazatelů a různých variant výběru vzorku pro odhad a vzorku kontrolního. Z výsledků je zřejmá nutnost provádět při odhadu modelů pečlivou analýzu vstupních poměrových ukazatelů, a to nejen z jejich matematického, ale také ekonomického hlediska. Jednotlivé ukazatele často nelze brát jako jednu veličinu, ale je nutné jejich rozdělení na kladnou a zápornou část. Rovněž odhad modelu pouze z jednoho náhodného vzorku může vést ke snížení celkové úspěšnosti modelu a jeho stability.

[4] Giniho koeficient je pouze obdobou plochy pod ROC křivkou (často se používá při interpretaci) a vypočítá se jako $G.K. = 2 \cdot AUCROC - 1$

References

[1] ALTMAN, E. I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance*, September, 1968, 589-609.

[2] ALTMAN, E. I., at al. *Application of Classification Techniques in Business, Banking and Finance*. JAI Press, Greenwich, 1981.

[3] HAIR, J. F., ANDERSON, R. E., TATHAM, R. L., BLACK, W. C. *Multivariate Data Analysis*. 6th ed. Prentice Hall, 2005.

[4] HANLEY, J. A., McNEIL, B. J. The Meaning and Use of the Area Under a Receiver Operating Characteristics (ROC) Curve. in *Radiology*, pp. 561-557, 1982.

[5] BEAVER, W. Financial ratios as predictors of failures. Empirical Research in Accounting: Selected Studies – 1966, supplement to *Journal of Accounting Research*, 4, 1967, 71-111.

[6] DURAND, B. Risk elements in consumer installments financing. Working paper, 1941, NBER.

[7] ENGELMANN, B., RAUHMEIER, R. (Eds.) *The Basel II Risk Parameters.* Springer Verlag, 2006.

[8] FISHER, R. The use of multiple measurements in taxonomic problems. *Annals of Eugenics,* 7, 1936, 179-188.

[9] GREEN, W. *Econometric Analysis*. 6th ed. Prentice Hall, 2008.

[10] GURNÝ, P., GURNÝ, M. Estimation of PD of financial institutions within linear discriminant analysis. *Mathematical Methods in Economics*. CZU Praha, 2009.

[11] QUEEN, M., ROLL, R. Firm mortality: using market indicators to predict survival. *Financial Analysts Journal 3*, 1987, 9–26.

[12] RESTI, A., SIRONI, A. *Risk management and Shareholders' value in banking*. Chichester: Wiley, 2007, 782 p.