INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 27, číslo 2, červen 2016

VYUŽITÍ LOGISTICKÉ REGRESE PŘI OBJEKTIVNÍM VYHODNOCOVÁNÍ VÝSLEDKŮ LÉKAŘSKÉ PÉČE

THE USE OF LOGISTIC REGRESSION IN OBJECTIVE EVALUATION OF THE RESULTS OF MEDICAL CARE

Marcela Rabasová

Adresa: VŠB-TU Ostrava, 17. listopadu 15, 708 33, Ostrava

E-mail: marcela.rabasova@vsb.cz

Abstrakt: Cílem tohoto příspěvku je stručné seznámení s principy logistické regrese a ukázka jejího použití v lékařské praxi. Logistická regrese patří mezi metody predikční diskriminační analýzy a v medicínské oblasti se používá nejen k tvorbě predikčních modelů, ale i četných skórovacích systémů. Tyto systémy, k nimž patří například systém POSSUM a systém tendenčních skóre, slouží k zajištění objektivity výsledků v nerandomizovaných lékařských studiích. V příspěvku je ukázána aplikace tendenčních skóre při porovnání výsledků laparoskopických a otevřených operací kolorekta z hlediska dlouhodobého přežívání. Analýza byla provedena na základě údajů o 850 pacientech s diagnózou kolorektálního karcinomu, kteří byli v letech 2001–2009 operováni na chirurgické klinice Fakultní nemocnice Ostrava. Doba přežití u obou operačních technik byla porovnána pomocí křivek přežití, k jejichž odhadům posloužila Kaplan-Meierova metoda. Vyhodnocení rozdílu mezi těmito křivkami bylo provedeno pomocí Breslowova a Mantel-Coxova (log-rank) testu.

Klíčová slova: logistická regrese, skórovací systémy, tendenční skóre, analýza přežívání, Kaplan-Meierova metoda.

Abstract: The aim of this paper is a brief introduction to the principles of logistic regression, and an illustration of its use in medical practice. Logistic regression belongs to the methods of predictive discriminant analysis and in the medical field it is used not only for creating predictive models, but also for creating numerous scoring systems. These systems, which include for example system POSSUM and propensity score system, serve to provide objectivity of results in non-randomized medical studies. The paper illustrates the application of propensity scores while results of laparoscopic and open colorectal surgeries are compared, as far as a long-term survival is concerned. The analysis is based on the data set of 850 patients diagnosed with colorectal cancer, who were operated on between 2001–2009 at the Surgical Clinic of the University Hospital Ostrava. The survival time due to both the surgical

techniques was compared, with survival curves estimated by Kaplan-Meier method. The differences between the curves has been assessed using Breslow and Mantel-Cox (log-rank) test.

Keywords: logistic regression, scoring systems, propensity scores, survival analysis, Kaplan-Meier method.

1. Úvod

Krátkodobé výsledky různých operačních technik jsou v dostupné literatuře často uváděny formou procentuálně vyjádřené morbidity (komplikací) a mortality (úmrtnosti). Porovnání takovýchto výsledků u různých operačních technik nebo mezi jednotlivými pracovišti event. chirurgy však může být zavádějící, protože nezohledňuje případnou odlišnost porovnávaných souborů v celé řadě důležitých charakteristik (tzv. „*case mix*“). Jednou z možností, jak objektivně porovnat morbiditu a časnou mortalitu, je hodnotit výsledky v souvislosti s individuálními riziky jednotlivých pacientů. K tomu slouží *skórovací systémy*, mezi něž patří například systém *POSSUM* a *tendenční skóre*. Oba tyto systémy využívají principů logistické regrese.

2. Logistická regrese

Logistická regrese patří mezi metody predikční diskriminační analýzy, jejímž hlavním cílem je zařazení objektů neznámého původu do předem vymezených skupin. Děje se tak prostřednictvím rozhodovacího pravidla, k jehož sestavení slouží skupina testovacích objektů. Jsou to objekty, u kterých známe hodnoty několika charakteristických veličin a jejich příslušnosti ke skupinám.

Předpokládejme situaci, kdy máme k dispozici n testovacích objektů s p naměřenými znaky, z nichž každý patří do jedné ze dvou skupin. Nechť naměřené znaky jsou u jednotlivých objektů reprezentovány p-rozměrnými náhodnými vektory $\mathbf{X}_1, \ldots, \mathbf{X}_n$ a příslušnost i-tého objektu k dané skupině nechť je vyjádřena hodnotou náhodné veličiny Y_i, která nabývá hodnot 0 nebo 1 podle toho, do které skupiny objekt náleží. U nového objektu, který chceme zařadit na základě vytvořeného rozhodovacího pravidla, nechť jsou naměřené znaky reprezentovány p-rozměrným náhodným vektorem $\mathbf{X}$ a příslušnost ke skupině náhodnou veličinou Y.

Model logistické regrese, popsaný např. v knihách [1] a [2], předpokládá, že $Y_1, \ldots, Y_n$ jsou nezávislé alternativní náhodné veličiny, jejichž podmíněná pravděpodobnost lze vyjádřit ve tvaru:

$$P(Y_i = 1 | \mathbf{X}_i = \mathbf{x}_i) = \frac{\mathrm{e}^{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i}}{1 + \mathrm{e}^{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i}},$$

$$P(Y_i = 0|\mathbf{X}_i = \mathbf{x}_i) = \frac{1}{1 + \mathrm{e}^{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i}},$$

pro $i = 1, \ldots, n$, kde $\mathbf{X}_i$ je p-dimenzionální náhodný vektor, $\mathbf{x}_i$ jeho realizace a $(\beta_0, \boldsymbol{\beta}^T)^T$ je neznámý, $(p+1)$-dimenzionální vektor parametrů. Jeho hodnoty odhadneme na základě známých hodnot $\mathbf{X}_i$ a Y_i u n testovacích objektů, čímž dostaneme i odhad funkce $\pi(\mathbf{x})$:

$$\pi(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{\mathrm{e}^{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}}}{1 + \mathrm{e}^{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}}}.$$

Ta udává pravděpodobnost toho, že objekt, kterému přísluší vektor pozorování $\mathbf{x}$, patří do skupiny 1. K nalezení optimálního rozhodovacího pravidla je pak využito *bayesovského přístupu*. Neznámým parametrem, o jehož hodnotě chceme rozhodnout, je náhodná veličina Y s oborem hodnot $\{0, 1\}$, která má pravděpodobnostní funkci $q(y)$. Rozhodnutí se provádí na základě hodnoty p-rozměrného náhodného vektoru $\mathbf{X}$ s hustotou $r(\mathbf{x})$. Nechť $r(\mathbf{x}|y)$ je podmíněná hustota $\mathbf{X}$ za podmínky $Y = y$, $\delta : \mathbb{R}^p \to \{0, 1\}$ rozhodovací funkce a $\mathcal{D}$ množina všech rozhodovacích funkcí $\delta : \mathbb{R}^p \to \{0, 1\}$. *Ztrátovou funkci* zavedeme jako:

$$L(Y, \delta(\mathbf{X})) = \begin{cases} 0 & \text{pokud } Y = \delta(\mathbf{X}), \\ 1 & \text{jinak,} \end{cases}$$

rizikovou funkci:

$$R(Y, \delta) = \mathsf{E}\,[L(Y, \delta(\mathbf{X}))|Y] = \int_{\mathbb{R}^p} L(Y, \delta(\mathbf{x})) r(\mathbf{x}|y)\, \mathrm{d}\mathbf{x},$$

a *bayesovské riziko*:

$$\rho(\delta) = \mathsf{E}\, R(Y, \delta) = \sum_{y=0}^{1} R(y, \delta) q(y).$$

Optimální rozhodovací funkcí je pak funkce:

$$\delta^* = \arg\min_{\delta \in \mathcal{D}} \rho(\delta).$$

V případě modelu logistické regrese tedy dostáváme: jestliže $\delta(\mathbf{x}) = j$, potom

$$\begin{aligned} \mathsf{E}\,[L(Y, \delta(\mathbf{X}))|\mathbf{X} = \mathbf{x}] &= \sum_{i=0}^{1} L(i, \delta(\mathbf{x})) P(Y = i|\mathbf{X} = \mathbf{x}) = \\ &= \begin{cases} \pi(\mathbf{x}) & \text{pro } j = 0, \\ 1 - \pi(\mathbf{x}) & \text{pro } j = 1, \end{cases} \end{aligned}$$

$$\min_{\delta \in \mathcal{D}} \mathsf{E}\left[L(Y, \delta(\mathbf{X}))|\mathbf{X} = \mathbf{x}\right] = \min\{\pi(\mathbf{x}), 1 - \pi(\mathbf{x})\}$$

a optimální rozhodovací funkce má tvar

$$\begin{aligned} \delta^*(\mathbf{x}) &= \arg\min_{\delta \in \mathcal{D}} \mathsf{E}\left[L(Y, \delta(\mathbf{X}))|\mathbf{X} = \mathbf{x}\right) = \\ &= \arg\min_{j=0,1} L(1-j, j) P(Y = 1 - j|\mathbf{X} = \mathbf{x}) = \\ &= \arg\max_{j=0,1} P(Y = j|\mathbf{X} = \mathbf{x}). \end{aligned}$$

To znamená, že objekty, kterým přísluší vektor pozorování $\mathbf{x}$ takový, že

$$\pi(\mathbf{x}) \geq 1 - \pi(\mathbf{x}),$$

(tj. $\beta_0 + \boldsymbol{\beta}^T \mathbf{x} \geq 0$), zařadíme do první skupiny, ostatní do nulté. Pokud $\pi(\mathbf{x}) = 1 - \pi(\mathbf{x})$, můžeme přitom objekt zařadit libovolně, aniž by se zvýšila pravděpodobnost chybné klasifikace. Místo neznámých parametrů β_0, $\boldsymbol{\beta}$ v praxi musíme použít jejich odhady $\widehat{\beta}_0$, $\widehat{\boldsymbol{\beta}}$, které získáme metodou maximální věrohodnosti. Model logistické regrese neklade žádné podmínky na rozdělení náhodných vektorů $\mathbf{X}_1, \ldots, \mathbf{X}_n$, ale předpokládá velmi specifický tvar pravděpodobnosti $P(Y = 1|\mathbf{X} = \mathbf{x})$, což vyžaduje ověření vhodným statistickým testem, např. *Hosmerovým-Lemeshowovým* [2].

3. Skórovací systém POSSUM

Skórovací systém POSSUM (a Physiological and Operative Severity Score for enUmeration of Mortality and morbidity) vznikl z potřeby jednoduchého skórovacího systému, který by byl použitelný napříč celým spektrem chirurgických výkonů. Byl vyvinut Copelandem a kol. [3] počátkem 90. let minulého století. K jeho odvození byly použity údaje 1372 pacientů operovaných v letech 1988 až 1989 ve Waltonské nemocnici v Liverpoolu. Původně tento systém sloužil jako nástroj pro porovnání výsledků mezi jednotlivými institucemi, ale jeho použití bylo později rozšířeno i na porovnání výsledků jednotlivých chirurgů a operačních technik.

Na začátku bylo do studie zahrnuto 62 rizikových faktorů pooperační morbidity a mortality, z nichž bylo diskriminační analýzou vybráno pouze 18 nejvýznamnějších, navzájem nezávislých faktorů, a to 12 faktorů souvisejících s fyziologickým stavem pacienta před operací (věk, kardiální příznaky, respirační příznaky, systolický krevní tlak, tepová frekvence, Glasgow coma score, hemoglobin, leukocyty, urea v séru, natrium v séru, kalium v séru, EKG) a 6 rizikových faktorů chirurgického výkonu (závažnost a rozsah operačního

výkonu, vícečetné operace v posledních 30 dnech, krevní ztráta, kontaminace peritoneální dutiny, přítomnost malignity, naléhavost operace). Každé z hodnot těchto faktorů, které významně ovlivňují pooperační morbiditu a mortalitu, jsou přiřazeny hodnoty 1, 2, 4 nebo 8, podle stupně rizikovosti. Součtem hodnot prvních 12 faktorů se získá tzv. *fyziologické skóre* (physiological score, PS) pacienta, součet hodnot zbývajících 6 faktorů tvoří tzv. *operační skóre* (operative score, OS) pacienta.

Logistickou regresí pak bylo vyjádřeno riziko morbidity R_1 vztahem:

$$\ln \frac{R_1}{1-R_1} = -5{,}91 + 0{,}16PS + 0{,}19OS$$

a riziko mortality R_2 vztahem:

$$\ln \frac{R_2}{1-R_2} = -7{,}04 + 0{,}13PS + 0{,}16OS,$$

kde PS je fyziologické skóre a OS operační skóre pacienta. Parametry fyziologického skóre se vztahují k okamžiku přijetí pacienta nebo k okamžiku bezprostředně před operací, operační skóre je doplněno po zákroku. POSSUM nezahrnuje takové faktory, jako např. rozdíly mezi jednotlivými chirurgy nebo anesteziology, ale je právě jedním z cílů tohoto systému na tyto rozdíly poukázat.

4. Tendenční skóre

Tendenční skóre byly představeny Rosenbaumem a Rubinem [4] v roce 1983 a v posledních desetiletích se rozšířily napříč celým spektrem lékařských studií. Podrobným popisem jejich výpočtu a použití se zabývá například Adamina a kol. ve své práci z roku 2006 [5].

Tendenční skóre nachází v medicínských aplikacích uplatnění zejména v situacích, kdy porovnáváme výsledky dvou nebo více léčebných postupů v nerandomizovaných studiích. Jedná se o studie, kde pacientům není určen léčebný postup náhodně, a které v praxi převažují, jak z etických tak z praktických důvodů. Může se tak stát, že ve skupině pacientů léčených metodou A je větší podíl rizikových pacientů, než je tomu u metody B, a při porovnání výsledků těchto metod z hlediska pooperační morbidity, mortality nebo pooperačního přežívání bychom na tento fakt měli brát zřetel.

Jednou z možností, jak vyřešit problém „nesourodosti“ porovnávaných skupin pacientů, je přiřadit každému pacientu tzv. tendenční skóre, které vystihuje pravděpodobnost (tendenci) toho, že pacient bude léčen konkrétní metodou. Tato pravděpodobnost může záviset na mnoha faktorech, jako jsou

například věk, pohlaví, diagnóza, komorbidita, počet předchozích operací a podobně. K výpočtu tendenčních skóre se používá logistická regrese. Zmíněné faktory, které mohou ovlivnit výběr léčebné metody, mají přitom funkci nezávislých proměnných, léčebná metoda samotná představuje závislou proměnnou. Každému pacientu je vytvořeným logistickým modelem vypočtena pravděpodobnost, že bude léčen konkrétní metodou (jeho tendenční skóre), a z původního nerandomizovaného výběru se provede výběr užší, ve kterém jsou zastoupeni pouze ti pacienti, kteří mají v druhé skupině vhodný protějšek – pacienta se stejným skóre. To znamená, že pacient léčený metodou A je porovnáván s pacientem, který měl stejnou šanci být léčen metodou A, ale ve skutečnosti byl léčen metodou B. Aplikace tendenčních skóre tak zajistí alespoň jistý stupeň randomizace a eliminuje vliv přidružených faktorů na výsledky analýzy.

V další kapitole je popsáno užití tendenčních skóre při porovnání dlouhodobého přežívání otevřených a laparoskopických operací kolorekta u skupiny pacientů operovaných na chirurgické klinice Fakultní nemocnice Ostrava.

5. Porovnání dlouhodobého přežívání otevřených a laparoskopických operací kolorekta

V letech 2001–2009 podstoupilo ve Fakultní nemocnici Ostrava *kolorektální operaci* 850 pacientů. Podle závěrů studie [6], která byla provedena na této klinice, byla z hlediska délky přežívání *laparoskopická technika* statisticky významně lepší než *otevřená* v případě operací karcinomu v oblasti kolon (tlustého střeva), v případě operací karcinomu v oblasti rekta (konečníku) se jevily obě operační techniky jako ekvivalentní. Jelikož tato studie *nebyla randomizovaná* (typ operační techniky nebyl pacientům přiřazován náhodně), nebyla zaručena stejnorodost porovnávaných skupin pacientů v celé řadě důležitých charakteristik. Proto vyvstala potřeba ověřit věrohodnost závěrů této studie použitím některé z technik umožňujících jistou *pseudorandomizaci*. Ve studii [7] byla jako prostředek k dosažení tohoto cíle použita tendenční skóre.

5.1. Popis analyzovaného souboru

Zdrojovými daty pro tuto analýzu byly údaje o 850 pacientech chirurgické kliniky Fakultní nemocnice Ostrava, kteří zde v letech 2001–2009 podstoupili operaci kolorektálního karcinomu. Z monitorovaných údajů byly pro naši studii podstatné: datum operace, datum poslední kontroly a informace, zda pacient zemřel či nikoliv. Pro aplikaci tendenčních skóre pak byly důležité hodnoty veličin: *věk, BMI, ASA klasifikace, ICHS* (výskyt ischemické cho-

roby srdeční), *DM* (diabetes mellitus), a *stádium nádoru*, které mohou výrazně ovlivnit pooperační mortalitu. Pacienti, u nichž některý z těchto údajů chyběl, nebyli do studie zařazeni. Z celkového počtu 850 pozorování tak bylo použito jen 809, z nichž 500 odpovídalo karcinomu tlustého střeva (colon) a 309 karcinomu konečníku (rectum). Doba přežití u obou operačních technik byla porovnána pomocí *křivek přežití*, k jejichž odhadům posloužila *Kaplan-Meierova metoda*. Vyhodnocení rozdílu mezi těmito křivkami bylo provedeno pomocí *Breslowova* a *Mantel-Coxova (log-rank) testu*.

5.2. Vyhodnocení výsledků operací v oblasti kolon

Operaci s diagnózou rakoviny tlustého střeva prodělalo 500 pacientů, z toho 202 mužů a 298 žen ve věkovém rozmezí 26–97 let. Jejich záznamy byly zkoumány s cílem zjistit, jestli se obě operační techniky – laparoskopická a otevřená, významně liší v době přežívání. Nejprve bylo pro konstrukci Kaplan-Meierových odhadů použito všech 500 pozorování. Z celkového počtu 500 operací bylo 272 provedeno laparoskopicky a 228 otevřeně (v první skupině bylo 183 cenzorovaných případů, ve druhé 113).

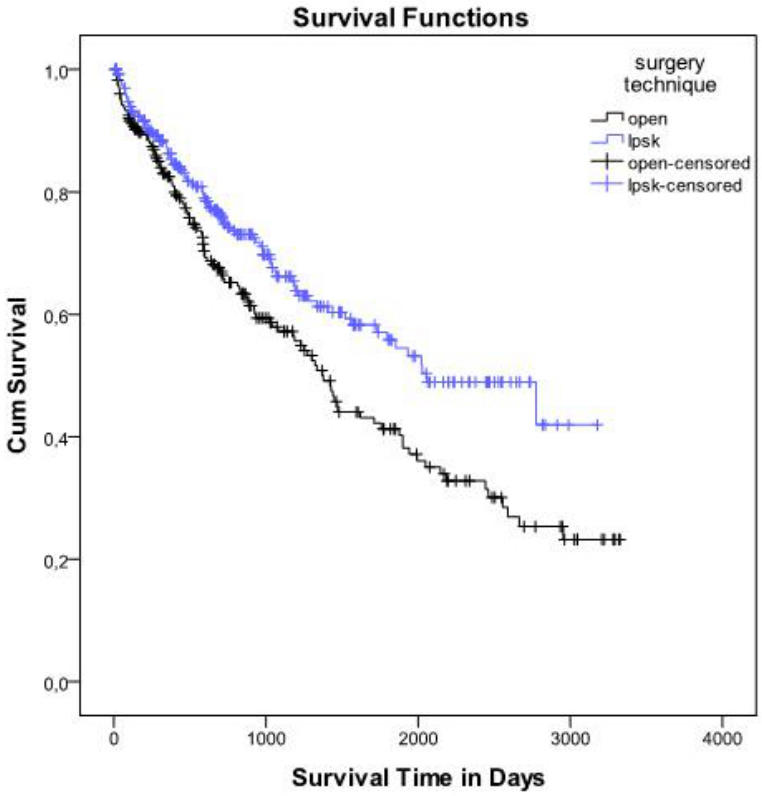
Tabulka 1: Mediány a střední doby přežívání – karcinom kolon (500 případů)

Střední doba	Odhad	Směr. odch.	95% konf. interval
Otevřené operace	1612,55	95,54	1425,30 – 1799,81
Laparoskopické operace	1979,99	96,56	1790,74 – 2169,24

Medián	Odhad	Směr. odch.	95% konf. interval
Otevřené operace	1372	102,88	1170,35 – 1573,65
Laparoskopické operace	2056	324,95	1419,11 – 2692,89

V tabulce 1 jsou uvedeny mediány a střední hodnoty doby přežití u obou skupin, včetně jejich směrodatných odchylek a intervalových odhadů. Již z těchto hodnot je patrné, že pacienti operovaní laparoskopicky přežívali déle (průměrně 1980 dnů od operace) než pacienti operovaní otevřeně (průměrně 1613 dnů). Ještě markantnějšího rozdílu si můžeme všimnout u mediánů obou skupin, jejichž hodnoty jsou 1372 dnů u otevřených a 2056 dnů u laparoskopických operací.

Názornou představu o dobách přežívání pacientů obou skupin získáme z vizuálního porovnání jejich křivek přežití, viz obrázek 1.



Obrázek 1: Křivky přežití – karcinom kolon (500 případů)

Jelikož křivka přežití pro skupinu 1 – laparoskopie, leží výrazně nad křivkou přežití pro skupinu 0 – otevřené operace, mohli bychom konstatovat, že laparoskopické operace v oblasti kolorekta dosahují z hlediska dlouhodobého přežívání mnohem lepších výsledků než operace otevřené. Že je rozdíl mezi nimi statisticky významný (na hladině významnosti 0,05), bylo potvrzeno jak Breslowovým testem ($p = 0{,}018$), tak log-rank testem ($p = 0{,}004$).

Tyto výsledky však mohou být zavádějící, jelikož studie není randomizovaná. Porovnávané skupiny pacientů se tak mohou významně lišit v celé řadě charakteristik, což může zapříčinit výraznou převahu pacientů s větší tendencí k pooperační mortalitě v některé z těchto skupin. Abychom se vypořádali s tímto problémem, byla pro každého pacienta vypočtena pravděpodobnost jeho úmrtí (jeho tendenční skóre) a z původního nerandomizovaného výběru se provedl výběr užší, ve kterém byli zastoupeni pouze ti pacienti, kteří měli ve druhé skupině vhodný protějšek – pacienta se stejným skóre. Aplikace tendenčních skóre tak zajistila alespoň jistý stupeň randomizace a eliminovala vliv přidružených faktorů na výsledky analýzy.

Pomocí normální diskriminační analýzy byly nalezeny parametry, které mají významný vliv na úmrtí pacienta po provedené operaci. Byly to veličiny:

věk, BMI, ASA klasifikace, ICHS, DM, a *stádium nádoru*. Na základě těchto šesti proměnných byl vytvořen model logistické regrese, který pro konkrétního pacienta vyjádřil riziko jeho pooperační mortality R vztahem:

$$\ln \frac{R}{1-R} = -3{,}79 + 0{,}02x_1 - 0{,}05x_2 + 0{,}11x_3 + 0{,}54x_4 + 0{,}41x_5 + 1{,}01x_6 \quad (1)$$

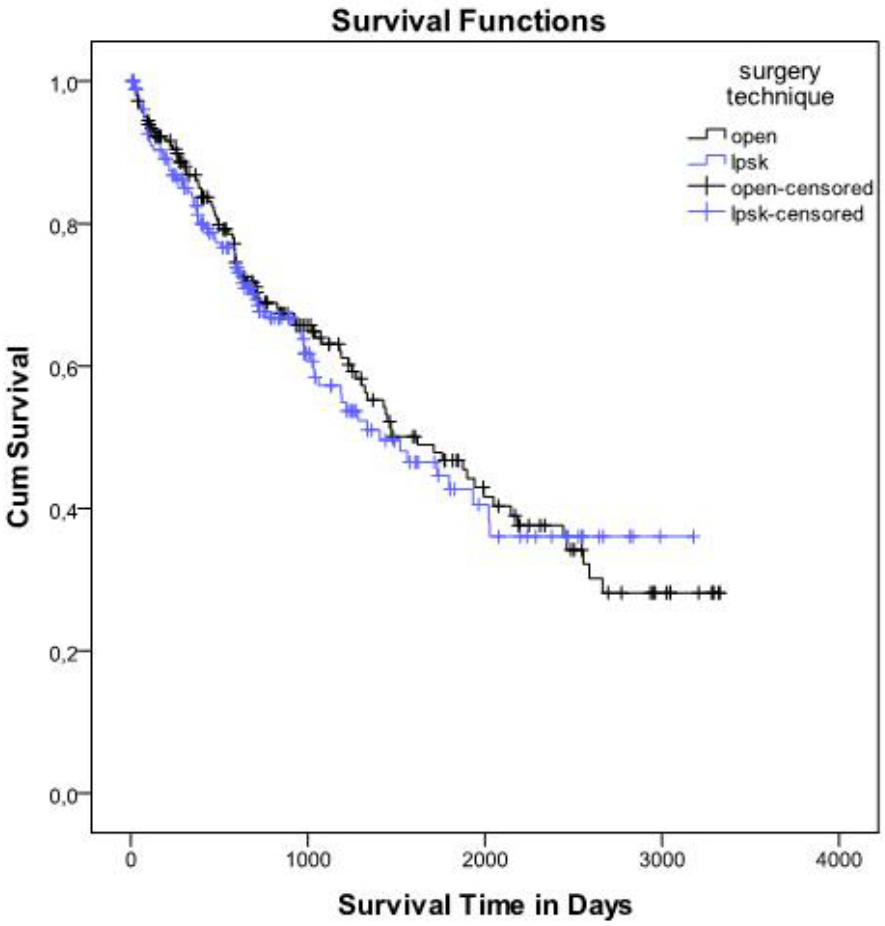
kde x_1 je hodnota proměnné *věk*, x_2 je hodnota proměnné *BMI*, x_3 je hodnota proměnné *ASA klasifikace*, x_4 je hodnota proměnné *ICHS*, x_5 je hodnota proměnné *DM*, x_6 je hodnota proměnné *stádium nádoru* pro daného pacienta. Pomocí rovnice (1) bylo pro každého pacienta vypočteno riziko jeho úmrtí, jeho tendenční skóre, a každému pacientu ze skupiny 1 (laparoskopické operace) byl přiřazen pacient ze skupiny 0 (otevřené operace) se stejným skóre (pokud takový pacient existoval). Touto cestou byl pořízen výběr 366 pacientů, rozdělených do dvou stejně velkých skupin, které byly srovnatelné ve smyslu tendence pacientů k pooperační mortalitě. V souboru získaném touto selekcí bylo 56,6 % cenzorovaných případů, ve skupině 1 tvořil jejich podíl 59,6 %, ve skupině 0 pak 53,6 %. Pro obě nové skupiny byly zkonstruovány Kaplan-Meierovy křivky přežití, jejichž grafy jsou zachyceny na obrázku 2.

Vidíme, že poloha obou křivek je nyní zcela odlišná. Křivka přežití u laparoskopických operací teď leží z velké části pod křivkou odpovídající operacím otevřeným. Rovněž medián (1406 dnů) a střední doba přežití (1703 dnů) u této metody jsou nižší než u operací otevřených, jak ukazuje tabulka 2. Z porovnání konfidenčních intervalů zmíněných charakteristik v obou skupinách je však patrné, že rozdíly mezi oběma operačními technikami nejsou příliš významné.

Tabulka 2: Mediány a střední doby přežívání – karcinom kolon (366 případů)

Střední doba	Odhad	Směr. odch.	95% konf. interval
Otevřené operace	1765	107,09	1555,11 – 1974,89
Laparoskopické operace	1703	117,40	1472,86 – 1933,07

Medián	Odhad	Směr. odch.	95% konf. interval
Otevřené operace	1619	232,57	1163,16 – 2074,84
Laparoskopické operace	1406	229,84	955,51 – 1856,49



Obrázek 2: Křivky přežití – karcinom kolon (366 případů)

To, že rozdíly v délce přežívání u obou operačních technik nejsou statisticky významné, bylo potvrzeno jak Breslowovým ($p_1 = 0{,}724$), tak log-rank testem ($p_2 = 0{,}503$). Laparoskopické operace prováděné na chirurgické klinice Fakultní nemocnice Ostrava u pacientů s karcinomem kolon tedy nebudeme z hlediska pooperačního přežívání hodnotit jako horší než otevřené, v žádném případě je však nemůžeme hodnotit jako lepší, jak je tomu ve studii [6], ve které nebyl zohledněn fakt, že studie nebyla randomizovaná.

5.3. Vyhodnocení výsledků operací v oblasti rekta

Pacientů s karcinomem rekta bylo celkem 309, z toho 201 mužů a 108 žen. Jejich věk se pohyboval v rozmezí 33–84 let, 189 operací bylo provedeno laparoskopicky a 120 otevřeně (v první skupině bylo 124 cenzorovaných případů, ve druhé 62). I zde nás zajímalo, jak se liší doba přežívání laparoskopických a otevřených operací. Nejprve byly Kaplan-Meierovy odhady křivek přežití sestrojeny na základě všech 309 vícerozměrných pozorování.

Z tabulky 3, ve které jsou uvedeny mediány a střední hodnoty doby přežití u obou skupin, je patrné, že se délka přežívání u porovnávaných operačních technik příliš neliší. Střední doba přežívání je větší u otevřených operací (1902 dnů oproti 1809 dnům u operací laparoskopických), zatímco medián je větší u laparoskopicých operací (1739 dnů oproti 1572 dnům u operací otevřených). Konfidenční intervaly v obou skupinách se pak u obou zmíněných charakteristik výrazně překrývají.

Tabulka 3: Mediány a střední doby přežívání – karcinom rekta (309 případů)

Střední doba	Odhad	Směr. odch.	95% konf. interval
Otevřené operace	1901,74	128,83	1649,24 – 2154,24
Laparoskopické operace	1809,15	117,17	1579,49 – 2038,80

Medián	Odhad	Směr. odch.	95% konf. interval
Otevřené operace	1572	346,60	892,70 – 2251,34
Laparoskopické operace	1739	413,79	927,97 – 2550,03

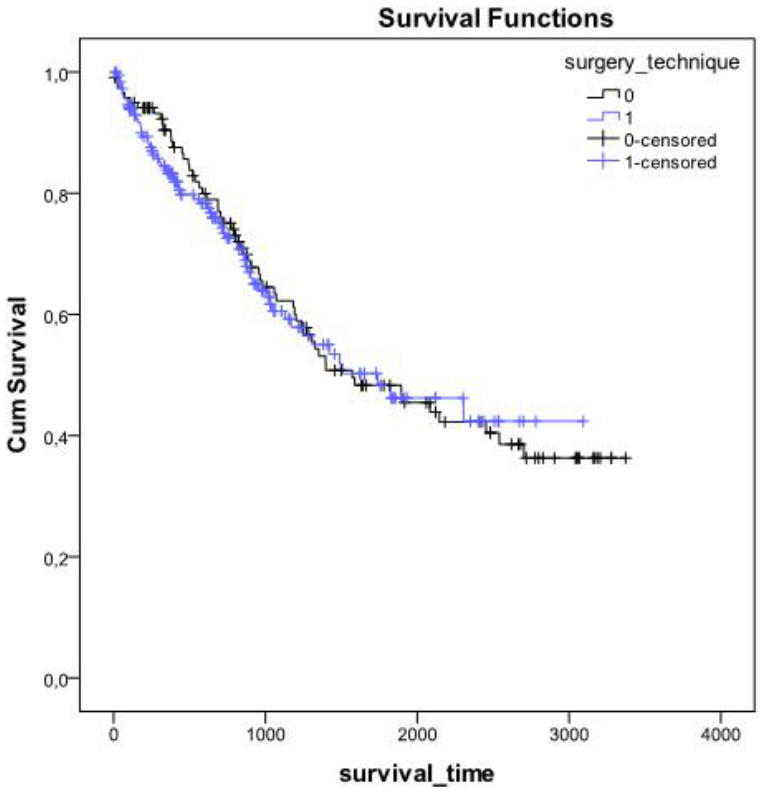
Že se obě operační techniky v době přežívání příliš neliší, dokládá i obrázek 3 s Kaplan-Meierovými odhady jejich křivek přežití.

Jelikož mají tyto křivky přibližně stejný průběh, můžeme předpokládat, že v době přežívání není mezi oběma operačními technikami statisticky významný rozdíl. Tento závěr byl potvrzen Breslowovým ($p = 0{,}839$) i log-rank ($p = 0{,}498$) testem.

Pro verifikaci tohoto závěru byla opět aplikována tendenční skóre. Byla přitom použita stejná skupina prediktorů úmrtí jako v případě karcinomu kolon. Na základě tendenčních skóre vypočtených podle vzorce (1) byl z původního souboru proveden výběr čítající 200 pacientů, seskupených do dvou stejně velkých skupin, které byly vyvážené z hlediska tendence pacientů k pooperačnímu úmrtí. Tím se staly obě skupiny pacientů srovnatelné bez nutnosti randomizace.

Podíl cenzorovaných případů činil v tomto zúženém výběru 58,5 %, ve skupině 0 (otevřené operace) 54,0 % a ve skupině 1 (laparoskopické operace) 63,0 %. Mediány a střední doby přežití pro obě skupiny jsou prezentovány v tabulce 4.

Tabulka nevypovídá o delší době přežívání u žádné z operačních technik. Střední doba přežívání je větší u otevřené techniky (1939 dnů), ale medián u techniky laparoskopické (1822 dnů). Rozdíly v těchto charakteristikách mezi



Obrázek 3: Křivky přežití – karcinom rekta (309 případů)

Tabulka 4: Mediány a střední doby přežívání – karcinom rekta (200 případů)

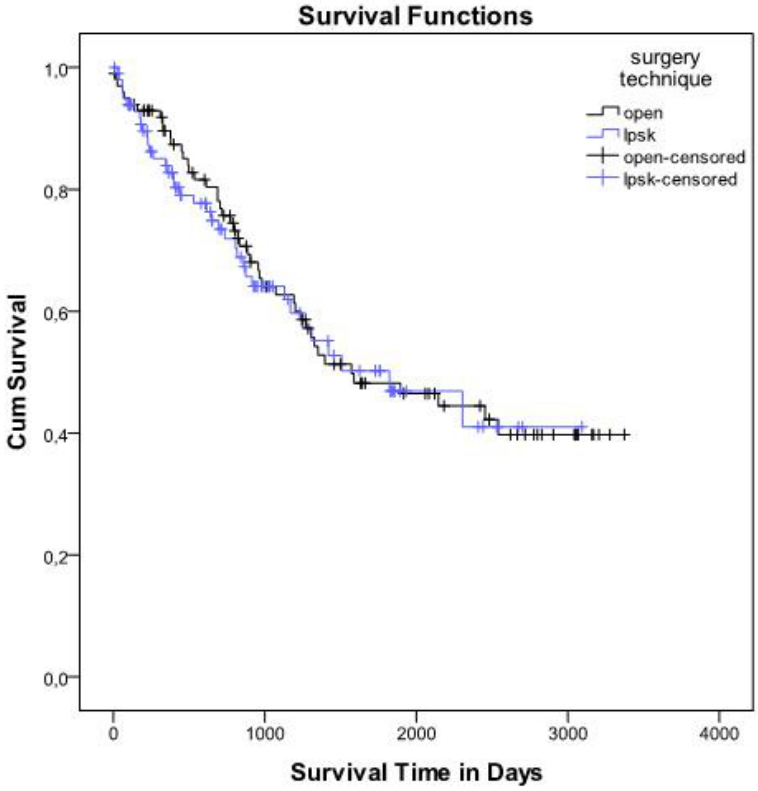
Střední doba	Odhad	Směr. odch.	95% konf. interval
Otevřené operace	1938,97	144,47	1655,80 – 2222,13
Laparoskopické operace	1803,29	152,78	1503,84 – 2102,74

Medián	Odhad	Směr. odch.	95% konf. interval
Otevřené operace	1572	427,49	734,12 – 2409,88
Laparoskopické operace	1822	492,57	856,57 – 2787,43

jednotlivými skupinami navíc nejsou nikterak velké a jejich konfidenční intervaly se výrazně překrývají.

Kaplan-Meierovy křivky přežití pro obě skupiny jsou zachyceny na obrázku 4. I zde je vidět, že rozdíly mezi oběma skupinami jsou minimální.

Naše závěry byly potvrzeny Breslowovým i log-rank testem, ani jeden z nich nevyhodnotil rozdíl mezi dobou přežívání u laparoskopických a otevřených operací v oblasti rekta jako statisticky významný ($p_1 = 0{,}774$, $p_2 =$



Obrázek 4: Křivky přežití – karcinom rekta (200 případů)

0,570). Při vyhodnocování operačních záznamů v případě karcinomu rekta jsme tedy došli ke stejným závěrům jako studie [6].

6. Shrnutí

Cílem tohoto příspěvku bylo ukázat přínos logistické regrese v medicínské oblasti, kde je tato metoda úspěšně používána nejen k tvorbě predikčních modelů, ale i četných skórovacích systémů, které usilují o kvantifikaci rizik chirurgických pacientů, čímž mohou ovlivnit rozhodování o rozsahu vyšetření, způsobu a agresivitě léčby, typu a rozsahu výkonu a předoperační přípravě a podílet se tak na racionalizaci nákladů a zkvalitnění nemocniční péče.

Jako ukázka byla použita aplikace tendenčních skóre při porovnání dlouhodobého přežívání laparoskopických a otevřených operací kolorekta. Studie [6], která nijak neřešila problém randomizace, vyhodnotila laparoskopické operace jako významně lepší v délce přežívání u operací v oblasti kolon. Jak se ale ukázalo díky použití tendenčních skóre, nebyly porovnávané skupiny pacientů rovnocenné v důležitých charakteristikách, které mají na pooperační přežívání významný vliv. U operací v oblasti kolon bylo ve skupině operované otevřenou technikou větší procento rizikových pacientů, což bylo příči-

nou častějších úmrtí v této skupině a vedlo k nesprávným závěrům o kratší době přežívání této operační techniky.

Aplikace tendenčních skóre umožnila pořídit z původního datového souboru výběr užší, ve kterém byly obě porovnávané skupiny rovnocenné ve smyslu tendence pacientů k úmrtí po provedené operaci, čímž byla zajištěna větší objektivita výsledků. Podle závěrů získaných po této tzv. pseudorandomizaci se laparoskopická a otevřená operační technika neliší v délce přežívání ani u operací v oblasti rekta, ani u operací v oblasti kolon.

Poděkování

Práce vznikla díky podpoře grantu The European Development Fund in the IT4Innovations Centre of Excellence, CZ.1.05/1.1.00/02.0070.

Literatura

[1] Hebák, P., Hustopecký, J., Pecáková, I., et al.: *Vícerozměrné statistické metody [3].* Praha, Informatorium, 2007. ISBN 80-7333-039-3.

[2] Hosmer, D. W., Lemeshow, S.: *Applied Logistic Regression.* New York, Wiley-Interscience, 2000. ISBN 0-471-35632-8.

[3] Copeland, G. P., Jones, D., Walters, M.: POSSUM: a scoring system for surgical audit. *British Journal of Surgery* 78, pp. 356–360, 1991.

[4] Rosenbaum, P. R., Rubin, D. B.: The central role of the propensity score in observational studies for casual affects. *Biometrika* 70, pp. 41–55, 1983.

[5] Adamina, M., Guller, U., Weber, W. P., Oertli, D.: Propensity scores and the surgeon. *British Journal of Surgery* 93, pp. 389–394, 2006.

[6] Briš, R., Praks, P., Janurová, K., et al.: Survival analysis on data of different surgery techniques to evaluate risk of postoperative complications. In: *Journal of Polish Safety and Reliability Association: Summer Safety and Reliability Seminars, SSARS 2011, Volume 1.* Gdynia, Polsko: Polish Safety and Reliability Association, 2011, pp. 33–38. ISBN 978-83-925436-2-6.

[7] Rabasová, M., Briš, R., Martínek, L.: Propensity score analysis as a tool for randomization in a non-randomized surgical study. *Proceedings of MENDEL 2012, 18th International Conference on Soft Computing, June 27–29, Brno, Czech Republic.* Brno: Vysoké učení technické v Brně, pp. 464–469, 2012. ISBN 978-80-214-4540-6.

MODELOVÁNÍ RIZIKA POOPERAČNÍCH KOMPLIKACÍ

MODELLING THE RISK OF POSTOPERATIVE COMPLICATIONS

Žaneta Miklová

Adresa: FEI Vysoká škola báňská – Technická univerzita Ostrava, 17. listopadu 2172/15, 708 00 Ostrava

E-mail: zaneta.miklova@vsb.cz

Abstrakt: V tomto článku bych chtěla odprezentovat výsledky analýzy přežití u pacientů po operaci kolorekta. Data, se kterými pracuji, mi poskytla Fakultní nemocnice Ostrava. Tato data jsou cenzorovaná. Cenzorovaným údajem je zde doba do pooperačních komplikací. Model, který k této analýze požívám, je Coxův proporcionální hazardní model. Tento model pak reprezentuje riziko pooperačních komplikací v závislosti na různých sledovaných proměnných. Cílem je potvrdit nebo vyvrátit, že typ operace (otevřená nebo laparoskopická) má u pacientů vliv na riziko pooperačních komplikací. V případě, že typ operace je statisticky významná proměnná, chci vyhodnotit, která z těchto operací je pro pacienta méně rizikovější.

Klíčová slova: analýza přežití, cenzorovaná data, Coxův proporcionální hazardní model, částečná věrohodnostní funkce, aproximace částečné věrohodnostní funkce.

Abstract: In this article I would like to present results of survival analysis of pacients after colorectum surgery. I am working with the data, which were provided by the University Hospital Ostrava. These data are censored namely the time until after-surgery complications. For purpose of this analysis I use Cox's proportional hazard model. This model is representing risk of after-surgery complications depending on different variables. The objective is to confirm or reject, that type of operation (laparoscopic or open) has an impact on the risk of after-surgery complications. In case that type of operation is statistically significant variable, I want to evaluate which kind of these types are less risky for pacients.

Keywords: survival analysis, censored dat, Cox proportional hazard model, partial likelihood function, approximation of partial likelihood function.

1. Úvod

V rámci své diplomové práce jsem se seznámila s teorii analýzy přežití, konkrétně s Cox proporcionálním hazardním modelem. V tomto článku předkládám závěry z analýzy přežití pacientů po operaci kolorekta, ke kterým jsem došla. A odpovím na otázku, jestli je laparoskopická metoda pro pacienty lepší než metoda otevřená.

2. Seznámení s daty

Díky spolupráci s Fakultní nemocnicí Ostrava jsem měla k dispozici data pacientů po operaci kolorekta. U pacientů jsou sledovány jednak základní údaje jako jsou *věk*, *pohlaví*, *BMI*, ale také proměnné typu *krevní ztráta* (v ml), *předchozí operace* (ano/ne) nebo *délka operace* (v minutách). Celkem se v datovém souboru nachází 23 proměnných včetně primární proměnné *typ operace* (laparoskopická/otevřená), vysvětlující proměnné *doba do pooperační komplikace* a informace o cenzorování.

Další proměnné, které mám k dispozici, jsou diagnóza (0 – c18, 1 – c19, 2 – c20), ASA (*Amer. Society of Anaesthesiology Classification*, 1–4), ICHS (0 – ne, 1 – ano), arytmie (0 – ne, 1 – ano), hypertenze (0 – ne, 1 – ano), cebrovaskulární (0 – ne, 1 – ano), pulmonální (0 – ne, 1 – ano), DM (0 – ne, 1 – ano), renální (0 – ne, 1 – ano), jaterní (0 – ne, 1 – ano), perop. komplikace (0 – ne, 1 – ano), konverze (kódováno hodnotami 0–2), stádium (1–4), grading (0–2).

Data, se kterými pracuji, jsou cenzorovaná. Cenzorování známe různého druhu [1]. Prakticky to znamená, že pokud je sledovaná proměnná doba do pooperační komplikace cenzorovaná, nevíme přesnou dobu do pooperačních komplikací, ale víme, že tato doba je větší nebo rovna než zaznamenaný údaj. Často je takovýto čas doprovázen symbolem +. V praxi k takové situaci může dojít, pokud ztratíme s pacientem kontakt, např. pacient se odstěhuje nebo nemá zájem pokračovat.

V tabulce 1 vidíme, kolik pacientů je v datovém souboru a kolik z nich má cenzorovaný čas do pooperačních komplikací.

Tabulka 1: Četnosti cenzorovaných/necenzorovaných údajů

Údaj	n	V procentech
Necenzorované časy $c = 1$	426	54,3%
Cenzorované časy $c = 0$	359	45,7%
Celkem	785	100,0%

3. Teoretický základ

Než přistoupím k interpretaci závěrů vyplývajících z modelu, shrnu pár důležitých poznatků, jak vlastně tento model vůbec vypadá a co bychom měli vědět ještě před tím, než spustíme software a příslušnou funkci, která sestaví model na základě dostupných dat.

3.1. Coxův proporcionální hazardní model

Tento model řadíme do třídy tzv. PH modelů (Proporcional Hazards Models), které na rozdíl od AFT modelů (Accelerated Failure Time Models) zahrnují vliv vstupních vysvětlujících proměnných jako modifikaci základní hazardní funkce [5]

$$h(t) = h_0(t)\mathrm{e}^{\beta' x}, \tag{1}$$

kde $h(t)$ je hazardní funkce, $h_0(t)$ je základní hazardní funkce, β je vektor neznámých koeficientů a x je vektor vstupních vysvětlujících proměnných [5].

3.1.1. Částečná věrohodnostní funkce. V souvislosti s tímto modelem se dále zavádí pojem částečné věrohodnostní funkce [3, 6]

$$\ell_p(\beta) = \prod_{i=1}^{n} \left[\frac{\mathrm{e}^{x_i\beta}}{\sum_{j\in R(t_i)} \mathrm{e}^{x_j\beta}} \right]^{c_i}, \tag{2}$$

kde $R(t_i)$ je riziková skupina v čase t_i, c_i je informace o cenzorování u i-tého pacienta. Na rozdíl, od klasické věrohodnostní funkce tohoto modelu, závisí částečná věrohodnostní funkce pouze na vstupních vysvětlujících proměnných a je tak možné odhadnout koeficienty β tohoto modelu pomocí metody maximální věrohodnosti (MLE).

3.1.2. Aproximace částečné věrohodnostní funkce. Tak jak je částečná věrohodnostní funkce napsaná, platí pouze pro případy, kdy se nám v datech neopakují pozorované časy do pooperačních komplikací. Tedy nejsou zde pacienti, kteří by tyto časy měli stejné. Aby se vyřešil tento problém, pracujeme s aproximací částečné věrohodnostní funkce. Těchto aproximací existuje více:

- Breslewova aproximace
- Efronova aproximace
- Coxova aproximace

Z nichž nejpoužívanější je Efronova, protože je z nich nejpřesnější [5].

3.1.3. Odhad koeficientů β a jejich významnost v modelu. Odhad koeficientů v aproximované částečné věrohodnostní funkci se provádí buď pomocí metody MLE [4] nebo pomocí metody Newton-Rhapsona (iterační metoda) [2].

Pro posouzení významnosti koeficientů v modelu, tedy zda daná vysvětlující proměnná má statisticky významný vliv na dobu do pooperačních komplikací, použiji test poměru částečné věrohodnostní funkce [5].

4. Výsledky

Pro zpracování výsledků jsem použila statistický program R, který má v balíčku survival naiplementovanou funkci k vytvoření Coxova modelu. Je nutné říct, že kategoriální proměnné se musí zakódovat pomocí nul a jedniček, viz tabulka 2, která ukazuje, jak se proměnná ASA (4 úrovně) rozpadne na tři samostatné proměnné. Dále je třeba si zvolit vhodnou aproximaci (zde jsem zvolila Efronovu) a určit, které proměnné do modelu zahrneme.

Tabulka 2: Kódování kategoriální proměnné

Úroveň ASA	ASA_1	ASA_2	ASA_3
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

K určení, které proměnné jsou statisticky významné, jsem využila test poměru částečné věrohodnostní funkce a smíšenou strategii výběru. Podrobný popis, jak tento výběr probíhal, můžete nalézt v [5].

Tabulka 3: Poslední krok obecné strategie

Model	$-2\ln\ell(\widehat{\beta})$	Pokles	df	p-hodnota
stádium + *věk* + *grading* + *arytmie*	4947,392			
typ operace	4945,889	1,503	1	0,220

Jako statisticky významné se jeví proměnné *stádium*, *věk*, *grading* a *arytmie*. Po přidání primární proměnné *typ operace* dospěji k závěru, že tato proměnná nemá statisticky významný vliv na riziko pooperačních kompli-

kací. Tedy neprokázala se mi domněnka, že by jedna z operativních metod (laparoskopická/otevřená) byla lepší než ta druhá. Nicméně jsem identifikovala, které proměnné a v jaké míře mají statisticky významný vliv na riziko pooperačních komplikací.

Pro model obsahující vysvětlující proměnné *stádium*, *arytmie*, *grading* a *věk* dostaneme odhady koeficientů β, které jsou uvedeny v tabulce 4.

Tabulka 4: Souhrn výsledků modelu

Vysvětlující proměnná	$\widehat{\beta}$	SE($\widehat{\beta}$)	$e^{\widehat{\beta}}$	95% CI pro $e^{\widehat{\beta}}$	
				dolní	horní
$stádium_1$	0,472	0,208	1,603	1,066	2,410
$stádium_2$	1,180	0,198	3,254	2,207	4,798
$stádium_3$	2,560	0,199	12,936	8,754	19,106
věk	0,026	0,005	1,026	1,016	1,036
arytmie	0,517	0,147	1,676	1,256	2,237
$grading_1$	−0,183	0,109	0,833	0,673	1,031
$grading_2$	0,278	0,159	1,321	0,969	1,801

Výsledný model má tvar

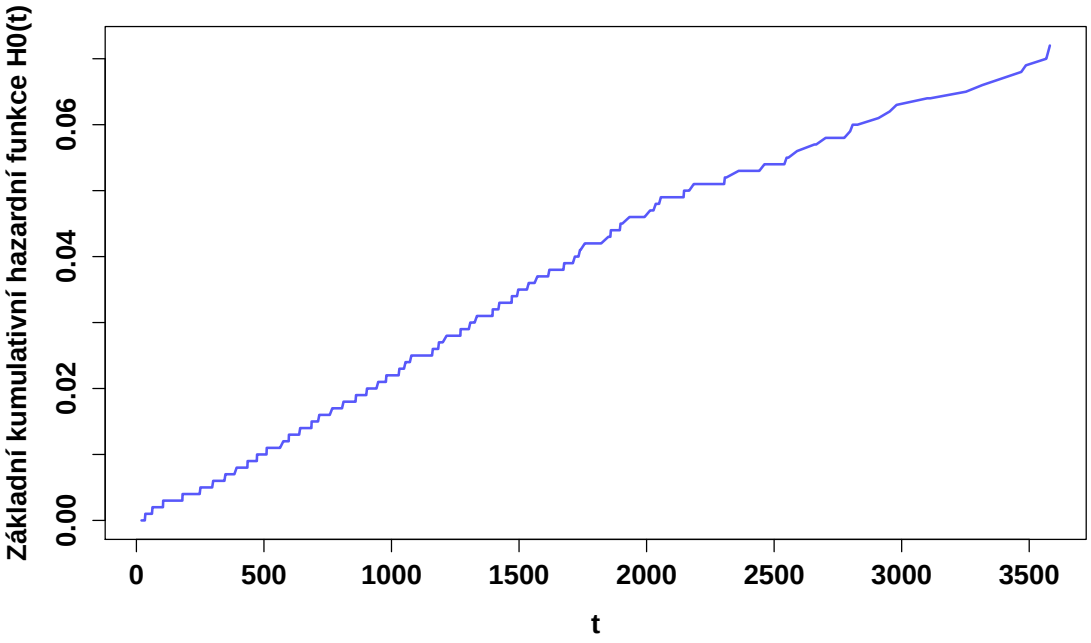
$$h(t) = h_0(t)\mathrm{e}^{0,472ST_1+1,180ST_2+2,560ST_3+0,517AR+0,026věk-0,183GR_1+0,278GR_2},$$

kde ST reprezentuje *stádium*, AR *arytmii* a GR *grading*.

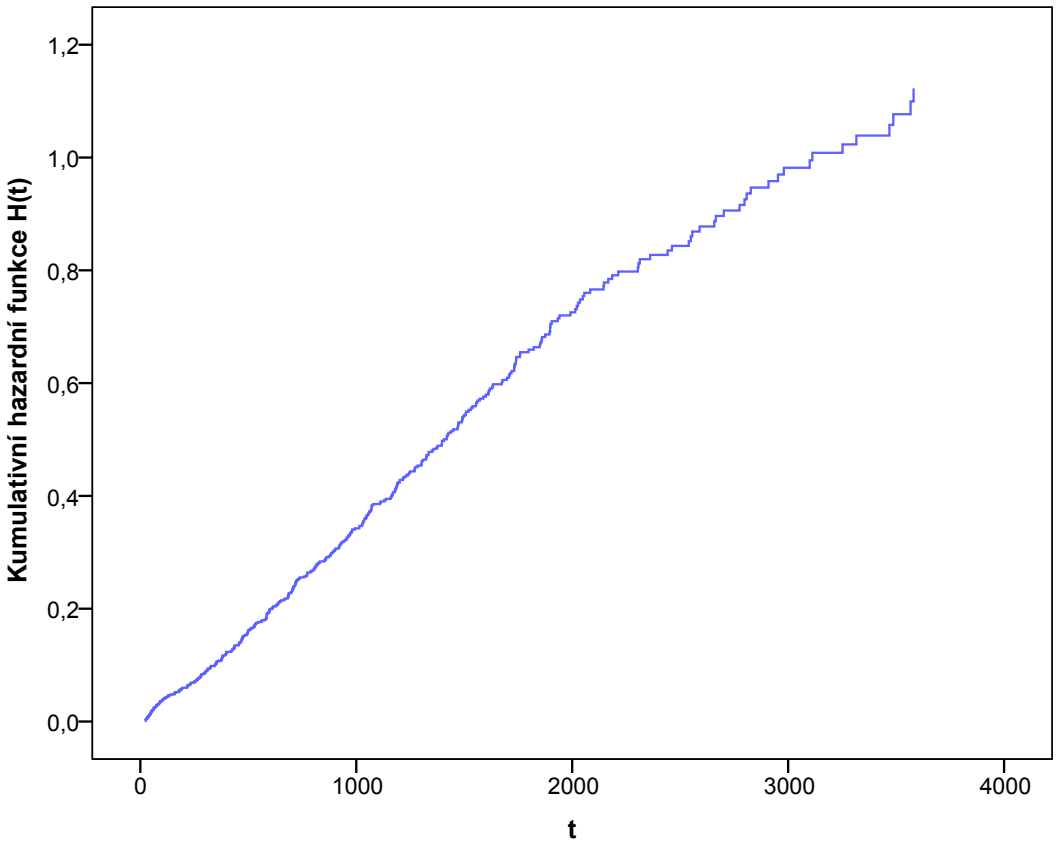
Na obrázku 1 vidíme odhad základní kumulativní hazardní funkce. Tato hazardní funkce odpovídá pacientům, jejichž vstupní vysvětlující proměnné jsou naměřeny na nule. Což pro nás nemá zrovna moc interpretační význam (pacient s věkem 0 let apod.). Nicméně statistický software SPSS nám umožňuje získat kumulativní hazardní funkci (obrázek 2) a funkci přežití (obrázek 3) pro pacienty, jejichž vstupní vysvětlující proměnné odpovídají průměrům těchto proměnných získaných z datového souboru – tabulka 5.

Tabulka 5: Průměrné hodnoty vysvětlujících proměnných

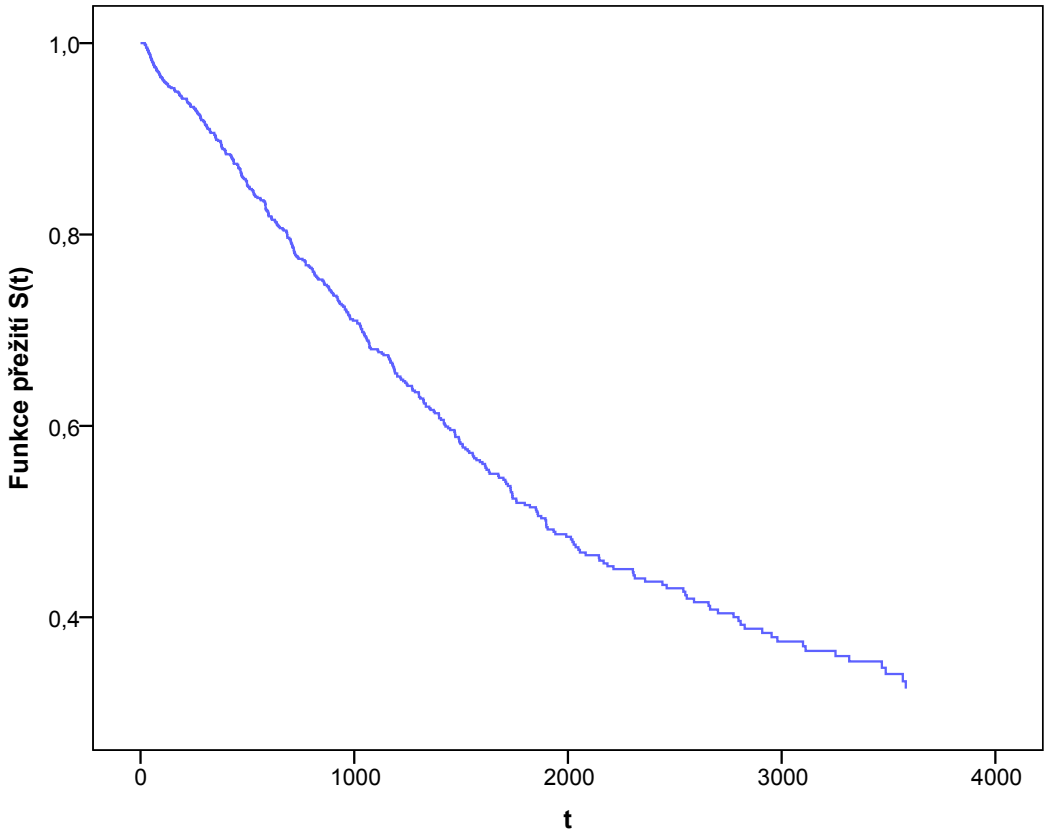
Údaj	ST_1	ST_2	ST_3	*věk*	*arytmie*	$grading_1$	$grading_2$
Průměr	0,284	0,311	0,234	65,403	0,116	0,582	0,102



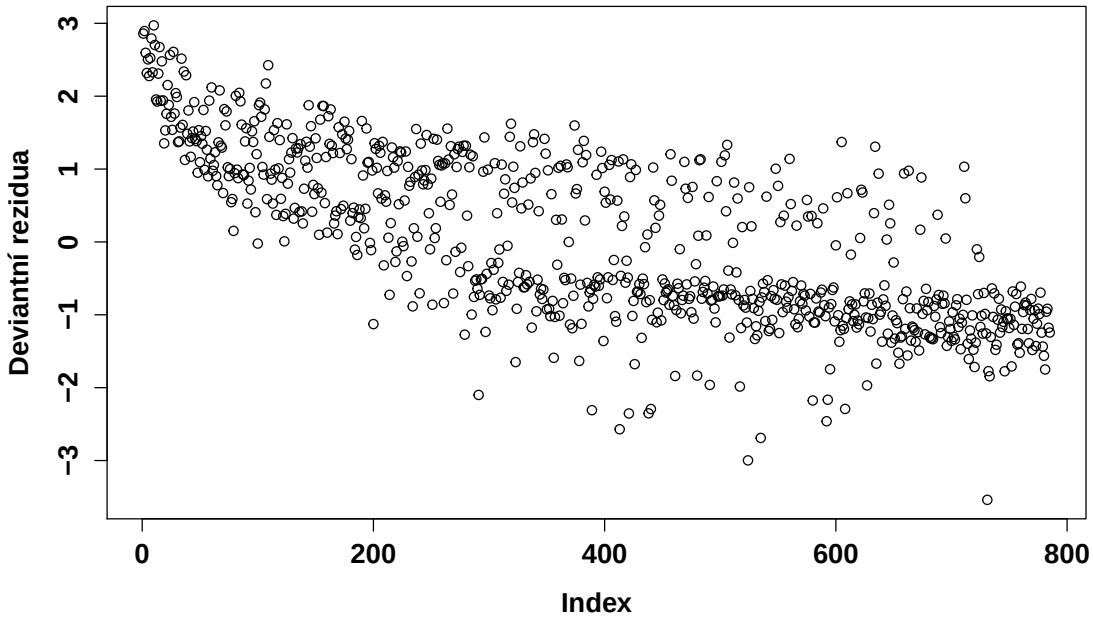
Obrázek 1: Odhad základní kumulativní hazardní funkce $\widehat{H}_0(t)$



Obrázek 2: Kumulativní hazardní funkce vyčíslená v průměrech vysvětlujících proměnných



Obrázek 3: Funkce přežití vyčíslená v průměrech vysvětlujících proměnných



Obrázek 4: Deviantní rezidua

Z tabulky 4 můžeme vyčíst, že pokud je pacient v posledním stadiu (4. stádium), může být riziko pooperačních komplikací u tohoto pacienta až třináctkrát vyšší než u pacienta v prvním stádiu (nejméně závažné stádium). Podobně pokud je u pacienta detekována arytmie, může být jeho riziko pooperačních komplikací až 1,6krát vyšší než u pacienta, který arytmii nemá.

Na obrázku 4 můžeme vidět vizuální kontrolu adekvátnosti modelu pomocí deviantních reziduí. U pacientů blízkých hodnotě 3 predikujeme delší dobu do pooperačních komplikací než ve skutečnosti byla a u pacientů blízkých hodnotě -3 naopak kratší dobu do pooperačních komplikací. Nicméně zdá se, že model pracuje dobře.

5. Závěr

Pomocí Coxova proporcionálního hazardního modelu jsem vyhodnotila riziko pooperačních komplikací u pacientů po operaci kolorekta. Nepotvrdila se mi domněnka lékařů, že by laparoskopická metoda byla méně rizikovější než metoda otevřená. Mým závěrem je, že typ operace nemá vliv na riziko pooperačních komplikací. Nejvíce toto riziko ovlivňuje proměnná stádium, kdy u pacientů v nejhorším stádiu může být riziko až třináckrát větší než u pacientů v prvním stadiu. Použitá vizualizační metoda, která ověřuje adekvátnost modelu, je sice dostačující, nicméně bych se ráda zaměřila i na jiné postupy, jak model zvalidovat. V budoucnu bych ráda vyzkoušela další modely, speciálně modely nelineární, a jejich výsledky mezi sebou porovnala.

Literatura

[1] Briš, R., Litschmannová, M.: *Statistika II.* Ostrava: Vysoká škola báňská – Technická univerzita, 2007, 1 CD-ROM. ISBN 978-80-248-1482-7.

[2] Collett, D.: *Modelling survival data in medical research.* Second edition. Boca Raton, Florida: Chapman, 2003, 391 p. Texts in statistical science. ISBN 15-848-8325-1.

[3] Hosmer, D. W., Lemeshow, S., May, S.: *Applied survival analysis: regression modeling of time-to-event data.* Second edition. Hoboken, New York: Wiley-Interscience, 2008, xiii, 392 p. ISBN 04-717-5499-4.

[4] Kotz, S., Balakrishnan, N., Read, C. B., Vidakovic, B.: *Encyclopedia of statistical sciences.* Second edition. Hoboken, New York: Wiley-Interscience, 2006, 16 vol. ISBN 047174402616.

[5] Miklová, Ž.: PH model v analýze přežití. VŠB-TU Ostrava, 2014. Diplomová práce. Vysoká škola báňská – Technická univerzita Ostrava. Vedoucí práce Prof. Ing. Radim Briš, CSc.

[6] Volf, P.: *Regresní modely v analýze přežití.* Jednota českých matematiků a fyziků. Praha, 1992 [cit. 2014-03-06].

GRANGEROVA KAUZALITA TROCHU INAK
GRANGER CAUSALITY FROM A DIFFERENT VIEWPOINT

Mária Bohdalová[1], Martin Kalina[2], Oľga Nánásiová[3]

Adresa: [1]Fakulta managementu, Univerzita Komenského, Odbojárov 10, 820 05 Bratislava
[2]SvF, Slovenská technická univerzita, Radlinského 11, 810 05 Bratislava
[3]FEI, Slovenská technická univerzita, Ilkovičova 3, 812 19 Bratislava

E-mail: maria.bohdalova@fm.uniba.sk, kalina@math.sk, olga.nanasiova@stuba.sk

Abstrakt: V mnohých vedeckých článkoch sa vedú diskusie o kauzálnych zákonoch, procesoch, vzťahoch, ... Pojem stochastickej kauzality nie je v nich jednoznačne zavedený. Základný prístup ku kauzalite nám ponúka regresná analýza. Je známe, že odhadnuté koeficienty regresnej závislosti náhodnej premennej Y od náhodnej premennej X nie sú vo všeobecnosti algebricky inverzné ku koeficientom odhadnutým z regresnej závislosti náhodnej premennej X od Y. Hoci regresný prístup vyjadruje vo svojej podstate kauzálny vzťah, dáva nám 2 rôzne miery symetrickej, resp. nekauzálnej korelácie. Granger využil regresný model a zaviedol pre analýzu ekonomických a finančných časových radov najznámejší prístup známy pod názvom Grangerova kauzalita. Grangerova kauzalita umožňuje modelovať nesymetrickú závislosť medzi dvoma stochastickými procesmi. V našom príspevku sa zameriame na modelovanie stochastickej kauzality na špeciálnych systémoch boolovych algebier (horizontálnych súčtoch booleovych algebier). Náš kauzálny model umožňuje skonštruovať také asymetrické rozdelenie, pre ktoré z $P(B|A) > P(B)$ nevyplýva $P(A|B) > P(A)$. Tiež poukazujeme na súvis takto modelovanej kauzality s Grangerovou kauzalitou.

Kľúčová slová: Grangerova kauzalita, kompatibilita, združené rozdelenie.

Abstract: Authors in many scientific papers discuss on causal laws, causal processes, causal relationships, etc. However, the notion of stochastic causality is not uniquely understood. Classical approaches to causality are related to regression analysis. It is well known that the regression of Y on X produces coefficient estimates that are not the algebraic inverses of those produced from the regression of X on Y. Although regressions may have a natural causal direction, there is nothing in the data on their own that reveals which direction is the correct one – each of them is an equally appropriate rescaling of a symmetrical and non-causal correlation. Within the existing approaches the best known is that by Granger which is used mainly in econometrics. In this paper we present a new theoretical model of stochastic causality on special systems of Boolean algebras (horizontal sums of Boolean algebras). Our model of causality enables to construct asymmetric distributions such that $P(B|A) > P(B)$ does not imply $P(A|B) > P(A)$. We show

a connection between the presented model of stochastic causality and the Granger's one.

Keywords: Granger causality, compatibility, joint distribution.

1. Introduction

In the 20th century causal inference was frequently associated with multiple correlation and regression. As it is well known, the regression of Y on X produces coefficient estimates that are not the algebraic inverses of those produced from the regression of X on Y. Although regression models may have a natural causal direction, there is nothing in the data on their own that reveals which direction is the correct one – each of them is an equally appropriate rescaling of a symmetrical and non-causal correlation. This problem is known as the problem of *observational equivalence*.

Clive W. J. Granger [2] introduced a data-based concept that is an example of the modern probabilistic approach to causality. This approach identifies cause with a factor that raises the probability of the effect: A causes B if $P(B|A) > P(B)$. The asymmetry of causality is secured by requiring the cause (A) to occur before the effect (B), see [4]. But the probability criterion is not enough on its own to produce asymmetry since $P(B|A) > P(B)$ implies $P(A|B) > P(A)$.

This paper is organized as follows. Next section introduces mathematical model of causality. A model of Granger causality on a system of Boolean algebras is described in Section 3. Conclusion concludes our findings.

2. Mathematical model of causality

There exists no unique mathematical definition of causality. Loosely understood, causality is a relationship between a cause and its effect. In statistics, the notion of causality is usually identified with a kind of stochastic dependence. Of course, this dependence (e.g. between two random variables) is a symmetric notion. In [2] Granger defined a causality between two stationary time-series $\mathbf{X} = \{X_t\}_{t\in\mathbb{Z}}$ and $\mathbf{Y} = \{Y_t\}_{t\in\mathbb{Z}}$ in a non-symmetric way. There are two basic principles upon which this notion of causality (and a relationship between a cause and its effect) is based.

- Cause always happens prior to its effect.
- Cause makes unique changes in the effect. In other words, the causal series contains unique information about the effect series that is not available otherwise.

The precise definition of Granger's causality is the following:

Definition 2.1 ([2]). *Denote by U_{t-1} all information of the universe that is accumulated by time $t-1$ and $\overline{X}_{t-1} = \{X_{t-i}\}_{i=1}^{\infty}$ (i.e., the past of $\mathbf{X}$). Let $\widetilde{U}_{\overline{X},t-1}$ denote all the information of the universe that is accumulated by time $t-1$ apart from the*

past $\overline{X}_{t-1}$. *Then if* $\sigma^2(Y_t|U_{t-1}) < \sigma^2(Y_t|\widetilde{U}_{\overline{X},t-1})$, *we say that* $\mathbf{X}$ *is causing* $\mathbf{Y}$, *denoted by* $\mathbf{X} \Rightarrow \mathbf{Y}$.

Remark. **(1)** Since we assume stationarity of the time series $\mathbf{X}$ and $\mathbf{Y}$ the condition $\sigma^2(Y_t|U_{t-1}) < \sigma^2(Y_t|\widetilde{U}_{\overline{X},t-1})$ is fulfilled for all t whenever it is fulfilled for one t. This means that the past of the time series $\mathbf{X}$ influences the future of $\mathbf{Y}$ (in Definition 2.1 expressed by means of the conditional variance).

(2) As Granger pointed out in [2], we could skip the condition that the time series $\mathbf{X}$ and $\mathbf{Y}$ are stationary, but then the causality between $\mathbf{X}$ and $\mathbf{Y}$ would become time-dependent, i.e., it might occur that at some time stamps t_i we would have $X_{t_i} \Rightarrow Y_{t_i}$ and at some other time stamps t_j X_{t_j} would not cause Y_{t_j}.

3. A model of Granger causality on a system of Boolean algebras

Before turning our attention to the Granger causality, we should say something on random vectors and stochastic processes as a generalization of random vectors.

3.1. Random vectors versus vectors of observables

We will deal with a measurable space $(\Omega, \mathcal{S})$ where $\mathcal{S}$ is a σ-algebra of measurable events. Denote $\mathcal{B}$ the σ-algebra of Borel subsets of $\mathbb{R}$. A *random variable* $\xi : \Omega \to \mathbb{R}$ is an $\mathcal{S}$-measurable function, i.e., $\xi^{-1}(B) \in \mathcal{S}$ for every $B \in \mathcal{B}$. The function ξ^{-1} is called an *observable*.

Further, let $\mathcal{B}^2$ and $\mathcal{S}^2$ denote the direct products $\mathcal{B} \times \mathcal{B} = \{A \times B; A, B \in \mathcal{B}\}$ and $\mathcal{S} \times \mathcal{S} = \{S_1 \times S_2; S_1, S_2 \in \mathcal{B}\}$, respectively. By $\sigma(\mathcal{B}^2)$ and $\sigma(\mathcal{S}^2)$ we will denote the least set σ-algebra containing the corresponding direct products.

Let ξ and η be $\mathcal{S}$-measurable functions. In the Kolmogorovian probability theory the random vector (ξ, η) is modelled as a bivariate function such that for every $B \in \sigma(\mathcal{B}^2)$ holds $(\xi, \eta)^{-1}(B) \in \sigma(\mathcal{S}^2)$. This model works perfectly if ξ and η are measurable simultaneously, e.g., two parameters measured on the same objects. But also in this case we are usually interested in knowing probabilities for $P(\xi^{-1}(A), \eta^{-1}(B))$ where $A, B \in \mathcal{B}$. This means that instead of constructing $\sigma(\mathcal{B}^2)$ and $\sigma(\mathcal{S}^2)$ it is enough (might be up to some exceptions) to work with the corresponding direct products $\mathcal{B}^2$ and $\mathcal{S}^2$. Thus the model becomes slightly different.

A different situation occurs if we consider a random vector (ξ, η), but ξ and η are not simultaneously measurable. Of course one possibility how to model this situation is to stay within the Kolmogorovian model. In this case we know that $P((\xi, \eta)^{-1} \in A \times B) = P((\eta, \xi)^{-1} \in B \times A)$, where $A, B \in \mathcal{B}$. But there exists another algebraic model which suits better for modelling this situation. We explain this algebraic model by means the following example.

Example 3.1. Let us consider a four-element set $\Omega = \{a, b, c, d\}$. We can make two different Boolean algebras of subsets of Ω, $\mathcal{S}_1 = \{\emptyset, \{a, b\}, \{c, d\}, \Omega\}$ and $\mathcal{S}_2 =$

$\{\emptyset, \{a, c\}, \{b, d\}, \Omega\}$. Denote $\mathcal{L} = \mathcal{S}_1 \cup \mathcal{S}_2$. Then $\mathcal{L}$ is called the *horizontal sum* of $\mathcal{S}_1$ and $\mathcal{S}_2$ (for more information on horizontal sums, but also on more general structures, see, e.g., [1, 9]). On $\mathcal{L}$ it is not possible to define random variables, we are able to define observables. Elements of the same Boolean algebra (sub-algebra of $\mathcal{L}$) are called *compatible*, and elements which do not belong to the same Boolean algebra are called *non-compatible*. E.g., $\{a, b\}$ and $\{a, c\}$ are non-compatible.

Of course, we can define infinitely many σ-additive functions on $\mathcal{L}$, i.e., probability measures. Let us now show that it is possible to define also bivariate σ-additive functions on $\mathcal{L}^2$, i.e., joint distributions (called also s-maps). We present one of such joint distributions in Table 1.

Table 1: S-map constructed on $\mathcal{L}^2$

s-map p	$\{a, b\}$	$\{c, d\}$	$\{a, c\}$	$\{b, d\}$	Ω
$\{a, b\}$	0.3	0	0.2	0.1	0.3
$\{c, d\}$	0	0.7	0.3	0.4	0.7
$\{a, c\}$	0.15	0.35	0.5	0	0.5
$\{b, d\}$	0.15	0.35	0	0.5	0.5
Ω	0.3	0.7	0.5	0.5	1

Let us mention that the s-map costructed in Table 1 is not symmetric, e.g., $p(\{a, b\}, \{a, c\}) = 0.2$, but $p(\{a, c\}, \{a, b\}) = 0.15 = p(\{a, c\}) \cdot p(\{a, b\})$. This means that $\{a, b\}$ depends on $\{a, c\}$, but $\{a, c\}$ is independent of $\{a, b\}$. Such situation cannot be modelled within the Kolmogorovian model. For more information on properties of s-maps see, e.g., [5, 6, 7, 8].

Let us now continue in our considerations. We assume that we have two random variables, ξ and η, but these variables are not measurable simultaneously. As the algebraic model for this situation we will use the horizontal sum of Boolean algebras. Then, instead of random variables ξ and η we must use observables ξ^{-1} and η^{-1}. The fact that observables ξ^{-1} and η^{-1} are not simultaneously measurable, can be interpreted as their non-compatibility. We have seen in Example 3.1 that unlike the probability measure, s-maps are not necessarily symmetric. This means, if we denote $a = \xi^{-1}(A)$ and $b = \eta^{-1}(B)$, we might get $p(a, b) \neq p(b, a)$.

We will use the horizontal sum of Boolean algebras $\mathcal{S}_1$ and $\mathcal{S}_2$ which we denote by $\mathcal{L}$. In such a way for arbitrary $A, B \in \mathcal{B}$ we have $(\xi^{-1}(A), \eta^{-1}(B)) \in \mathcal{L} \times \mathcal{L}$ and $(\eta^{-1}(B), \xi^{-1}(A)) \in \mathcal{L} \times \mathcal{L}$ and we have one s-map p modelling the (possibly non-symmetric) distribution of both vectors of observables, (ξ^{-1}, η^{-1}) and (η^{-1}, ξ^{-1}).

Remark (Interpretation of the non-symmetric distribution). We design two different experiments. In experiment Nr. 1 we measure first a parameter corresponding to ξ^{-1} and then η^{-1}. In the second experiment we change the order of ξ^{-1} and η^{-1}. We

admit that the relative frequencies of $(\xi^{-1}, \eta^{-1}) \in A \times B$ and that of $(\eta^{-1}, \xi^{-1}) \in B \times A$, might be different (order-dependent).

3.2. Modelling of Granger causality

Assume that $\{\mathbb{X}_t\}_{t\in T}$ is a stochastic process. For every time-stamp $t \in T$ X_t is an $\mathcal{S}_t$-measurable random variable where $\mathcal{S}_t$ is a Boolean σ-algebra. If we want to model causality (in the sense of non-symmetric dependence), we have to make the same procedure as above (with random vectors) when we have abandoned Boolean algebras and considered their horizontal sums instead.

We will consider a family $\{\mathcal{S}_t\}_{t\in T}$ of σ-algebras and we construct their horizontal sum which we denote by $\widehat{\mathcal{S}}$. For every time-stamp $t \in T$ and every Borel set $A \in \mathcal{B}$ we will have $X_t^{-1}(A) \in \widehat{\mathcal{S}}$. Then, for $s \neq t$, X_t^{-1} and X_s^{-1} are non-compatible observables. We know already that there exists joint distribution of X_t^{-1} and X_s^{-1} (or equivalently, conditional distribution $f(X_s^{-1}|X_t^{-1})$ which is interesting especially for $s > t$).

Granger causality. Assume that we have two (not necessarily stationary) stochastic processes, $\{\mathbf{X}_t\}_{t\in T}$ and $\{\mathbf{Y}_t\}_{t\in T}$, where T is a set of all possible time-stamps. According to Definitions 2 and 5 in [3, pp. 336–337], $\{\mathbf{Y}_t\}_{t\in T}$ causes $\{\mathbf{X}_t\}_{t\in T}$ if $F(X_{t+1}|Y_t) \neq F(X_{t+1})$, where $F(\cdot|\cdot)$ is a conditional distribution function $F(\cdot)$ is an unconditioned distribution function.

To model causality between stochastic processes $\{\mathbf{X}_t\}_{t\in T}$ and $\{\mathbf{Y}_t\}_{t\in T}$, we need to have an equivalent of a measurable space such that for every $t, s \in T$ observables X_t^{-1} and Y_s^{-1} are non-compatible. This means that we need two copies of the above mentioned horizontal sum $\widehat{\mathcal{S}}$ (and make their horizontal sum). We will denote this space by $\widehat{\mathcal{S}}_2$. Thus we get that $F_{(X_t^{-1}, Y_t^{-1})}$ and $F_{(Y_t^{-1}, X_t^{-1})}$ may be different functions.

In this experiment we are not able to distinguish the order (X_t^{-1}, Y_t^{-1}) and (Y_t^{-1}, X_t^{-1}) if we measure X and Y at the same time stamp. This means that, as we have already commented in Remark on this page, measuring the non-symmetric causality experimentally has to follow exactly what Granger proposed in [2, 3].

4. Conclusion

We have introduced a new (theoretical) model of causality that enables to construct asymmetric distributions such that $P(B|A) > P(B)$ does not imply $P(A|B) > P(A)$. This model can serve as a theoretical background for modelling of time series (where we have natural occurrence of causality).

Acknowledgements

The authors kindly announce the support of the Science and Technology Assistance Agency under the contract No. APVV-14-0013, and of the VEGA grant agency, grant numbers VEGA 2/0059/12 and VEGA 1/0710/15.

References

[1] Dvurečenskij, A., Pulmannová, S.: *New Trends in Quantum Structures*. Kluwer, Dodrecht, 2000.

[2] Granger, C. W. J.: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37(3), pp. 424–438, 1969.

[3] Granger, C. W. J.: Testing for causality. A personal viewpoint. *Journal of economic Dynamics and Control* 2, pp. 329–352, 1980.

[4] Hoover, K. D.: *Causality in economics and econometrics. The New Palgrave Dictionary of Economics.* Second edition. Steven N. Durlauf and Lawrence E. Blume (eds.), Palgrave Macmillan, 2008. The New Palgrave Dictionary of Economics Online. doi:10.1057/9780230226203.0209 [cit. 13. 9. 2014]. URL: `http://www.dictionaryofeconomics.com/article?id=pde2008_C000569`

[5] Nánásiová, O.: Map for simultanous measuremants for a quantum logic. *International Journal of Theoretical Physics* 42, pp. 1889–1903, 2003.

[6] Nánásiová, O.: Principle conditioning. *International Journal of Theoretical Physics* 43, pp. 1757–1767, 2004.

[7] Nánásiová, O., Kalina, M.: Calculus for non-compatible observables, construction through conditional states. *International Journal of Theoretical Physics* 54, pp. 506–518, 2015.

[8] Nánásiová, O., Pulmannová, S.: S-maps and tracial states. *Information Sciences* 179, pp. 515–520, 2009.

[9] Pták, P., Pulmannová, S.: *Quantum Logics.* Kluwer Academic Press, Bratislava, 1991.

URČENIE POTENCIÁLU KLIENTSKEHO PORTFÓLIA

DETERMINING THE POTENTIAL OF CLIENT PORTFOLIO

Marianna Godišová[1], **Iveta Stankovičová**[2]

Adresa: [1]Fakulta matematiky, fyziky a informatiky, Univerzita Komenského v Bratislave, Mlynská dolina, 842 48 Bratislava, Slovenská republika
[2]Fakulta managementu, Univerzita Komenského v Bratislave, Odbojárov 10, 820 05 Bratislava 25, Slovenská republika

E-mail: godisova.m@gmail.com, iveta.stankovicova@fm.uniba.sk

Abstrakt: Príspevok sa zaoberá problematikou identifikácie najhodnotnejších zákazníkov komerčnej poisťovne. Samotná orientácia na zákazníka na súčasnom trhu nepredstavuje konkurenčnú výhodu. Je síce základom a dôležitosť zákazníka uznávajú všetky inštitúcie pôsobiace v podnikateľskej sfére, ale často nestačí. Dnes sa hlavne treba venovať „tým správnym“ zákazníkom. Cieľom príspevku bude s využitím kvantitatívnych metód identifikovať klientov s najvyššou celoživotnou hodnotou, ktorí prinášajú poisťovni najvyšší profit. V prostredí SAS Enterprise Miner bol vytvorený predikčný model na dátach z oblasti poisťovníctva. Výstupom je skórovacia karta, ktorá obsahuje významné činitele (premenné) vo vzťahu k modelovanej premennej. Poisťovňa môže informácie získané zo skórovacej karty využiť na prijímanie dôležitých rozhodnutí týkajúcich sa marketingových aktivít smerom k potenciálne najlepším klientom. Význam a dôležitosť zamerania sa na túto skupinu klientov možno podporiť aj zistením z našej práce, že očakávaný výnos plynúci z „prémiových“ klientov tvorí značnú časť celkového očakávaného výnosu. V prípade nášho modelu, asi 15 % klientov vytvorí viac ako 40 % celkového očakávaného zisku poisťovne.

Kľúčová slová: riadenie vzťahov so zákazníkmi, celoživotná hodnota zákazníka, credit scoring, predikčný model, skórovacia karta.

Abstract: The paper deals with the topic of identifying the most valuable customers of a commercial insurance company. The actual orientation toward the customer in the current market does not constitute a competitive advantage. Even though it is the basis and the importance of the customer is recognized by all institutions involved in the business sphere, this is often not enough. Today special attention should be given to „the right“ customers. The aim of the contribution will be to identify clients with the highest lifetime

value using quantitative methods which bring the highest profit to the insurance company. In the SAS Enterprise Miner environment a prediction model based on insurance data was created. The output is a scorecard containing important factors (variables) in relation to the modelled variable. The insurance company may use the information obtained from the scorecard to make critical decisions on marketing activities directed towards the potential best customers. The significance and importance of focusing on this group of customers can be supported by our findings that the expected revenue derived from the „premium“ rating customers constitutes a significant part of the total expected profit. In our case, approximately 15 % of clients constituted more than 40 % of the total expected profit for the insurance company.

Keywords: customer relationship management, scorecard, customer lifetime value, credit scoring, predictive model.

1. Úvod

Vďaka neustálemu rozvoju metód riadenia vzťahov so zákazníkmi a možnosti využívať rôzne technologické novinky sa organizácie dnes čoraz častejšie zamýšľajú nad tým, ako zvýšiť svoj zákaznícky podiel a rozšíriť svoje portfólio o „kvalitných“ zákazníkov. Snažia sa o to, aby vo svojom klientskom portfóliu mali zákazníkov, ktorí organizácii prinesú zisk a zároveň s ňou zotrvajú v dlhodobom vzťahu.

Problematika určenia budúcej hodnoty klienta pre podnik je v odborných kruhoch diskutovanou témou. Organizácie sa pomocou rôznych metód, techník a softvérových riešení snažia nájsť spôsob, ako „oceniť“ už existujúceho, ale aj potenciálneho klienta. Snažia sa predikovať, aký zisk zákazník prinesie a aké budú náklady na jeho obsluhu. Modely slúžiace na odhad budúcej hodnoty klienta sú pre podnik časovo, finančne a aj dátovo náročné. V praxi je dôležité vedieť určiť celoživotnú hodnotu zákazníka (z angl. *CLV = Customer Lifetime Value*). Podstata celoživotnej hodnoty zákazníka tkvie v tom, že zákazník nepredstavuje jednorazovú transakciu. Reprezentuje vzťah, ktorý je oveľa cennejší ako akákoľvek jednorazová výmena tovarov a finančných prostriedkov medzi podnikom a klientom [8].

Dôležitosť kalkulácie a odhadu *CLV* vystúpila do popredia kvôli niekoľkým faktorom. Štúdie ukázali, že ziskovosť plynúca z jednotlivých zákazníkov nie je rovnaká a tak ani hodnota každého zákazníka pre podnik nie je rovnaká. Podľa Paretovho pravidla 80 % zisku tvorí iba 20 % zákazníkov. V modernom marketingu sa objavuje aj upravené Paretovo pravidlo, ktoré namiesto známeho pomeru 80 : 20 poukazuje na pomer 80 : 20 : 30. Znamená to, že 20 % najlepších klientov tvorí 80 % zisku, pričom 30 % klientov z portfólia je stra-

tových [4]. Na základe tohto pravidla (či už pôvodného, alebo jeho upravenej verzie) je zjavné, že je potrebné venovať sa dvadsiatim percentám zákazníkov, ktoré sú relevantní [1].

Cieľom príspevku je odhaliť a poukázať na faktory, ktoré popisujú existujúceho klienta komerčnej poisťovne s potenciálom patriť k jej „najhodnotnejším“ klientom s využitím metód hĺbkovej analýzy údajov, štatistických a modelovacích metód a metód oceňovania.

2. Dáta a použitá metodológia

Mali sme k dispozícii dátový súbor o rozsahu 7 500 klientov komerčnej poisťovne. Každý klient v súbore bol charakterizovaný množstvom premenných (spolu 90 premenných), ktoré je možné rozdeliť do nasledovných skupín:

1. premenné o výške aktuálnej sumy poistného pre jednotlivé druhy poistenia,
2. premenné o počte uzavretých a zrušených poistných zmlúv pre jednotlivé druhy poistenia a pripoistenia,
3. premenné o počte výskytu poistných udalostí,
4. premenné o dĺžke vzťahu medzi klientom a poisťovňou,
5. premenné o postavení klienta v poisťovni, jeho vzťahu s poisťovňou a profitabilite,
6. premenné o kontaktných kanáloch,
7. demografické premenné.

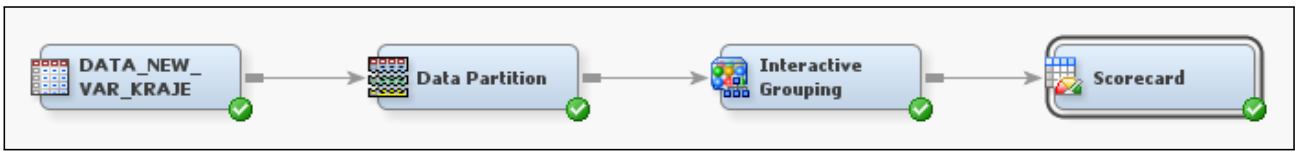
Z dôvodu požiadavky na vytvorenie modelu, ktorý bude možné jednoducho interpretovať, zvolili sme metódu *credit scoring*, ktorá predstavuje skupinu prediktívnych modelov na riadenie rôznych druhov rizík, ktoré sa používajú hlavne vo finančných organizáciách [6]. Prediktívne modely je možné vytvoriť rôznymi modelovacími metódami. Medzi najpoužívanejšie patria logistická regresia, rozhodovacie stromy, neurónové siete a lineárna diskriminačná analýza.

Prediktívny credit scoringový model umožňuje nájsť najvýznamnejšie vysvetľujúce premenné (faktory) vzhľadom k vysvetľovanej (modelovanej, cieľovej) premennej. Táto metóda patrí medzi najúspešnejšie aplikácie štatistických techník vo finančnom sektore a patrí medzi najstaršie aplikácie data miningu [5]. Je založená na skúmaní a analyzovaní historických údajov o správaní sa klientov, teda o ich nákupnej histórii. Údaje sa v procese data miningu rozdelia na niekoľko dátových množín, aby sa dala vyhodnotiť kvalita modelovania. Dáta zvyčajne rozdeľujeme na tréningovú, validačnú a prípadne testovaciu množinu, napríklad v pomere $40 : 40 : 20$.

Najpoužívanejšou formou výstupu z credit scoringových modelov je skórovacia karta (angl. *scorecard*). Ide o štatistický model v špeciálnom formáte využívaný najmä kvôli jednoduchosti interpretácie. Skórovacia karta umožňuje uskutočniť distribúciu bodov pre klientov na základe hodnôt signifikantných premenných vzhľadom ku modelovanej premennej.

V našom dátovom súbore sa priamo nenachádzala vhodná závislá premenná. Vzhľadom na cieľ, sme ju odvodili od spojitej premennej „PROFITABILITA“. Táto premenná v dátovom súbore predstavovala výnosy z produktov daného klienta očistené o náklady poisťovne na tohto klienta. Použili sme Paretovo pravidlo 80 : 20. Hraničná hodnota profitability, ktorá rozdelila klientov na „prémiových“ a „ostatných“ bola definovaná tak, aby „prémioví“ klienti tvorili 20 % z celkového počtu klientov (7 500 klientov). Predstavovalo to 1 500 klientov s najvyššou profitabilitou. Pre týchto klientov cieľová premenná (CLV) nadobúdala hodnotu 1 a u ostatných klientov hodnotu 0.

Na modelovanie sme použili data miningový softvér SAS Enterprise Miner (ďalej SAS EM), konkrétne 4 uzly špeciálne vyvinuté pre credit scoring: *Interactive Grouping, Scorecard, Reject Inference, Credit Exchange*. Využili sme funkcionalitu hlavne prvých dvoch uzlov. Uzol *Interactive Grouping* umožňuje vykonať kategorizáciu kvantitatívnych premenných do intervalov a tiež zoskupiť hodnoty kvalitatívnych premenných tak, aby čo najlepšie predikovali závislú premennú. Uzol *Scorecard* potom umožní vytvoriť skórovaciu kartu na základe modelu logistickej regresie (obrázok 1).



Obrázok 1: Proces tvorby prediktívneho modelu v SAS EM

Binárna logistická regresia predikuje podmienenú pravdepodobnosť p výskytu želanej udalosti ($Y = 1$) v závislosti od vysvetľujúcich premenných (X_i) kategoriálneho aj spojitého typu, čiže $p = P(Y = 1/X_i)$. Tento vzťah je však nelineárny, preto sa používa tzv. logitová transformácia podmienenej pravdepodobnosti p [8]:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k.$$

Základom tejto transformácie je prirodzený logaritmus zo šance, čo je podiel $p/(1-p)$. Parametre β_i sa odhadujú metódou maximálnej vierohodnosti.

Spätnou transformáciou získame želanú podmienenú pravdepodobnosť p:

$$p = \frac{\mathrm{e}^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k)}}{1+\mathrm{e}^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k)}} = \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k)}}.$$

V procese data miningu príprava údajov na modelovanie predstavuje 60 – 80 % z celkového času. Aj v našom prípade to tak bolo. Museli sme z pôvodných premenných odvodiť nové premenné vhodnejšie na modelovanie (napr. z dátumu narodenia klienta odvodiť jeho vek, resp. vhodne vek kategorizovať, z adresy bydliska odvodiť premenné o kraji, resp. okrese). Tiež sme museli nahradiť chýbajúce údaje vhodnými metódami, alebo vytvoriť rôzne indikátorové premenné. Mnohé premenné sme museli transformovať, aby sme odstránili ich zošikmenie a podobne.

Logistická regresia tiež vyžaduje nekorelované vstupy, preto sme premenné do modelu vyberali pomocou zhlukovej analýzy, ktorej podstatou je rozklad súboru objektov na niekoľko relatívne rovnorodých podmnožín tak, aby objekty patriace do rôznych zhlukov si boli čo najmenej podobné a objekty patriace do toho istého zhluku si boli podobné čo najviac [8]. Do modelu vstupovali len také vysvetľujúce premenné, ktoré reprezentovali jednotlivé zhluky premenných.

Úlohou počiatočnej analýzy bolo aj posúdiť predikčnú silu vysvetľujúcich premenných a eliminovať tie premenné, ktoré sú pre účely modelu nelogické, alebo sú málo významné. V SAS EM na to slúžia dve miery a to informačná hodnota (IV) a štatistika *WOE*. Sú to štatistiky, ktoré sa veľmi často používajú pri vývoji skórovacích kariet.

Štatistika *WOE* (z angl. *Weight of Evidence*) je mierou predikčnej sily jednotlivých kategórií (hodnôt) vstupnej premennej v schopnosti odlíšiť „prémiových“ klientov od ostatných. Porovnáva pomer „prémiových“ klientov a ostatných klientov pre každú kategóriu hodnôt skúmanej premennej. Na výpočet sa používa nasledovný vzorec:

$$WOE = \ln\left(\frac{Distr\ not\ CLV_i}{Distr\ CLV_i}\right),$$

kde $Distr\ CLV_i$ znamená distribúciu prémiových klientov pre i-tu kategóriu, teda pomer počtu „prémiových“ klientov v kategórii i k celkovému počtu prémiových klientov a $Distr\ not\ CLV_i$ je distribúcia „neprémiových“ klientov pre kategóriu i, teda pomer počtu neprémiových klientov v kategórii i k celkovému počtu neprémiových klientov. Počet kategórií označme ako m, potom $i = 1, 2, \ldots, m$. Priebeh štatistiky *WOE* by mal byť monotónny vzhľadom k vytvoreným kategóriám, čiže by mal byť buď klesajúci alebo rastúci.

Predikčná sila jednotlivých vstupných premenných sa meria pomocou štatistiky informačná hodnota IV (z angl. *Information Value*), ktorá je váženým súčtom štatistiky WOE cez všetky kategórie vstupnej premennej:

$$IV = \sum_{i=1}^{m} \left[(Distr\ not\ CLV_i - Distr\ CLV_i) \cdot \ln \left(\frac{Distr\ not\ CLV_i}{Distr\ CLV_i} \right) \right].$$

Do modelu zvyčajne nevstupujú premenné, pre ktoré nadobúda štatistika IV hodnoty nižšie ako 0,1. Premenné, ktoré dosahujú hodnoty IV vyššie ako 0,5, sú podozrivé a vyžadujú ďalšiu analýzu [3, 7].

V prostredí SAS Enterprise Miner sa skórovacia karta vytvára pomocou logistickej regresie v uzle *Scorecard*. Do modelu vstupujú len štatisticky významné premenné so silnou predikčnou schopnosťou. Vo všeobecnosti sa uvádza, že skórovacia karta by mala obsahovať 8–15 premenných z dôvodu zabezpečenia jej stability. Výhodou stabilnej skórovacej karty sú konzistentné výsledky aj pri prípadných zmenách na vstupoch [7].

3. Interpretácia výsledkov

V procese modelovania sme vytvorili niekoľko modelov a porovnali sme ich na základe súboru štatistík kvality modelov, ktoré sú implementované v SAS EM do uzla *Model Comparison*. Hlavným kritériom výberu modelu bola miera chybovosti na validačnej množine (tabuľka 1). Modely s označením SC_1f a SC_2s dosiahli rovnakú hodnotu pre túto štatistiku (13,59 %). Na základe lepšej interpretovateľnosti sme vybrali model SC_2s. Finálny model je tvorený desiatimi premennými, ktoré sú štatisticky významné vzhľadom k modelovanej premennej *CLV*. Ich zoznam usporiadaný zostupne podľa ich významnosti v modeli SC_2s uvádzame v tabuľke 2.

Pre ilustráciu uvádzame, ako sú body distribuované v skórovacej karte pre premennú, ktorá popisuje, či má klient aktuálne uzavretú poistnú zmluvu na poistenie motorového vozidla (tabuľka 3). V prípade, že klient má uzavretý práve tento typ zmluvy (1 – má), priradí sa mu 44 bodov, v opačnom prípade (0 – nemá) dostáva záporný počet bodov. Analogicky je skóre priraďované aj u ostatných významných faktorov. Celá skórovacia karta je uvedená v práci [2].

Na to, aby bol zákazník identifikovaný ako potenciálne „prémiový“, teda existuje u neho pomerne vysoká pravdepodobnosť, že v priebehu svojho vzťahu s poisťovňou bude prinášať vysokú profitabilitu, musí jeho celkové skóre prekročiť stanovenú bodovú hranicu identifikovanú pomocou hodnoty Kolmogorovej-Smirnovej štatistiky. Hraničné skóre predstavuje bod na horizontálnej osi, v ktorom Kolmogorov-Smirnov graf (obrázok 2) nadobúda

Tabuľka 1: Porovnanie mier kvality vytvorených modelov

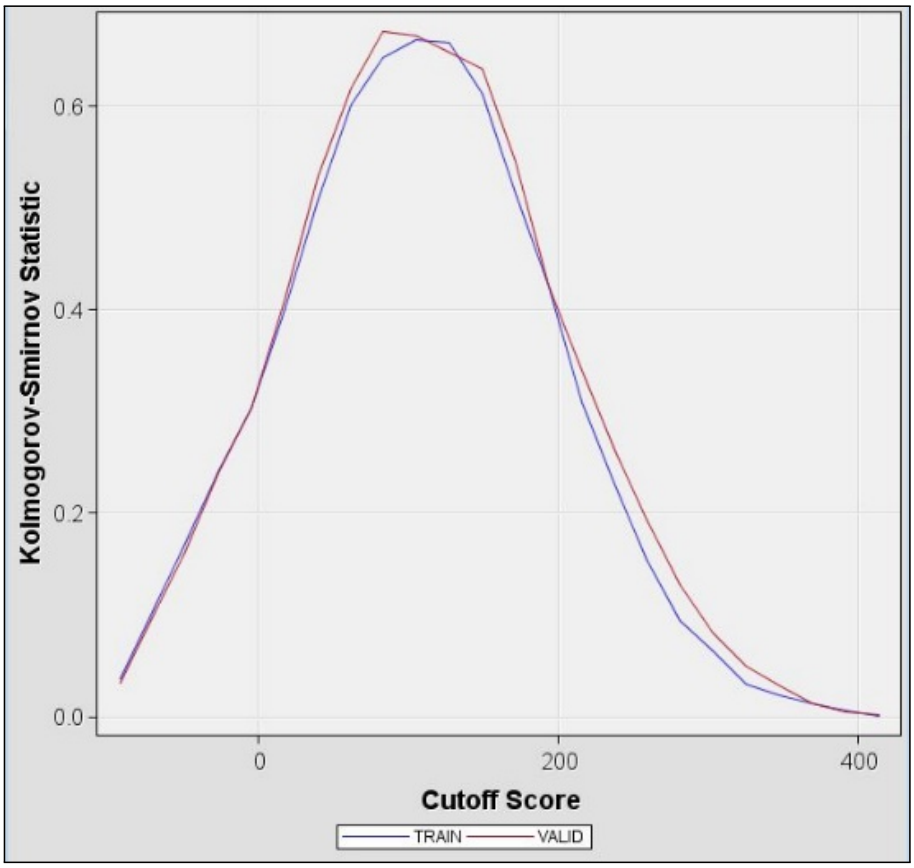
	Model					
Dáta	Štatistika	SC_1f	**SC_2s**	SC_3f	SC_6f	SC_5f
Tréningová množina	AIC	3245,09	3246,04	3357,54	3297,51	3318,02
	BIC	3337,01	3318,26	3462,59	3369,73	3416,50
	Miera chybovosti	0,1366	0,1391	0,1431	0,1391	0,1389
	AUC	0,9060	0,9050	0,8980	0,9010	0,9030
	KS štatistika	0,6660	0,6640	0,6400	0,6590	0,6640
	Kumulatívny lift na 10 %	3,6938	3,6462	3,6462	3,7129	3,6510
Validačná množina	Miera chybovosti	**0,1359**	**0,1359**	0,1394	0,1399	0,1399
	AUC	0,9110	0,9120	0,9040	0,9060	0,9090
	KS štatistika	0,6650	0,6730	0,6420	0,6600	0,6760
	Kumulatívny lift na 10 %	3,8444	3,8444	3,8886	3,8665	3,7561

Tabuľka 2: Zoznam významných premenných v modeli SC_2s zoradených zostupne podľa veľkosti hodnôt štatistiky informačná hodnota (IV)

P. č.	Názov premennej	IV
1	Čas od posledného nákupu	0,911
2	Aktuálna výška poistného	0,909
3	Indikátor vlastníctva produktu – poistenie motorového vozidla	0,857
4	Výška brutto poistného – poistenie majetku	0,616
5	Počet zrušení akejkoľvek poistnej zmluvy	0,611
6	Indikátor, či nastala poistná udalosť – poistenie motorového vozidla	0,510
7	Indikátor, či v prípade poslednej uzavretej poistnej zmluvy išlo o poistenie motorového vozidla	0,385
8	Pohlavie	0,366
9	Výška brutto poistného – kapitálové životné poistenie	0,366
10	Veková kategória	0,332

Tabuľka 3: Distribúcia bodov vybranej premennej v skórovacej karte

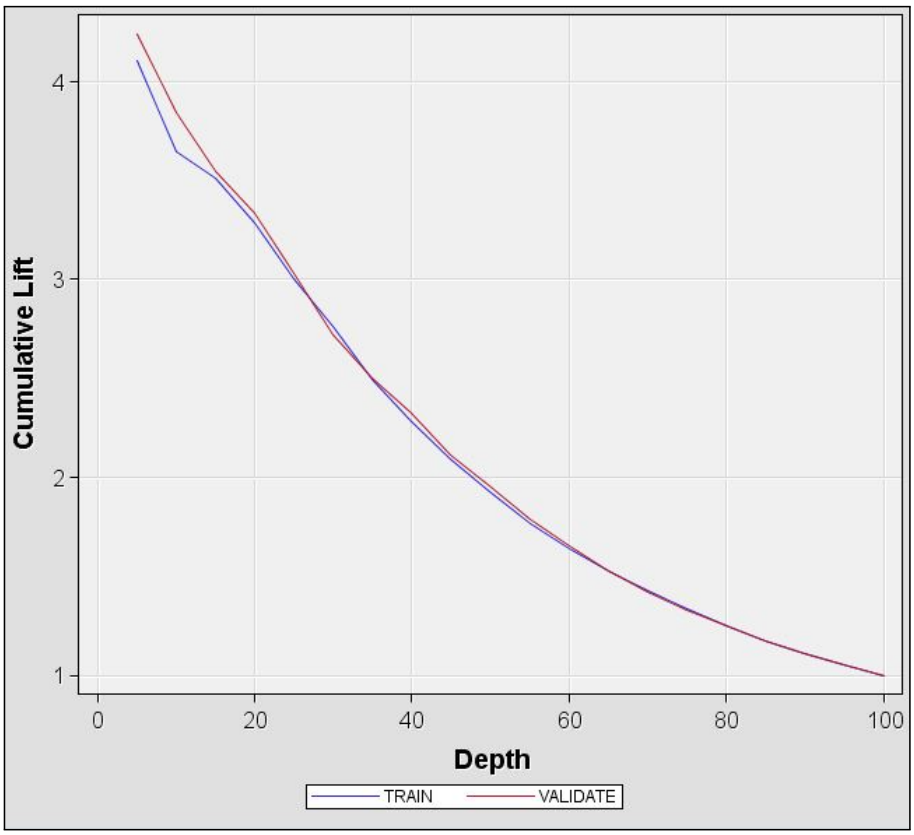
Klient má v súčasnosti uzavreté poistenie motorového vozidla	Body
1	44
0	−5
Chýbajúce alebo neznáme hodnoty	−5



Obrázok 2: Kolmogorov-Smirnov graf predikčného modelu pre *CLV*

svoju maximálnu hodnotu. Hranica, ktorá oddelila „prémiových“ klientov od ostatných je 83 bodov.

Kvalitu vytvoreného modelu možno posúdiť aj pomocou hodnoty kumulatívneho liftu. Priebeh grafu kumulatívneho liftu umožňuje posúdiť kvalitu modelu oproti náhodnému výberu. Na piatom percentile nadobudol kumulatívny lift pre náš model na validačnej množine hodnotu 4,24 (obrázok 3). Čiže pri snahe o identifikáciu 5 % klientov s najvyššiu profitabilitou na zá-



Obrázok 3: Priebeh kumulatívneho liftu predikčného modelu pre *CLV*

klade nášho modelu, by sme mali byť viac ako štvornásobne úspešnejší, ako keby sme hľadali „prémiových“ klientov náhodne.

Implementovaním tohto modelu by poisťovňa bola schopná v existujúcom portfóliu na základe významných premenných ohodnotiť všetkých svojich klientov a rozhodnúť, ktorým z nich má zmysel venovať zvýšenú pozornosť. Model poukazuje na to, ktoré poistné produkty najviac vplývajú na vysokú hodnotu profitability. Zo skórovacej karty vieme určiť, že ide hlavne o poistenie motorového vozidla, kapitálové životné poistenie a poistenie majetku. Toto zistenie dáva poisťovni indíciu, v prípade ktorých produktov má zmysel zamerať sa na podporu predaja, prípadne ich ponúkať ako balíček služieb.

Charakteristika „prémiového“ klienta, resp. segmentu klientov s najvyšším potenciálom je zrejmá zo štatistík v tabuľke 4.

4. Očakávaná profitabilita portfólia klientov

Keďže klienti predstavujú pre poisťovňu aktíva, má zmysel zaoberať sa odhadom objemu finančných prostriedkov, ktoré poisťovni prinesú.

Výstupom modelu logistickej regresie je pravdepodobnosť, že klient bude patriť medzi „prémiových“ klientov. Vychádzajúc z Paretovho pravidla, tak

by 20 % „prémiových“ klientov mohlo poisťovni priniesť až 80 % celkového zisku. Pri výpočte sme vychádzali z očakávaného (stredného) profitu „prémiových“ a ostatných klientov.

Očakávaný zisk (označenie *priem_profit*$(CLV = i)$) pre tieto skupiny uvádzame v tabuľke 5. Očakávaný profit pre každého klienta z portfólia bol vypočítaný podľa vzťahu:

$$\textit{Očakávaný profit} = \sum_{i=0}^{1} P(CLV = i) \cdot \textit{priem\_profit}(CLV = i),$$

Tabuľka 4: Charakteristiky portfólia klientov s najvyšším potenciálom

P. č.	Názov numerickej premennej	Medián
1.	Indikátor vlastníctva produktu – poistenie motorového vozidla	1
2.	Aktuálna výška poistného (ročná v EUR)	295
3.	Čas od posledného nákupu (v rokoch)	3
4.	Počet zrušení akejkoľvek poistnej zmluvy	1
5.	Indikátor, či nastala poistná udalosť – poistenie motorového vozidla	0
6.	Výška brutto poistného – poistenie majetku (ročná v EUR)	414
7.	Výška brutto poistného – kapitálové životné poistenie (ročná v EUR)	1478
8.	Indikátor, či v prípade poslednej uzavretej poistnej zmluvy išlo o poistenie motorového vozidla	0

P. č.	Názov kategoriálnej premennej	Modus
9.	Pohlavie	M
10.	Veková kategória (v rokoch)	38–50

Tabuľka 5: Očakávaný (stredný) profit na 1 klienta

Klienti	Hodnota target premennej	Očakávaný profit
Prémioví	$CLV = 1$	2513,07 €
Ostatní	$CLV = 0$	240,92 €

kde $P(CLV = i)$ je pravdepodobnosť, že cieľová premenná nadobudne hodnotu 0, resp. 1 a *priem_profit*$(CLV = i)$ je priemerná profitabilita pre skupinu klientov, kde má cieľová premenná hodnotu 0, resp. 1 ($i = 0, 1$). Celková profitabilita portfólia je potom sumou profitability všetkých klientov v portfóliu.

Na vzorke tvorenej 2252 zákazníkmi náš model za „prémiových" klientov identifikoval približne 15 % vzorky. Odhadnutím očakávanej profitability tejto vzorky sme zistili, že 15 % klientov identifikovaných ako prémiových, má potenciál poisťovni priniesť zisk, ktorý tvorí až 40 % celkového očakávaného zisku (tabuľka 6). Platnosť Paretovho pravidla sa síce nepotvrdila, napriek tomu možno konštatovať, že „prémioví" zákazníci sú tou skupinou klientov, ktorej by poisťovňa mala venovať zvýšenú starostlivosť.

Tabuľka 6: Očakávaný profit z portfólia klientov (2252 zákazníkov)

Hodnota	Podiel klientov	Objem zisku	Podiel zisku
$CLV = 1$	15 %	646 359,66 €	41 %
$CLV = 0$	85 %	931 348,27 €	59 %
Spolu	100 %	1 577 707,93 €	100 %

5. Záver

Vytvorený model poukazuje na vzťah medzi finančným prínosom zákazníka pre poisťovňu a jeho charakteristikami. Je tvorený najvýznamnejšími premennými z hľadiska predikčnej sily vzhľadom na modelovanú premennú.

Profitabilita zákazníka je na základe nášho modelu dobre vysvetľovaná hlavne hodnotou celkovej aktuálnej výšky poistného a predpísaného poistného pre poistenie majetku a životné kapitálové poistenie. Štatisticky významné sa tiež javia premenné týkajúce sa poistenia motorového vozidla. Vplyv má tiež čas, kedy klient naposledy uzatvoril nejaký druh poistnej zmluvy a koľkokrát už počas vzťahu s poisťovňou zmluvu stornoval. Z demografických ukazovateľov na profitabilitu podľa modelu vplývajú vek (resp. veková kategória) a pohlavie.

Model má dobrú predikčnú silu a dokáže odlíšiť „prémiových" klientov od ostatných zákazníkov v existujúcom portfóliu. Zaradenie klientov do jednej z týchto skupín môže predstavovať podklad pre ďalšie marketingové aktivity poisťovne. Význam zameriavania sa na „najlepších" klientov je tiež podporený zistením, že hoci ich podiel v celkovej klientskej báze nemusí byť vysoký, potenciálny profit, ktorý môžu zabezpečiť, nemá zanedbateľnú výšku.

Literatúra

[1] Blažková, M.: *Marketingové řízení a plánovaní pro malé a střední firmy.* Grada Publishing, Praha, 2007.

[2] Godišová, M.: *Určenie potenciálu klientskeho portfólia.* Diplomová práca. Univerzita Komenského v Bratislave, 2015.

[3] Lin, A. Z.: *Variable Reduction in SAS by Using Weight of Evidence and Information Value.* SAS Global Forum 2013. Data Mining a Text Analytics. Paper 095-213 [online]. [cit. 14. 2. 2015]. URL: `http://support.sas.com/resources/papers/proceedings13/095-2013.pdf`

[4] Lošťáková, H.: *B-to-B marketing: Strategická marketingová analýza pro vytváření tržních příležitostí.* Professional Publishing, Praha, 2005.

[5] Perline, R.: *Development of Credit Scoring Applications Using SAS Enterprise Miner.* SAS Institute Inc., Cary, North Caroline, USA, 2011.

[6] Řezáč, M.: *Credit scoringové modely – vývoj, implementace, praxe* [online]. Bratislava, 2013. [cit. 1. 2. 2015]. URL: `http://www.math.muni.cz/~mrezac/prezentace/VS_2013_MRezac_s.pdf`

[7] Siddiqi, N.: *Credit Risk Scorecards. Developing and Implementing Intelligent Credit Scoring.* John Wiley & Sons, New Jersey, USA, 2006.

[8] Stankovičová, I., Vojtková, M.: *Viacrozmerné štatistické metódy s aplikáciami.* Iura Edition, Bratislava, 2007.

[9] Toporek, A.: *Understanding Customer Lifetime Value: A Non-Geek's Guide.* CTS Service Solutions, LLC [online], 2011. [cit. 4. 1. 2015]. URL: `http://customersthatstick.com/blog/customer-service-techniques/understanding-customer-lifetime- value-a-non-geek-guide/`

Obsah

Vědecké a odborné články

Informační bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Vydavatelem je Česká statistická společnost, IČ 00550795, adresa společnosti je Na padesátém 81, 100 82 Praha 10. Evidenční číslo registrace vedené Ministerstvem kultury ČR dle zákona č. 46/2000 Sb. je E 21214. Časopis je na Seznamu recenzovaných neimpaktovaných periodik vydávaných v ČR, více viz server `http://www.vyzkum.cz/`.

The Information Bulletin of the Czech Statistical Society is published quarterly.
The contributions in the journal are published in English, Czech and Slovak languages.

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.