VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Logistická regrese jako nástroj pro diskriminaci v lékařských aplikacích

Logistic regression – a tool for discrimination in surgery

2009 Pavlína Kuráňová

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně. Uvedla jsem všechny literární prameny a publikace, ze kterých jsem čerpala.

V Ostravě dne 7. května 2009 ……

Na tomto místě bych ráda poděkovala za cenné rady, užitečné informace a příkladné vedení při tvorbě diplomové práce svému vedoucímu **doc.Ing. Radimu Briši, Csc**. a své rodině a blízkým za trpělivost, kterou se mnou během této doby měli.

Abstrakt

Práce se investigativně zabývá analýzou lékařských dat pocházejících z chirurgických zákroků kolorekta z let 2001-2006. Tato data byla nejdříve transformována do podoby operačních faktorů a fysiologických faktorů a posléze použita pro účely generace logistického regresního modelu, který byl verifikován standardními statistickými, jakož i doplňkovými referenčními verifikačními technikami. Jako doplňková verifikační technika byla použita exponenciální analýza s vylepšením, která byla zalgoritmizována. Získané logistické regresní modely byly konfrontovány s dodaným referenčním modelem.
Nalezené logistické regresní modely jsou posléze využity pro účely diskriminační analýzy, která spočívá v možnosti zařazení pacienta s neznámou příslušností do jedné ze dvou skupin podle operačních komplikací. Na vzorku dodaných dat je provedena prognóza operačních komplikací s vyhodnocením chyby správnosti zařazení. Nalezený postup pro toto zařazování pacientů včetně vyhodnocení chyb je algoritmicky zpracován v prostředí MATLAB.

Klíčová slova: diskriminační analýza, logistická regrese, morbidita, lékařská data, exponenciální analýza, metoda maximální věrohodnosti.

Abstract

This thesis is investigativly engaged in analysis of medical data derivable from surgical intervention colectomy in years 2001 – 2006. These data was first transformed into form of surgical factors and physiological factors and finally these data was used for generation of logistic regression model purposes that was verified by standard statistical by way of supplementary reference verifications technique.
As a supplementary verifications technique was used an exponential analyses in improvement, which were algorithmized. Obtained logistic regressive models were confronted by supplied referential model.
Found logistics regressive models are finally utilized for discriminating analysis purposes that consist in possibility of unknown pertinence of patient which is formate into one of two groups according to surgical complications. In the sample of delivered data is performed prognosis of surgical complications with errors evaluation of formatting correctness. The process found for this patient formatting including mistake evaluation is algorithmic processed in MATLAB.

Keywords: discriminating analysis, logistics regression, morbidity, medical data, exponential analysis, maximum likelihood estimators.

Seznam použitých zkratek a symbolů

F	- fisherovo diskriminační kritérium
POSSUM	- Physiological and Operative Severiny for enUmeration of Mortality and Morbidity
PS	- fysiologické skóre
OS	- operační skóre
$S(X)$	- regresní funkce
ln	- přirozený logaritmus
Σ	- kovarianční matice
χ^2	- chí-kvadrát
$\frac{\partial L}{\partial \theta}$	- parciální derivace L podle parametru θ

Obsah

1 Úvod 4

2 Skórovací systémy a analýza dat 5

2.1 Skórovací systémy ……. 5

2.2 POSSSUM…….. 6

2.3 Analýza dodaných chirurgických dat……. 7

3 Diskriminační analýza…….. 8

3.1 Kanonická diskriminační analýza…… 8

3.2 Diskriminační funkce, diskriminace mezi dvěma skupinami……. 9

4 Logistická regrese 12

4.1 Statistické rozhodovací funkce…… 12

4.2 Logistická regrese…… 13

4.3 Logistická regrese jako nástroj pro diskriminaci…… 13

4.4 Metoda maximální věrohodnosti……. 15

4.5 Vícerozměrná metoda maximální věrohodnosti…….. 15

4.6 Metoda maximální věrohodnosti pro logistickou regresi…… 16

5 Testování modelů 18

5.1 Testování hypotéz …….. 18

5.2 Čistý test významnosti…… 18

5.3 Alternativní hypotézy……. 19

5.4 Stanovení testové statistiky…….. …….. 20

5.5 Pearsonův χ^2 test dobré shody……..……. 21

5.6 Pearsonův χ^2 test pro logistickou regresi…… 22

6 Generování nového modelu 24

6.1 Logistické rovnice modelů…….. 25

6.2 Testování vytvořených logistických rovnic…… 27

7 Algoritmické zpracování nových regresních modelů 30

7.1 Využití modelů pro predikci morbidity…….. 30

8 Exponenciální analýza a její modifikace 33

8.1 Exponenciální analýza…… 33

8.2 Vylepšená exponenciální analýza…… 33

8.3 Aplikace exponenciální a vylepšené exponenciální analýzy…….. 34

9 Závěr 39

Seznam použité literatury…… 40

Seznam příloh……. 41

Seznam obrázků

Obr. 1: Rozhodovací oblast dle P_{VALUE} …… **19**
Obr. 2: Znázornění datového souboru (2001-2006)…….. **25**
Obr. 3: Výsledný model (2001-2006)…….. **26**
Obr. 4: Grafické znázornění zařazení pacientů ze všech dat (364)…….. **31**
Obr. 5: Grafické znázornění zařazení pacientů z roku 2006 (36)……. **32**
Obr. 6: Znázornění počtu komplikací v modelu M-POSSUM……. **35**
Obr. 7: Znázornění počtu komplikací v modelu z let 2001-2006……. **36**
Obr. 8: Znázornění počtu komplikací v modelu z let 2001-2005……. **37**

Seznam tabulek

Tab. 1: Možnosti výsledku testu hypotéz..... …… **21**
Tab. 2: Tabulka odhadnutých četností…… **22**
Tab. 3: Tabulka empirických četností…… **23**
Tab. 4: χ^2 test modelu z dat z let 2001-2006…… **27**
Tab. 5: χ^2 test modelu z dat z let 2001-2005…… **27**
Tab. 6: Modely testovány na všech datech z let 2001-2006 (364)…….. **31**
Tab. 7: Modely testovány jen na datech z roku 2006 (36)……. **31**
Tab. 8: Exponenciální analýza a vylepšená exponenciální analýza využitím modelu M-POSSUM …… **35**
Tab. 9: Exponenciální analýza a vylepšená exponenciální analýza s využitím rovnice z dat 2001- 2006……. **36**
Tab.10: Exponenciální analýza a vylepšená exponenciální analýza s využitím rovnice z dat 2001- 2005…… **37**

Úvod

Počátkem 90.let 20. století způsobil nástup miniinvazivních technik v chirurgii převratné změny a významně ovlivnil další směry vývoje chirurgie. Tyto změny v různé míře zasáhly prakticky všechny oblasti chirurgie a nečekaně rychle nahrazují „klasické otevřené“ operační postupy.Obecně je miniinvazivní chirurgie spojována s menším operačním stresem a příznivějším pooperačním průběhem. Tato skutečnost je opakovaně uváděna i v řadě publikací věnovaných laparoskopické chirurgii v oblasti kolorekta.

Krátkodobé výsledky miniinvazivní chirurgie kolorekta jsou v literatuře nejčastěji uváděny formou procentuálně vyjádřené morbidity a časné mortality (mortalita do 30-ti dnů od operace). Tyto výsledky jsou pak srovnávány s výsledky souborů pacientů operovaných v rámci randomizovaných studií. Tento způsob hodnocení sebou nese určitá rizika. Existují již sice statisticky dobře kontrolované studie, zatím je jich však málo a obsahují jen menší soubory pacientů. Navíc u mnohých studií je nepřesně vymezen výklad pojmu morbidita. Výsledná čísla se pak pohybují v širokém rozmezí. Snažme se tedy o vytvoření sofistikovaného matematického modelu, který umožní srovnání aktuální a predikované morbidity.

Vypracování skórovacích systémů prošlo a stále prochází vývojem. Výsledkem je velké množství nejrůznějších systémů. V poslední době často testovanými jsou POSSUM (Physiological and Operative Severity Score for enUmeration of Mortality and Morbidity) a jeho modifikace.

Tato práce se zaměřuje na vytvoření objektivního skórovacího systému vytvořeného z konkrétních lékařských dat. Ty jsou porovnávány s modelem POSSUM převzatým z anglického časopisu

$$POSSUM \qquad \ln\left(\frac{R}{1-R}\right) = -5{,}91 + (0{,}16 \cdot PS) + (0{,}19 \cdot OS),$$

který popisuje riziko morbidity. Je sestaven z výsledků pacientů, kteří se podrobili operacím v anglických nemocnicích.

Cílem práce bylo vytvořit model na základě vícerozměrné logistické regrese, dle dat z dodaného datového souboru. Soubor pochází z Fakultní nemocnice Ostrava-Poruba z chirurgické kliniky a obsahuje veličiny: PS (předoperační skóre), OS (operační skóre) a veličinu morbidita (0 nebo 1). Hlavním cílem této práce bylo využití takto generovaného modelu pro účely co nejefektivnějšího zařazení pacienta s neznámou příslušností do jedné ze dvou skupin klasifikované morbidity (0 nebo 1). Dalším cílem byla algoritmizace tohoto procesu.

Verifikační část práce se věnuje exponenciální analýze a její modifikaci. Zabývá se vzájemným srovnáním a ověřením vytvořených modelů i modelu POSSUM.

2 Skórovací systémy a analýza dat

Zde vysvětlujeme pojmy vyskytující se v lékařské praxi, zejména týkající se kolorektální chirurgie. V druhé části kapitoly jsou popsány veličiny vyskytující se v datovém souboru.

2.1 Skórovací systémy [1]

Jednou z možností objektivního hodnocení dosažených krátkodobých výsledků (časné mortality, morbidity) a jejich porovnávání je stratifikace pacientů v souboru podle rizik založených na predikci individuální pravděpodobnosti komplikací. Předpokladem tohoto přístupu je robustní a ověřený matematický model vycházející z funkčního skórovacího systému, umožňující objektivní přiřazení rizika jednotlivým pacientům.

Obecně mohou být chirurgické skórovací systémy také klasifikovány jako předoperační, které se snaží o kvantifikaci rizik pacientů podstupujících operační výkon a fysiologické hodnotící závažnost stavu nebo onemocnění, vycházející z odpovědi organismu na nemoc nebo poranění (včetně operace), snažící se o prognózu průběhu a výsledku. Tyto systémy se často významně překrývají. Skóre vztahující se na individuálního pacienta určuje jeho individuální prognózu. Získaný výsledek tak může ovlivnit rozhodování o rozsahu vyšetřování, o způsobu a agresivitě léčby, o rozsahu výkonu a předoperační přípravě. Může se takto podílet i na racionalizaci nákladů. Skórovací systémy aplikované na skupiny pacientů dovolí smysluplnou analýzu dosažených výsledků morbidity. Umožní stratifikaci pacientů podle závažnosti stavu, podle jejich rizik a pravděpodobnosti komplikací. Pouze takto lze objektivně srovnávat dosažené výsledky. Skórovací systémy se tak mohou stát objektivním nástrojem zhodnocení kvality chirurgické péče.

2.1.1 Předoperační skóre

Je kvantitativní charakteristika pacienta, která se snaží předoperačně stanovit rizika jednotlivého pacienta, který podstupuje operaci. Vzhledem k tomu, že nejčastější pooperační komplikace jsou kardiorespirační a zároveň jsou tyto komplikace i častou příčinou smrti, zaměřují se některé systémy právě na predikci těchto rizik.

Např. - American Society of Anesthesiologists (ASA), kde je pacient zařazen do jedné z pěti kategorií na základě anamnézy a jednoduchého vyšetření
- Cardiac Risk Index (CRI) je systém, který je zaměřen na stanovení rizika kardiálních komplikací
- Prognostic Nutritional Index (PNI) je navrhnut pro stanovení rizik komplikací chirurgických pacientů ve vztahu k nutričnímu stavu

2.1.2 Fysiologické skóre

Fysiologické skóre závažnosti stavu je kvantitativní charakteristika pacienta, vycházející z fysiologické odpovědi organismu na nemoc nebo poranění. Většinou bylo toto skóre zaměřováno na skupiny kriticky nemocných s tendencí prognózovat spíše pro skupinu pacientů než pro jednotlivce.

Mohou se však podílet na rozhodování a eventuální změně rozsahu resuscitační péče u pacientů v kritickém stavu.

Např.:

- Acute Physiology and Chronic Health Evalution (APACHE), jehož cílem je klasifikace pacientů na základě závažnosti stavu
- APACHE II, je osvědčen pro použití u pacientů s nitrobřišní sepsí, diskutována je i jeho spolehlivost u traumatologických pacientů s vysokým nebo naopak nízkým rizikem
- Z dalších jsou to například: APACHE III., SAPS, AOSF apod.

2.1.3 Morbidita a Mortalita

Morbidita a časná mortalita jsou v chirurgii velmi často publikované krátkodobé výsledky, které jsou pak následně vyhodnocovány a srovnávány.

Pro laparoskopické operace v oblasti kolon se morbidita pohybuje v širokém intervalu 7 – 28 % a mortalita v rozmezí 0 – 3 %. Morbidita u laparoskopických operací rekta se pohybuje ještě v širším pásmu 7 – 44 % a mortalita u laparoskopických operací karcinomu rekta je udávána v rozmezí 0 - 7 %.

Pojem morbidita je často nedostatečně nebo nepřesně definován, což přispívá k omezené možnosti srovnávání výsledků.

Mortalita představuje objektivnější charakteristiku než morbidita. Pro tyto účely byla zvolena časná mortalita, tedy mortalita do 30 dnů od operace.

Výsledná morbidita a mortalita nezohledňuje odlišnosti v jednotlivých souborech pacientů. Často jsou vylučování pacienti s výraznou komoditou, pacienti s nepříznivým lokálním nálezem nebo naopak jsou zařazování pouze přesně definování pacienti.

2.2 POSSUM

Hrubá čísla uvádějící morbiditu jsou často zavádějící a zkreslená neboť nezohledňují variabilitu zdravotního stavu hodnoceného souboru pacientů, stav v okamžiku operace, obecný zdravotní stav populace a například ani zavedenou chirurgickou praxi. Neumožňují tak objektivní srovnání dosažených výsledků a problematické je i hodnocení nově zavedených postupů.

Skórovací systém POSSUM byl vyvinut Copelandem a kol. [2] počátkem 90. let. Původně měl sloužit jako nástroj pro srovnávání výsledků mezi jednotlivými institucemi zvláště v případech s výraznými rozdíly v morbiditě a mortalitě.

POSSUM vznikl z potřeby jednoduchého skórovacího systému, který by byl použitelný napříč celým spektrem chirurgických výkonů. Retrospektivní lineární multivariantní diskriminační analýzou bylo stanoveno 12 nezávislých faktorů závažnosti fysiologického stavu a 6 nezávislých faktorů závažnosti operačního výkonu (každý faktor s možností čtyř exponenciálních koeficientů 1, 2, 4 a 8), které se signifikantně podílí na pooperační mortalitě a morbiditě.

Logistická regresní analýza na hladině významnosti $\alpha = 0{,}001$ vyjádřila riziko morbidity následovně:

Riziko morbidity (***R***):

$$\ln\left(\frac{R}{1-R}\right) = -5{,}91 + \left(0{,}16 \cdot PS\right) + \left(0{,}19 \cdot OS\right),$$

kde PS znamená fysiologické skóre a OS znamená operační skóre.

POSSUM byl ověřován na řadě chirurgických pacientů a byl srovnáván s jinými skórovacími systémy. Byl použit ke srovnání výsledků kolorektální chirurgie, jak mezi různými pracovišti, tak mezi jednotlivými chirurgy.

2.3 Analýza dodaných chirurgických dat

V datovém souboru máme tyto skupiny dat: rok, OS, PS, morbidita, POSSUM a skupiny dat označeny exp(1), exp(2), exp(POSSUM), na těchto datech je využita exponenciální a vylepšená exponenciální analýza pro jejich pozdější zařazení do tabulek. Exponenciální analýza je detailně popsaná v kapitole **8**.

2.3.1 Skupiny OS a PS

Parametr fysiologického skóre (PS) se vztahuje k přijetí pacienta nebo k okamžiku bezprostředně předcházejícímu operaci. Parametr operační skóre (OS) lze doplnit po operačním výkonu.

Potřebná data jsou lehce dostupná a lze je získat i retrospektivně u většiny chirurgických pacientů.

2.3.2 Skupina morbidita

Parametr nabývající pouze dvou hodnot 0 nebo 1. Charakterizuje nám výsledek operace a pooperační průběh léčení pacienta. Hodnota 0 znamená, že pacient neměl žádné pooperační komplikace. Naopak hodnota 1, hodnotí stav, kdy došlo k jakýmkoliv komplikacím. Komplikace jsou přesně definovány a hodnoceny lékaři dle daných tabulek.

3 Diskriminační analýza

Problémem souvislosti skupiny kvantitativních a jedné alternativní či vícehodnotové nominální proměnné se zabývá, mimo jiné také diskriminační analýza. Primárně její úlohou bylo zkoumat schopnost sledovaných proměnných přispět k odlišení jednotlivých skupin jednotek v souboru (jak ji formuloval R.A.Fisher). Diskriminační analýza velmi často směřuje ke klasifikaci jednotek s neznámou skupinovou příslušností (jejíž zařazení se kupříkladu týká nějaké budoucí události). Konkrétní podoby klasifikační pravidlo nabývá na základě hodnot souboru kvantitativních proměnných a skupinové příslušnosti u těch jednotek, u nichž jsou všechny potřebné údaje k dispozici. Použije se pak k třídění nezařazených jednotek.

3.1 Kanonická diskriminační analýza [3]

Základem Fisherova pojetí diskriminační analýzy je nalézt takovou lineární kombinaci p sledovaných proměnných, tedy $Y = b^T x$, kde $b^T = [b_1, b_2, \ldots, b_p]$ je vektor parametrů, aby lépe než jakákoliv jiná lineární kombinace separovala uvažovaných H skupin, tak aby její vnitroskupinová variabilita byla co nejmenší a meziskupinová variabilita co největší.

Jestliže celkovou variabilitu původních sledovaných proměnných vyjádříme maticí T:

$$T = \sum_{h=1}^{H} \sum_{i=1}^{n_h} (x_{ih} - \bar{x})(x_{ih} - \bar{x})^T ,$$

rozložme ji na součet matice E vyjadřující vnitroskupinovou variabilitu,

$$E = \sum_{h=1}^{H} \sum_{i=1}^{n_h} (x_{ih} - \bar{x}_h)(x_{ih} - \bar{x}_h)^T ,$$

a matice B vyjadřující meziskupinovou variabilitu,

$$B = \sum_{h=1}^{H} \sum_{i=1}^{n_h} (\bar{x}_h - \bar{x})(\bar{x}_h - \bar{x})^T = \sum_{h=1}^{H} n_h (\bar{x}_h - \bar{x})(\bar{x}_h - \bar{x})^T ,$$

tedy $E + B = T$, pak součty čtverců $Q_B(Y)$ a $Q_E(Y)$, představují míru meziskupinové a vnitroskupinové variability pro novou veličinu Y, lze zapsat jako

$$Q_B(Y) = b^T B b \quad \text{a} \quad Q_E(Y) = b^T E b .$$

Největší meziskupinové a nejmenší vnitroskupinové variability veličiny Y dosáhneme při maximálním podílu

$$F = \frac{Q_B(Y)}{Q_E(Y)} = \frac{b^T B b}{b^T E b}$$

známém jako Fisherovo diskriminační kritérium. Pro stanovení veličiny $Y = b^T x$, která by postihovala odlišnosti mezi skupinami, je tedy třeba určit prvky vektoru b tak, aby maximalizoval diskriminační kritérium F.

Nalézt vektor b znamená položit parciální derivace diskriminačního kritéria podle prvků vektoru b rovny nule. Po provedení parciálních derivací lze rovnice spojit a maticově zapsat jako:

$$(B - \lambda E)b = 0 \text{, resp. } (BE^{-1} - \lambda I)b = 0; \quad |E| \neq 0 .$$

Soustava rovnic má netriviální řešení, pokud charakteristický polynom $|BE^{-1} - \lambda I|$ je roven nule. Tato charakteristická rovnice má r řešení, kterými jsou charakteristická čísla $\lambda_1, \lambda_2, \ldots, \lambda_r$ matice $BE^{-1}(\lambda_1 > \lambda_2 > \ldots > \lambda_r)$. Největšímu z těchto charakteristických čísel λ_1 odpovídá charakteristický vektor b_1, který maximalizuje diskriminační kritérium F.

Charakteristická rovnice $|BE^{-1} - \lambda I|$ ovšem neurčuje vektor b_1 jednoznačně, ale pouze stanovuje poměr mezi jeho prvky. Konkrétní hodnoty prvků vektoru b_1, tedy koeficienty pro hledanou lineární kombinaci, lze získat například tak, že volíme b_1, tak aby

$$\frac{1}{n-H} b_1^T E b = 1$$

(vnitroskupinovou variabilitu nové veličiny $Y_1 = b_1^T x$ vyjadřuje jednotkový rozptyl). Za tohoto předpokladu lze Fisherovo diskriminační kritérium F zapsat jako

$$F = \frac{1}{n-H} b_1^T B b ,$$

A tedy příslušné charakteristické číslo λ_1 vyjadřuje míru meziskupinové variability veličiny Y_1.

Lineární kombinace Y_1 se označuje jako první diskriminant. Geometricky ji lze chápat jako projekci bodů reprezentujících jednotlivé jednotky v p-rozměrném prostoru na přímku, do značné míry zachovávající rozdílnost skupin, do nichž jsou jednotky roztříděny. Umístění jednotek na této přímce vyjadřuje jejich diskriminační skóre. Střední hodnota tohoto skóre je podstatnou informací o tom nakolik jsou skupiny pomocí diskriminantu odlišeny.

3.2 Diskriminační funkce, diskriminace mezi dvěma skupinami [3]

Alternativním použitím diskriminační analýzy je klasifikace objektů neznámého původu do dvou nebo několika možných skupin. Diskriminační kritérium pro zařazování neznámých objektů do skupin je přitom funkce původních proměnných odhadnutá na základě výběrového souboru jednotek se známou příslušností ke skupinám.

Předpokládejme, že populace je rozdělena do dvou skupin (*v našem případě morbidita 1 nebo 0*) a rozdělení vícerozměrné náhodné veličiny x ve dvou skupinách je vícerozměrné normální s vektory

středních hodnot μ_1 a μ_2 a shodnými kovariančními maticemi Σ. Charakteristický vektor b, který maximalizuje Fisherovo diskriminační kritérium F, lze v tomto případě vyjádřit jako

$$b = k\Sigma^{-1}(\mu_1 - \mu_2) \quad \text{a} \quad \frac{1}{k}b = a = \Sigma^{-1}(\mu_1 - \mu_2),$$

kde k je libovolná konstanta. Pokud $b^T Eb/(n-H) = 1$, je

$$k = \left[(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)\right]^{-½} \quad \text{a} \quad k > 0.$$

Diskriminační funkce

$$x^T a = x^T \Sigma^{-1}(\mu_1 - \mu_2)$$

má ve skupinách normální rozdělení se středními hodnotami

$$\mu_1^T a = \mu_1^T \Sigma^{-1}(\mu_1 - \mu_2) \quad \text{a} \quad \mu_2^T a = \mu_2^T \Sigma^{-1}(\mu_1 - \mu_2),$$

jejíchž vzdálenost

$$\left(\mu_1^T a - \mu_2^T a\right) = [(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)]$$

je tedy vzdálenost dvou skupin. Střed mezi těmito středními hodnotami lze zapsat jako

$$c = \frac{1}{2}\left(\mu_1^T a + \mu_2^T a\right) = \frac{1}{2}(\mu_1 + \mu_2)\Sigma^{-1}(\mu_1 - \mu_2).$$

Pokud $x^T a > c$, má klasifikovaná jednotka blíže k první skupině, v opačném případě má blíže ke druhé skupině. Diskriminační pravidlo tedy vypadá takto

$$x^T \Sigma^{-1}(\mu_1 - \mu_2) > \frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) \quad \rightarrow \quad \text{skupina 1;}$$

$$x^T \Sigma^{-1}(\mu_1 - \mu_2) < \frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) \quad \rightarrow \quad \text{skupina 2.}$$

Lze ukázat, že za předpokladu vícerozměrné normality společnou kovariančí maticí toto kritérium zařazuje x do skupiny 1, pokud $f_1(x)/f_2(x) > 1$, jinak do skupiny 2. $f_1(x)$ a $f_2(x)$ jsou zde hustoty pravděpodobnosti vícerozměrných normálních rozdělení pro skupinu 1, resp. 2.

V předchozích uvedených kritériích předpokládáme stejné zastoupení z obou skupin v populaci, tedy i stejnou pravděpodobnost mylného zařazení jednotky pocházející z první skupiny do druhé a naopak. Skupiny, ale stejně velké být nemusí.

Předpokládejme, že π_1, resp. π_2, je rozsahu skupiny odpovídající apriorní pravděpodobnost příslušnosti jednotky k určité skupině. Na základě hodnot p veličin zjištěných u některé jednotky můžeme uvažovat podmíněnou aposteriorní pravděpodobnost této příslušnosti a podle Bayesova vzorce ji vyjádřit jako

$$\frac{\pi_h f_h(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}, \quad h = 1, 2.$$

Jednotka neznámého původu by pak zřejmě měla být zařazena do skupiny s vyšší aposteriorní pravděpodobností. Tedy do skupiny 1, pokud

$$f_1(x)/f_2(x) > \pi_2/\pi_1, \qquad (*)$$

v opačném případě do skupiny 2.

Toto pravidlo, ale může vést k chybnému zařazení jednotky z první skupiny do skupiny druhé, a naopak, s pravděpodobnostmi

$$\int_{\Omega_{21}} f_1(x)dx, \text{ resp. } \int_{\Omega_1} f_2(x)dx,$$

kde Ω_1, resp. Ω_2 jsou dvě nepřekrývající se oblasti všech možných výsledků x (je-li x z Ω_1, zařadíme jednotku do první skupiny, je-li z Ω_2, do druhé). Celkovou pravděpodobnost mylné klasifikace lze vyjádřit jako

$$\pi_1 \int_{\Omega_2} f_1(x) + \pi_2 \int_{\Omega_1} f_2(x).$$

Lze ukázat, že pravidlo (*) celkovou pravděpodobnost mylné klasifikace minimalizuje.

4 Logistická regrese

V této kapitole se zabýváme teoretickým základem logistické regrese jejím odvozením a využitím pro diskriminaci. V další části kapitoly je naformulována metoda maximální věrohodnosti a její použití pro logistickou regresi.

4.1 Statistické rozhodovací funkce [4]

Jednotlivé diskriminační procedury jsou odvozeny na základě teorie statistických rozhodovacích funkcí.

K nalezení optimálního rozhodovacího pravidla je využito bayesovského přístupu. Roli neznámého parametru, o jehož hodnotě chceme rozhodnout, hraje náhodná veličina $Y \in \{0; 1\}$, která má pravděpodobnostní funkci $q(y)$. Rozhodnutí je prováděno na základě hodnoty náhodného vektoru $X \in R^p$, jenž má hustotu $r(x)$. Podmíněnou hustotu $\underline{X}$ za podmínky $Y = y$ označíme $r(\underline{x}|\, y)$. Nechť $\delta : R^p \to \{0;1\}$ je rozhodovací funkce a H je množina všech funkcí $\delta : R^p \to \{0;1\}$.

Ztrátovou funkci zavedeme jako

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

Riziková funkce je definována jako

$$R(Y, \delta) = E\left[L(Y, \delta(\underline{X}))|\ Y\right] = \int_{R^p} L(Y, \delta(\underline{x})) \cdot r(\underline{x}|\, y)\, d\underline{x}$$

Baysovské riziko se určí jako

$$\rho(\delta) = ER(Y, \delta) = \sum_{y=0}^{1} R(y, \delta) \cdot q(y)$$

Optimální rozhodovací funkci δ^* potom dostaneme jako

$$\delta^* = \arg\min_{\delta \in H} \rho(\delta)$$

Označme podmíněnou pravděpodobnostní funkci Y za podmínky $\underline{X} = \underline{x}$ jako $p(y|\underline{x})$. Nechť $p(1|\underline{x}) = P(Y = 1|\underline{X} = \underline{x}) = \pi(\underline{x})$ a $p(0|\underline{x}) = P(Y = 0|\underline{X} = \underline{x}) = 1 - \pi(\underline{x})$. Existuje-li pro všechna $x \in R^p$

$$\hat{\delta}(\underline{x}) = \arg\min_{\delta \in H} E\left[L(Y, \delta(\underline{X}))|\underline{X} = \underline{x}\right] = \sum_{y=0}^{1} L(Y, \delta(\underline{x})) \cdot p(y|\underline{x}),$$

Lze snadno pomocí Bayesovy věty ukázat, že $\hat{\delta} = \delta^*$ .Přímým výpočtem lze dále nalézt vyjádření rizikové funkce a bayesovského rizika:

$$R(0, \delta) = P(\delta(\underline{X}) = 1 | Y = 0),$$
$$R(1, \delta) = P(\delta(\underline{X}) = 0 | Y = 1),$$
$$\rho(\delta) = P(\delta(\underline{X}) = 1, Y = 0) + P(\delta(\underline{X}) = 0, Y = 1).$$

Vidíme tedy, že bayesovské riziko lze v této situaci interpretovat jako pravděpodobnost špatného rozhodnutí o hodnotě Y , tj. o zařazení daného objektu do skupiny. Dále budeme bayesovské riziko nazývat jako pravděpodobnost chyby.

4.2 Logistická regrese [5]

Logistickou regresí rozumíme takový model, kde nezávislá proměnná Y je kódována pouze hodnotami 0 a 1. Hodnota 0 udává, že daná pozorovaná situace nenastala a hodnota 1 představuje, že daný pozorovaný jev nastal. Nezávislou proměnnou označíme vektor $x = (x_1, x_2, \ldots, x_n)$ a závislou proměnnou označíme y a píšeme $y = g(x)$. Hledaný vícerozměrný logistický model má tvar:

$$g(x) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n$$

kde,

$\beta_i \quad i = 1, 2, \ldots p$ jsou koeficienty

$x_i \quad i = 1, 2, \ldots p$ jsou nezávislé proměnné

Výsledný logistický regresní model, pak bude mít tvar:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

Hledané koeficienty $\beta_i, i = 1, 2, \ldots p$ odhadneme pomocí metody maximální věrohodnosti. $\pi(x)$ je tzv. logitová transformace, která je dominantní při logistické regresi. Její důležitost spočívá v tom, že $g(x)$ má mnohé žádané vlastnosti lineárního regresního modelu. Logit $g(x)$ je lineární v parametrech, může být spojitý a může se pohybovat v intervalu $(-\infty, \infty)$ v závislosti na x .

4.3 Logistická regrese jako nástroj pro diskriminaci [4]

Vychází z teorie diskriminační analýzy popsané v kapitole 3.2. Logistická regrese nebyla původně vytvořena pro diskriminaci, ale lze ji pro ni úspěšně použít.

Model logistické regrese, který je upravený pro účely diskriminace, je definován následovně. Nechť $Y_1, \ldots, Y_n$ je posloupnost nezávislých náhodných veličin s alternativním rozdělením, jehož parametr splňuje:

$$P(Y_i = 1|\underline{X}_i = \underline{x}_i) = \left[1 + e^{(-\beta_0 - \beta' x_i)}\right]^{-1},$$
$$P(Y_i = 0|\underline{X}_i = \underline{x}_i) = \left[1 + e^{(-\beta_0 + \beta' x_i)}\right]^{-1},$$

kde $(\beta_0, \beta')'$ je neznámý $(p+1)$ rozměrný parametr a $X_1, \ldots, X_n$ je posloupnost nezávislých náhodných veličin. Tento model má tzv. učící fázi, ve které známe u každého objektu jak hodnoty $\underline{X}_i$, tak hodnoty Y_i (tj. víme, do které skupiny ten který objekt patří). Na základě této znalosti odhadneme parametry β_0, β a poté dostaneme odhad $\tilde{\pi}(\underline{x})$ funkce $\pi(\underline{x})$,
kde

$$\pi(\underline{x}) = P(Y = 1|\underline{X} = \underline{x}) = \left[1 + e^{(-\beta_0 - \beta' \underline{x})}\right]^{-1}.$$

Další objekt, u kterého neznáme jeho zařazení a u něhož jsme naměřili hodnotu $\underline{X}$ pomocných znaků, přiřadíme do jedné ze dvou skupin podle hodnoty rozhodovací funkce.

Zde sice neznáme apriorní hustotu veličiny Y ani podmíněnou hustotu náhodné veličiny $\underline{X}$ ze podmínky $Y = y$, ale pro výpočet optimální rozhodovací funkce nám postačí znalost podmíněné hustoty Y za podmínky $\underline{X} = \underline{x}$, která je určena hodnotou funkce $\pi(\underline{x})$. Pokud $\delta(x) = j$, je totiž

$$E\left[L(Y, \delta(\underline{X}))|\underline{X} = \underline{x}\right] = \sum_{y=0}^{1} L(y, \delta(\underline{x}))p(y|\underline{x}) = \sum_{y=0}^{1} L(y, j)p(y|\underline{x}) =$$

$$L(1-j, j)p(1-j|\underline{x}) = \begin{cases} \pi(x), & j = 0, \\ 1 - \pi(x), & j = 1. \end{cases}$$

Tedy

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

Toto minimum existuje $\forall x \in R^P$ a tudíž můžeme psát

$$\delta^*(\underline{x}) = \arg\min_{j=0,1} L(1-j, j)p(1-j|\underline{x}).$$

Tedy objekt, na němž jsme naměřili hodnotu $\underline{X}$ pomocných znaků, zařadíme do první skupiny, pokud

$$\pi(\underline{X}) \geq 1 - \pi(\underline{X}), \; tj.$$
$$\beta_0 + \beta' \underline{X} \geq 0$$

Tudíž objekt zařadíme do první skupiny, pokud $S(\underline{X}) \geq 0$ a do nulté skupiny, pokud $S(\underline{X}) < 0$.
Přitom $S(\underline{x}) = \beta_0 + \beta' \underline{x} < 0$. Pokud $S(\underline{X}) = 0$, tj. $\pi(\underline{X}) = \frac{1}{2}$, můžeme objekt zařadit do libovolné skupiny, aniž bychom zvýšili hodnotu pravděpodobnosti chyby.

K vlastní diskriminaci však musíme použít odhad $\widetilde{S}(\underline{x})$ funkce $S(\underline{x})$, ve kterém jsou neznámé parametry β_0, β nahrazeny odhady $\widetilde{\beta}_0, \widetilde{\beta}$, tedy

$$\widetilde{S}(\underline{x}) = \widetilde{\beta}_0 + \widetilde{\beta}' \underline{x}.$$

Hlavní výhodou tohoto modelu je fakt, že neklade žádné podmínky na rozdělení náhodných vektorů. $X_1, \ldots, X_n$.

4.4 Metoda maximální věrohodnosti [4]

Odhady získané touto metodou se všeobecně vyznačují dobrými statistickými vlastnostmi. Metoda maximální věrohodnosti je principielně jednoduchá metoda pro konstrukci odhadů neznámých parametrů známých rozdělení pravděpodobnosti, která je založena na maximalizaci věrohodností funkce.

Nechť $(t_1, \ldots, t_n)^T$ je náhodný výběr z rozdělení s hustotou $f(t;\Theta)$, kde Θ je neznámý parametr. Nyní nalezněme funkci věrohodnosti danou

$$L(t_1, \ldots, t_n; \Theta) = f(t_1, \Theta) \cdot f(t_2, \Theta) \ldots f(t_n, \Theta) = \prod_{i=1}^{n} f(t_1; \Theta)$$

a z ní získejme Θ tak, aby $\hat{\Theta} = \hat{\Theta}(t_1, \ldots, t_n)$ bylo nejlepším odhadem pro Θ. Pravá strana rovnice je sdružená hustota pravděpodobnosti *n*-nezávislých proměnných $(t_1, \ldots, t_n)$ se stejným rozdělením.

Tedy L je funkcí neznámého parametru Θ, který je odhadován, metoda maximální věrohodnosti je založena na získání takové hodnoty Θ, která maximalizuje L. Při praktických výpočtech se ukázalo jako výhodnější maximalizovat spíše funkci $\ln L$, namísto L. Což je možné, protože obě operace jsou ekvivalentní.
Rovnici tedy zlogaritmujeme, položíme rovnu 0 a vypočteme neznámé parametry Θ.

$$\frac{\partial \ln L(t_1, \ldots, t_n)}{\partial \Theta} = 0$$

4.5 Vícerozměrná metoda maximální věrohodnosti

Mějme vektor neznámých koeficientů $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_p)$ a vektor nezávislých proměnných $X = (x_1, x_2, \ldots, x_n)$.
Abychom nalezli vektor neznámých $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_p)$, řešíme soustavu rovnic:

$$\frac{\partial \ln L(x_1, x_2, \ldots, t_n)}{\partial \Theta_i} = 0, \qquad i = 1, 2, \ldots, p.$$

4.6 Metoda maximální věrohodnosti pro logistickou regresi [5]

Metodu maximální věrohodnosti pro logistickou regresi provedeme pro dvě nezávislé proměnné x_1, x_2 a pro závislou proměnnou y. Počet hledaných koeficientů je $p+1$, kde p je počet závislých proměnných. V našem případě budeme hledat tři koeficienty β_0, β_1, β_2 pro dvě nezávislé proměnné.

Pro logistickou regresi bude mít věrohodnostní funkce tvar:

$$L(x_1, x_2; \beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}$$

kde

$$\pi(x_i) = \frac{e^{g(x_i)}}{1 + e^{g(x_i)}}$$

a

$$g(x_i) = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i}$$

a kde $\beta = (\beta_0, \beta_1, \beta_2)$ je vektor neznámých koeficientů.

Po zlogaritmování dostaneme tuto rovnici:

$$\ln L(x_1, x_2; \beta) = \ln \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}$$

Po úpravě dostaneme tuto rovnici:

$$\ln L(x_1, x_2; \beta) = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}$$

kde

$$\pi(x_i) = \frac{e^{g(x_i)}}{1 + e^{g(x_i)}}$$

a

$$g(x_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

a kde $\beta = (\beta_0, \beta_1, \beta_2)$.

Abychom nalezli věrohodnostní funkci budeme řešit soustavu tří rovnic o třech neznámých:

$$\begin{aligned}\sum_{i=1}^{n}\left[y_i - \pi(x_i)\right] &= 0\\ \sum_{i=1}^{n} x_{1i}\left[y_i - \pi(x_i)\right] &= 0\\ \sum_{i=1}^{n} x_{2i}\left[y_i - \pi(x_i)\right] &= 0\end{aligned}$$

Z této soustavy vypočteme odhady neznámých parametrů $\beta_0, \beta_1, \beta_2$.

Po nalezení koeficientů bude mít regresní model tvar:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}},$$

kde $g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

5 Testování modelů

V této kapitole jsou popsány metody testování hypotéz, definování správné hypotézy, použití vhodného testovacího modelu a nalezení hodnoty P_{VALUE}. Dále se zabýváme teoretickým testováním modelu, tedy rozhodujeme o pravdivosti hypotéz. V tomto rozhodovacím procesu proti sobě stojí nulová a alternativní hypotéza a naším cílem je rozhodnout, zda data z výběrového souboru odpovídají hypotéze.

5.1 Testování hypotéz

Statistické hypotézy – jsou takové domněnky o populaci, jejichž pravdivost lze ověřovat prostřednictvím statistických testů významnosti.
Testy významnosti – postupy či procedury, které na základě náhodného výběru objektivně rozhodnou, zda má být ověřována hypotéza zamítnuta či nikoliv.
Nulová hypotéza H_0 - ověřována hypotéza, o jejímž zamítnutí rozhoduje test významnosti.
Alternativní hypotéza H_A - hypotéza, kterou přijímáme, zamítneme-li nulovou hypotézu.

5.2 Čistý test významnosti [6]

Čistý test významnosti zodpovídá otázku, zda získaný náhodný výběr $\underline{X}$ (jeho pozorované hodnoty) je či není extrémní s ohledem na nějakou testovanou hypotézu o populaci.

Čistý test významnosti se skládá z následujících kroků:

1. Nulová hypotéza (H_0) – vyjadřuje domněnku o povaze populace. Musí být specifikovaná dostatečně přesně, aby definovala tzv. pravděpodobnostní míru na populaci.
2. Stanovení testové statistiky $T(\underline{X})$ - je nějaká funkce náhodného výběru, odvozena z populace. Výběr funkce je determinován charakteristikami pravděpodobnostního rozdělení z něhož populace pochází a jehož se zároveň týká nulová hypotéza.
3. Stanovení tzv. *nulového pravděpodobnostního rozdělení* $H_0 : F_0(x) = P(T(\underline{X}) < x \mid H_0)$. Přitom H_0 musí být specifikována natolik přesně, aby toto pravděpodobnostní rozdělení bylo jednoznačně určeno.
4. Výpočet pozorované hodnoty testové statistiky: $t = x_{OBS}$ a statistiky zvané P_{VALUE}. Pozorovanou hodnotu této statistiky P_{VALUE} vypočteme v závislosti na kontextu nulové

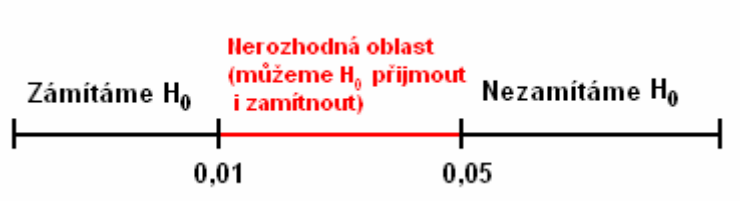
hypotézy, avšak interpretace této hodnoty je vždy stejná. Pokud H_0 platí, pak se dá ukázat, že rozdělení statistiky P_{VALUE} je uniformní (rovnoměrné). Tedy:

$$P\left(P_{VALUE}\left(\underline{X}\right) < p \mid H_0\right) = p$$

Proto P_{VALUE} má stejnou interpretaci pro všechny nulové hypotézy nezávislé na nulovém rozdělení. Je jasné, že čím menší P_{VALUE}, tím extrémnějších hodnot nabývá testová statistika s ohledem na nulové rozdělení, tedy tím silnější je výpověď náhodného výběru proti nulové hypotéze. Jestliže váha výpovědi proti nulové hypotéze roste při klesajícím P_{VALUE}, bylo by nerozumné označit nějaké P_{VALUE} jako mez, pod kterou už budou nulové hypotézy zamítány, a analogicky, nad kterou budou tyto hypotézy akceptovány. Proto vymezíme pro P_{VALUE} tzv. nerozhodnou oblast (interval), oddělující oblast zamítnutí od oblasti přijetí.

5. Rozhodnutí na základě P_{VALUE}

$P_{VALUE} < 0{,}01$	zamítáme H_0
$0{,}01 < P_{VALUE} < 0{,}05$	nerozhodná oblast
$P_{VALUE} > 0{,}05$	nezamítáme H_0



Obr. 1: Rozhodovací oblast dle P_{VALUE}.

5.3 Alternativní hypotézy [6]

Z definice P_{VALUE} je jasné, že čistý test významnosti pro účely testování hypotéz vyžaduje nejen přesnou specifikaci nulové hypotézy, ale také komentář k tomu, která alternativní hypotéza může být správná, v případě, že nulová hypotéza je zamítnuta. Tato alternativní hypotéza ovšem nemusí být specifikována natolik přesně, jako nulová hypotéza. Při volbě vhodné definice P_{VALUE} je nutno znát pouze to, v jakém vztahu je alternativní hypotéza vzhledem k nulové hypotéze. Stanovení alternativní hypotézy však musí být v souladu s pozorovanou hodnotou testové statistiky.

Tyto hodnoty testové statistiky, které mají malé P_{VALUE} při platnosti H_0 by měly mít tendenci k většímu P_{VALUE} při platnosti alternativy a naopak. (Velká P_{VALUE} při platnosti H_0 by měla odpovídat malým P_{VALUE} při platnosti alternativy.)

5.4 Stanovení testové statistiky [6]

Pro rozhodnutí zda testovanou nulovou hypotézu přijmeme nebo zamítneme použijeme testovou statistiku. Statistický test rozhodne zda data z výběrového souboru $(\underline{X})$ odpovídají nulové hypotéze.

Pro logistickou regresi se například používají:

- Pearsonův χ^2 test
- Test nulovosti koeficientů
- Hosmer-Lemeshowovy testy

Jelikož při rozhodování o nulové hypotéze vycházíme z výběrového souboru, který nemusí dostatečně přesně odpovídat vlastnostem základního souboru, můžeme se při rozhodování dopustit 2 druhů chyb.

Chyba I. druhu - vzniká, je-li nulová hypotéza ve skutečnosti platná a my ji přesto zamítneme, pravděpodobnost, že k této chybě dojde značíme α

Chyba II. druhu - je nezamítnutí nulové hypotézy v případě, že je platná hypotéza alternativní, pravděpodobnost této chyby značíme β

Abychom mohli tyto chyby blíže determinovat, musíme vyjít z tzv. teoretického modelu pro testování hypotéz. V němž se zvolí pevné číslo α … hladina významnosti, což bývá nejčastěji hodnota 0,05 (popř. 0,01). Rozhodování pak probíhá dle následujícího klíče:
Pokud

$P_{VALUE} < \alpha$ zamítáme H_0

$P_{VALUE} > \alpha$ nezamítáme H_0

V následující tabulce jsou popsány všechny situace, které mohou při rozhodování vzniknout:

SKUTEČNOST	**ZAMÍTÁME** H_0	**NEZAMÍTÁME** H_0
Platí H_0	**chyba I.druhu,** pravděpodobnost: α	**správné rozhodnutí,** pravděpodobnost: $1-\alpha$
Platí H_A	**správné rozhodnutí,** pravděpodobnost: $1-\beta$	**chyba II.druhu,** pravděpodobnost: β

Tab. 1: Možnosti výsledku testu hypotéz

5.5 Pearsonův χ^2 test dobré shody [7]

Pearsonův χ^2 test dobré shody se používá k testování hypotézy o procentním rozdělení v základním souboru nebo o pravděpodobnostním rozdělení náhodné veličiny. Je to jednoduchý test založený na rozdílu mezi pozorovanými (empirickými) a očekávanými (teoretickými) četnostmi.

Na předem zvolené hladině významnosti budeme testovat nulovou hypotézu H_0, že náhodná veličina (základní soubor) má určité rozdělení při alternativní hypotéze H_A, že náhodná veličina (základní soubor) má rozdělení jiné než to, které je specifikováno nulovou hypotézou.

Chceme-li zjistit, jak dobře se pozorované a očekávané četnosti shodují, zkoumáme je pomocí rozdílu těchto četností.

Pro pozorované (empirické) četnosti v rámci daného testu platí $\sum_{i=1}^{k} n_i = n$, kde $k \geq 2$, je počet disjunktních tříd.

Výrazy (np_i) se nazývají v rámci daného testu očekávané (teoretické) četnosti a platí $\sum_{i=1}^{k} np_i = n$.

Je-li nulová hypotéza pravdivá, pak pozorované a očekávané četnosti by měly být zhruba stejné a statistika bude mít malou hodnotu. Jinými slovy velké hodnoty poskytují argumenty proti nulové hypotéze.

Obecný tvar χ^2 statistiky: $\chi^2 = \sum_{i=1}^{n} \frac{(n_i - np_i)^2}{np_i}$.

Proč testovat předpoklady modelů? Logistický model sice neklade žádné zvláštní požadavky na rozdělení náhodných veličin $\underline{X}_1,\ldots,\underline{X}_n$, ale zato předpokládá specifický tvar pravděpodobnosti $P(Y=1|\underline{X}=\underline{x})$. Skutečnost, že opravdu platí $P(Y=1|\underline{X}=\underline{x})=\left[1+e^{(-\beta_0-\beta'\underline{x})}\right]^{-1}$, bychom měli ověřit nejlépe pomocí nějakého statistického testu.

5.6 Pearsonův χ^2 test pro logistickou regresi [4]

Naměřené hodnoty $X_1,\ldots,X_n$ se mohou opakovat, tedy nemusejí být nutně náhodné. Nechť I je počet různých hodnot veličin $\underline{X}_1,\ldots,\underline{X}_n$ v učící skupině a $\underline{x}_1,\ldots,\underline{x}_n$ jsou tyto hodnoty. Celý model logistické regrese pracuje totiž s podmíněným rozdělením veličin $Y_1,\ldots,Y_n$ za podmínky $\underline{X}_1=\underline{x}_1,\ldots,\underline{X}_n=\underline{x}_n$. Hodnoty $Y_1,\ldots,Y_n$ vyjadřující zařazení i-tého objektu do jedné ze dvou skupin přeznačme $\underline{Y}_{i,j}$, $i=1,\ldots,I,\quad j=1,\ldots,m_i$, kde m_i je počet objektů s hodnotou vysvětlujících znaků $\underline{X}_i$ a $\underline{Y}_{i,j}$, $j=1,\ldots,m_i$ označuje zařazení objektu, u nichž mají vysvětlující znaky hodnotu $\underline{X}_i=\underline{x}_i$. Dále označujeme $n_1=\sum_{i=1}^{I}\sum_{j=1}^{m_i}Y_{i,j},\quad n_0=\sum_{i=1}^{I}\sum_{j=1}^{m_i}(1-Y_{i,j})$.

Metodou maximální věrohodnosti získáme odhady $\tilde{\beta}_0,\tilde{\beta}$ parametrů β_0,β. Pomocí těchto odhadů spočítáme odhady logistických pravděpodobností $\tilde{\pi}(\underline{x}_i)=\tilde{\pi}_i=\left[1+e^{(-\tilde{\beta}_O-\tilde{\beta}'\underline{x}_i)}\right]^{-1}$. Přímo z věrohodnostních rovnic plynou vztahy

$$n_1=\sum_{i=1}^{I}\sum_{j=1}^{m_i}Y_{i,j}=\sum_{i=1}^{I}m_i\tilde{\pi}_i,\qquad n_0=\sum_{i=1}^{I}\sum_{j=1}^{m_i}(1-Y_{i,j})=\sum_{i=1}^{I}m_i(1-\tilde{\pi}_i)$$

Celou situaci lze popsat kontingenční tabulkou typu $2\times I$ s danými marginálními sloupcovými četnostmi. Přitom i-tý sloupec tabulky reprezentuje binomické rozdělení s parametry m_i, $\pi(\underline{x}_i)=\pi_i$.

Tabulka odhadnutých četností:

		X			
		x_1	…	x_I	
Y	**1**	$m_1\tilde{\pi}_1$	…	$m_I\tilde{\pi}_I$	n_1
	0	$m_1(1-\tilde{\pi}_1)$	…	$m_I(1-\tilde{\pi}_I)$	n_0
		m_1	…	m_I	n

Tab. 2: Tabulka odhadnutých četností

Tabulka empirických (pozorovaných) četností:

		X			
		x_1	…	x_I	
Y	**1**	$Y_{1\bullet}$	…	$Y_{I\bullet}$	n_1
	0	$m_1 - Y_{1\bullet}$	…	$m_I - Y_{I\bullet}$	n_0
		m_1	…	m_I	n

Tab. 3: Tabulka empirických četností

Jelikož máme dány marginální sloupcové četnosti, jsou hodnoty v naší tabulce vázány I podmínkami $\left(m_1\pi_1 + m_1(1-\pi_1) = m_1, \ldots, m_I\pi_I + m_I(1-\pi_I) = m_I\right)$. Tento údaj budeme potřebovat pro výpočet stupňů volnosti v Pearsonově testové statistice, která má tvar:

$$Z^2 = \sum_{i=1}^{I} \frac{\left(Y_{i\bullet} - m_i\tilde{\pi}_i\right)^2}{m_i\tilde{\pi}_i} + \sum_{i=1}^{I} \frac{\left(m_i - Y_{i\bullet} - m_i(1-\tilde{\pi}_i)\right)}{m_i(1-\tilde{\pi}_i)} = \sum_{i=1}^{I} \frac{\left(Y_{i\bullet} - m_i\tilde{\pi}_i\right)^2}{m_i\tilde{\pi}_i(1-\tilde{\pi}_i)}.$$

Pomocí této statistiky můžeme testovat shodu dat v pozorované tabulce s tabulkou teoretickou, která je v našem případě založena na modelu logistické regrese.

Při hypotéze H_0: platí logistický model, má statistika Z^2 asymptoticky rozdělení χ^2 s následujícím počtem stupňů volnosti:

Velikost kontingenční tabulky – počet vazeb v teoretické tabulce – počet odhadnutých parametrů =

$$= 2I - I - (p+1) = I - (p+1).$$

Tedy při platnosti H_0 je $Z^2 \to \chi^2\left(I - (p+1)\right)$. Nutné je splnit podmínku $I > (p+1)$.

6 Generování nového modelu

Pro vytvoření nového modelu použijeme metodu logistické regrese viz kapitola 4. Naše obecná regresní rovnice bude mít tvar:

$$S(x) = \beta_0 + \beta_1 \cdot 0S + \beta_2 \cdot PS$$

kde OS je operační skóre a PS je fyziologické skóre.

Z poskytnutých lékařských dat z Fakultní nemocnice Ostrava-Poruba z chirurgické kliniky vytvoříme model logistické regrese pomocí programu Statgraphics.

Konstrukce generování nového modelu:

- načteme data z datového souboru „data.xls“ do programu Statgraphics, nezávislé proměnné OS, PS a závislou proměnnou morbidita
- pomocí vícerozměrné logistické regrese sestavíme model
- výsledný model logistické regrese zobrazuje závislost mezi dvěma nezávislými proměnnými OS a PS a mezi závislou proměnnou morbidita

6.1 Logistické rovnice modelů

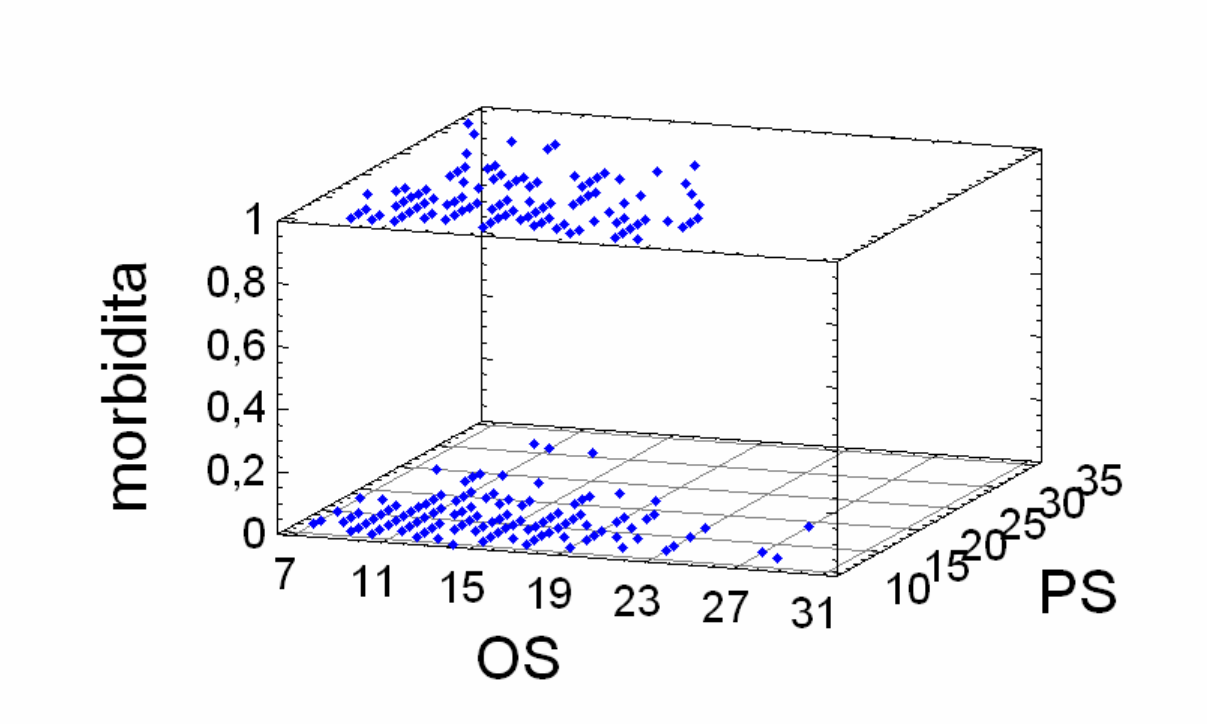
Rovnice jsou sestaveny z poskytnutých dat z Fakultní nemocnice Ostrava-Poruba z chirurgické kliniky:

a) pro data z let 2001 – 2006

Výsledná rovnice:

$$\ln\left(\frac{R}{1-R}\right) = -2{,}19973 + 0{,}0548604 * OS + 0{,}0725761 * PS \quad (1)$$

Kde PS je parametr fysiologického skóre a OS je parametr operačního skóre. Znázornění datového souboru z let 2001–2006 na obrázku 2.



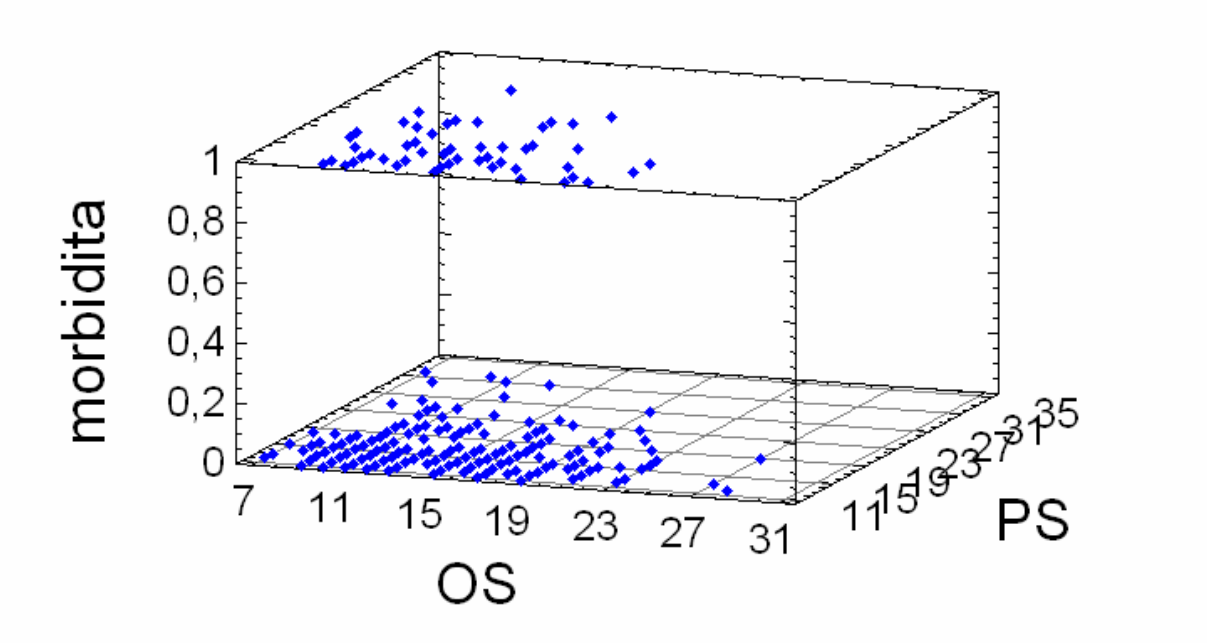
Obr. 2: Znázornění datového souboru(2001-2006)

b) pro data z let 2001 – 2005

Výsledná rovnice:

$$\ln\left(\frac{R}{1-R}\right) = -2{,}33399 + 0{,}0702439 * OS + 0{,}0637289 * PS \quad (2)$$

Kde PS je parametr fysiologického skóre a OS je parametr operačního skóre. Znázornění datového souboru z let 2001–2005 na obrázku 3.



Obr. 3: Znázornění datového souboru(2001-2005)

6.2 Testování vytvořených logistických rovnic

Zde se budeme zabývat tím, zda je výsledný model kompaktní s daty a jestli jsou koeficienty OS a PS statisticky významné a zda model na těchto datech závisí nebo ne.

6.2.1 χ^2 test dobré shody modelu

Tento test ověřuje hypotézu o tom, zda jsou nové logistické regresní funkce adekvátní s pozorovanými daty.

a) pro data z let 2001-2006 *(viz rovnice (1))*

			Morbidita 1		Morbidita 0	
Třída	**Logitový interval**	**n**	**Pozorováno**	**Predikce**	**Pozorováno**	**Predikce**
1	< -0,635056	83	22	26,6839	61	56,3161
2	-0,635056 až -0,433331	66	29	24,6184	37	41,3816
3	-0,433331 až -0,249321	69	31	28,8279	38	40,1721
4	-0,249321 až 0,0181291	70	29	32,954	41	37,046
5	> 0,0181291	71	42	39,9158	29	31,0842
Celkem		359	153		206	

Tab. 4: χ^2 test modelu z dat z let 2001-2006

χ^2 = 3,88165 … počet stupňů volnosti = 3 … P_{VALUE} ***= 0,274527***

b) pro data z let 2001-2005 *(viz rovnice (2))*

			Morbidita 1		Morbidita 0	
Třída	**Logitový interval**	**n**	**Pozorováno**	**Predikce**	**Pozorováno**	**Predikce**
1	< -0,739345	65	17	19,5343	48	45,4657
2	-0,739345 až -0,477914	67	25	23,4161	42	43,5839
3	-0,477914 až -0,298337	65	25	26,2976	40	38,7024
4	-0,298337 až -0,0383263	64	32	29,2647	32	34,7353
5	> -0,0383263	64	35	35,1302	29	28,8689
Celkem		325	134		191	

Tab. 5: χ^2 test modelu z dat z let 2001-2005

χ^2 =1,21442 … počet stupňů volnosti = 3 … P_{VALUE} ***= 0,749546***

Jelikož hodnota P_{VALUE} je u obou modelů (*rovnice (1) i rovnice (2)*) větší než ***0,1***, není důvod zamítat vhodnost použitého modelu pravděpodobnosti na 90% hladině významnosti.

6.2.2 χ^2 test deviance koeficientů

Provádí se pro účely testování nulovosti koeficientů. Zkoumáme nulovost koeficientů OS a PS. Bude-li kterýkoliv z nich označen za nulový, nebude výsledný model na tomto koeficientu záviset a bude ho možno z modelu vyloučit.

Testujeme koeficienty OS a PS:

- nulová hypotéza: $H_0: \beta = 0$
- alternativní hypotéza: $H_A : \beta \neq 0$

a) pro data z let 2001-2006 *(viz rovnice (1))*

Zde se $\beta = (\beta_{0S}, \beta_{PS}),\ \beta_{0S} = 0{,}0548604, \quad \beta_{PS} = 0{,}0725761$

χ^2 testy deviance:

OS:

- χ^2 statistika : $\chi^2 = 4{,}20424$
- počet stupňů volnosti = 1
- $P_{VALUE} = 0{,}0403$

PS:

- χ^2 statistika : $\chi^2 = 7{,}80888$
- počet stupňů volnosti = 1
- $P_{VALUE} = 0{,}0052$

b) pro data z let 2001-2005 *(viz rovnice (2))*

Zde se $\beta = (\beta_{0S}, \beta_{PS}),\ \beta_{0S} = 0{,}0702439, \quad \beta_{PS} = 0{,}0637289$

χ^2 testy deviance:

OS:

- χ^2 statistika : $\chi^2 = 6{,}42149$
- počet stupňů volnosti = 1
- $P_{VALUE} = 0{,}0113$

PS:

- χ^2 statistika : $\chi^2 = 5{,}37696$
- počet stupňů volnosti = 1
- $P_{VALUE} = 0{,}0204$

Jelikož vždy u parametrů OS i PS vyšly P_{VALUE} menší než hladina významnosti $\alpha = 0{,}05$, zamítáme ve všech případech nulovou hypotézu H_0 o nulovosti koeficientů. Přijímáme alternativní hypotézu H_A, oba koeficienty v obou rovnicích jsou statisticky významné na 95% konfidenční úrovni a nemůžeme je z modelu odstranit.

Po provedení testů bylo zjištěno, že naše modely jsou kompaktní s daty z Fakultní nemocnice Ostrava-Poruba a můžeme je dál využít.

7 Algoritmické zpracování nových regresních modelů

Pomocí logistické regrese jako nástroje pro diskriminaci jsme sestavili program „***morbidita.m***“ na výpočet predikce morbidity pacientů.

Dle metodiky uvedené v kapitole ***4.3***, kdy regresní rovnice v našem případě vypadá takto:

$$S(x) = \beta_0 + \beta_1 \cdot 0S + \beta_2 \cdot PS$$

Tedy objekt bude zařazen do první skupiny (morbidita=1), pokud $S(\underline{X}) \geq 0$. Do nulté skupiny (morbidita=0), pokud $S(\underline{X}) < 0$.

7.1 Využití modelů pro predikci morbidity

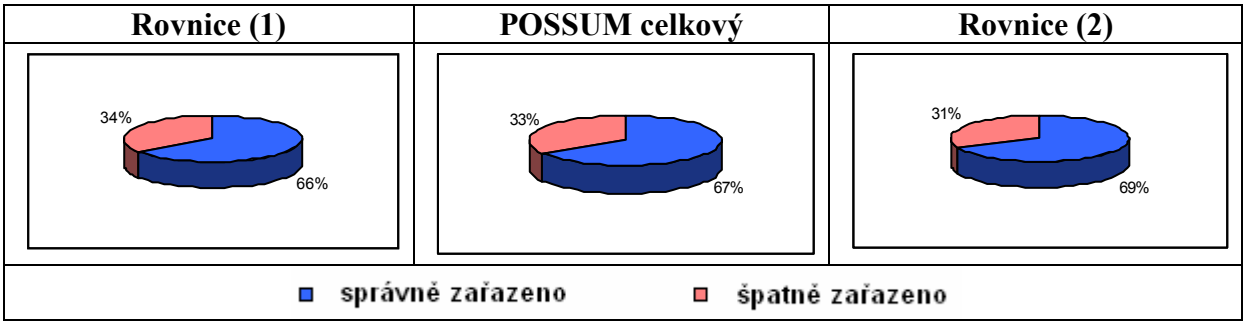
Na nově vytvořené modely použijeme program „***morbidita.m“*** a otestujeme ho na příslušných datech. Výsledky predikce morbidity (0 nebo 1) porovnáme se skutečným sloupcem morbidita v našem datovém souboru „***data_nova.xls***“. Na závěr vypočteme procento chyby modelu. Všechny výsledky jsou přehledně zaznamenány v následujících tabulkách.

Jednotlivé sloupce znamenají:

- ***Model:*** jsou vytvořené modely
- ***Správně zařazeno:*** zde je počet správně zařazených hodnot (0 nebo 1), dle testovacího souboru „data.xls“, tzn. program správně spočítal hodnotu 0 nebo 1 jako byla uvedena v testovacím souboru na stejném místě
- ***Špatně zařazeno:*** zde jsou ty hodnoty, které byly určeny špatně tzn. program spočítal hodnotu 1 (respektive 0) a v testovacím souboru byla uvedena 0 (respektive 1)
- ***Predikovaný počet 1:*** je součet všech 1, které program určil pro pacienty s daným PS a OS
- ***Skutečný počet 1:*** součet všech jedniček v dané skupině podle testovacího souboru
- ***Procento chyby:*** udává jaká je možnost toho, že daný pacient byl zařazen do nesprávné skupiny, vypočte se jako podíl ***Špatně zařazeno / počet všech pacientu ve skupině***

Model	Správně zařazeno	Špatně zařazeno	Skutečný počet 1	Predikovaný počet 1	Procento chyby
Rovnice (1)	240	124	74	88	34%
POSSUM celkový	243	121	74	80	33,2%
Rovnice (2)	250	114	74	64	31,3%

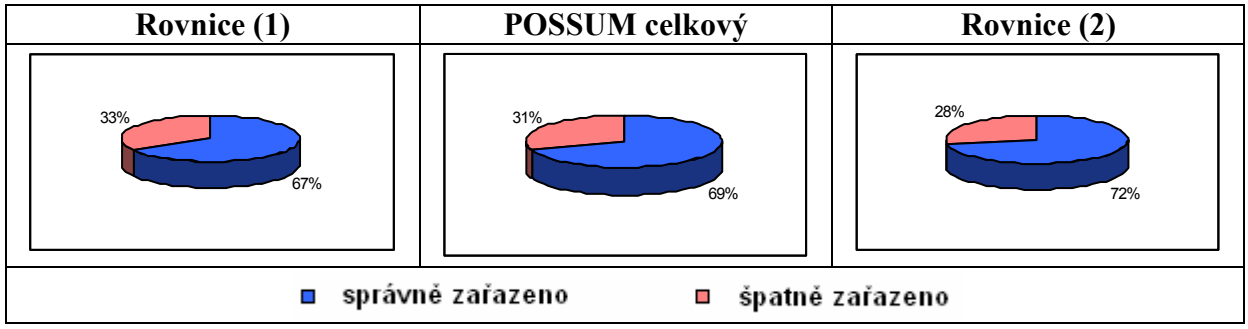
Tab. 6: Modely testovány na všech datech z let 2001-2006 (364)



Obr. 4: Grafické znázornění zařazení pacientů ze všech dat (364)

Model	Správně zařazeno	Špatně zařazeno	Skutečný počet 1	Predikovaný počet 1	Procento chyby
Rovnice (1)	24	12	12	8	33,3%
POSSUM 2006	25	11	12	5	30,5%
Rovnice (2)	26	10	12	2	27,7%

Tab. 7: Modely testovány jen na datech z roku 2006 (36)



Obr. 5: Grafické znázornění zařazení pacientů z roku 2006 (36)

Z výsledků je patrné, že nejefektivnější prognózu morbidity poskytuje model sestavený z dat z let 2001-2005. V obou případech testování u něj vychází nejmenší procento špatného zařazení pacienta.

8 Exponenciální analýza a její modifikace

Nyní data získaná z Fakultní nemocnice Ostrava-Poruba analyzujme ještě pomocí verifikační metody exponenciální analýzy popsané v [8]a vylepšené exponenciální analýzy viz. [9].

Exponenciální analýzu a vylepšenou exponenciální analýzu budeme aplikovat na model POSSUM a na námi vytvořené modely, dané rovnicemi (1) a (2) pro predikci morbidity.

8.1 Exponenciální analýza

Exponenciální analýza byla poprvé provedena v [8] pro účely porovnání očekávané a skutečně se vyskytující morbidity a mortality.

Algoritmus exponenciální analýzy:

1. Jednotlivé pacienty rozdělíme do skupin na základě aktuálních hodnot PS a OS, které vyjádříme v procentech.
2. Celou analýzu provádíme od skupiny 90 – 100 a postupně zvyšujeme interval po deseti procentech až do doby, kdy bude vyhovovat podmínce, aby předpovídaný počet pacientů s pooperačními komplikacemi měl vzrůstající tendenci (nebo byl stejný) vzhledem k předchozí skupině.
3. Analýzu provádíme zdola nahoru.
4. Dojde-li k porušení podmínky, analýza se zastaví a skupina, u které došlo k porušení podmínky se do analýzy nepočítá. Pokračuje se dále analýzou shora dolů.
5. Důležitá je poslední skupina analýzy zdola nahoru u které nedošlo k porušení podmínky. Její dolní hranice určuje horní hranici skupiny u analýzy shora dolů.
6. predikovaný počet = počet pacientů v dané skupině * dolní hranice této skupiny / 100

8.2 Vylepšená exponenciální analýza [8]

Vylepšená exponenciální analýza stejně jako původní exponenciální analýza dělí pacienty do skupin, dle aktuálního PS a OS a provádí jejích analýzu. Ale místo bodu 6 v původním algoritmu je využita věta o úplné pravděpodobnosti.

Označme x modelem předpokládaný počet komplikací ve skupině *(a,b)*, která obsahuje n pacientů a *P(A)* pravděpodobnost, že náhodně vybraný pacient ze skupiny *(a,b)* bude mít komplikace.
Potom:

$$x = n.P(A)$$

Jestliže se ve skupině *(a,b)* vyskytlo r různých pravděpodobností komplikací $p_1, \ldots, p_r$ a n_i je počet pacientů s pravděpodobností komplikace p_i, pak můžeme psát :

$$x = n\left(\sum_{i=1}^{r} p_i \frac{n_i}{n}\right) \quad \text{tedy} \qquad x = \sum_{i=1}^{r} p_i n_i \ .$$

Pokud provedeme přeznačení a každému (i-tému) pacientovi ze skupiny *(a,b)* přiřadíme pravděpodobnost komplikace p_i, pak je zřejmé, že

$$x = \sum_{i=1}^{n} p_i \ .$$

8.3 Aplikace exponenciální a vylepšené exponenciální analýzy

Pomocí programu ***„tabulka1.m“*** porovnáme výsledky jednotlivých analýz a zhodnotíme výsledky.

Jednotlivé slupce tabulky popisují:

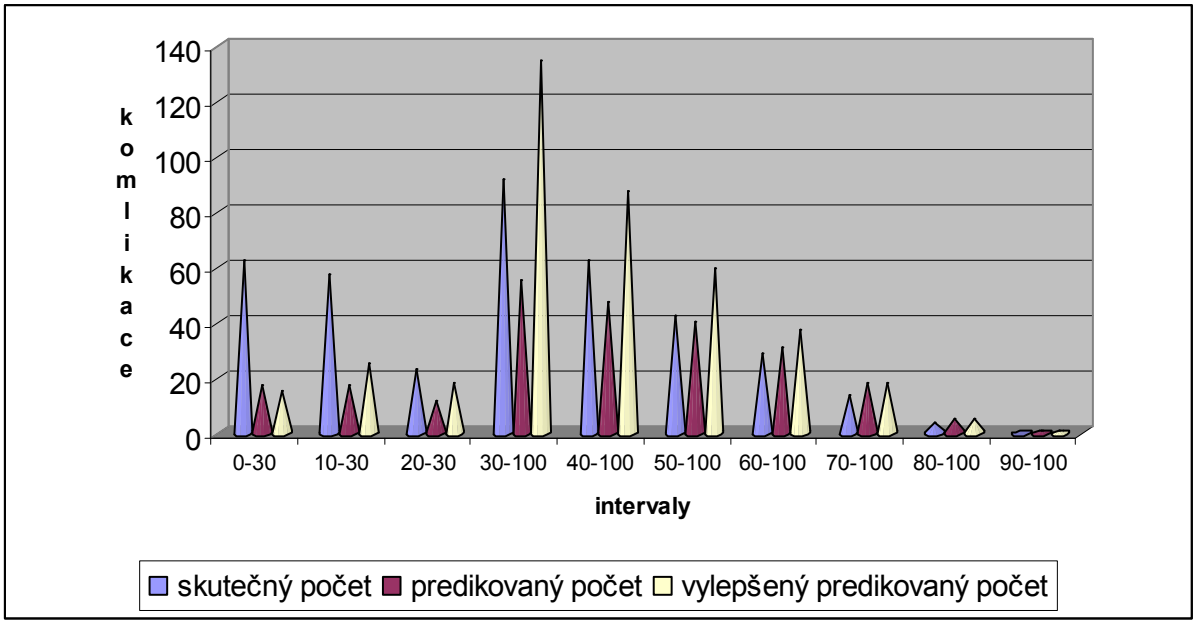
- skupina – udává intervaly pro morbiditu pacientů v procentech
- počet pacientů – udává, kolik pacientů dané skupiny mělo skutečně pooperační komplikace
- předpovězený počet – udává předpovídaný počet pacientů s pooperačními komplikacemi, dle vztahu: ***předpovídaný počet = počet pacientů * dolní hranice skupiny / 100***, až na první skupinu, která začíná hodnotou 0%, jelikož by 0 nemělo smysl násobit určí se predikovaný počet jako medián z dat ve skupině – použijeme střední hodnotu vypočítaného rizika předpovězený počet u 10-ti % pásma a touto hodnotou budeme násobit počet pacientů
- poměr – udává poměr mezi skutečným a předpovídaným počtem pacientů, kteří měli pooperační komplikace
- předpovězený počet modifikovaný - udává předpovídaný počet pacientů s pooperačními komplikacemi, dle vztahu: $x = \sum_{i=1}^{r} p_i n_i$
- poměr modifikovaný - udává poměr mezi skutečným a předpovídaným počtem pacientů, kteří měli pooperační komplikace, dle modifikované exponenciální analýzy

1. Model POSSUM

$$M - POSSUM \qquad \ln\left(\frac{R}{1-R}\right) = -5{,}91 + (0{,}16 \cdot PS) + (0{,}19 \cdot OS)$$

skupina (%)	počet pacientů	skutečný počet	předpovězený počet	poměr	předpovězený počet vylepšený	poměr vylepšený
0 – 30	181	62	17	3.6471	15	4.1333
10 - 30	165	57	17	3.3529	25	2.28
20 - 30	60	23	12	1.9167	18	1.2778
30 - 100	183	91	55	1.6545	134	0.6791
40 - 100	118	62	47	1.3191	87	0.7126
50 - 100	80	42	40	1.05	59	0.7119
60 - 100	51	29	31	0.9355	37	0.7838
70 - 100	25	14	18	0.7778	18	0.7778
80 - 100	6	4	5	0.8	5	0.8
90 - 100	1	0	1	0	1	0

Tab. 8: Exponenciální analýza a vylepšená exponenciální analýza využitím modelu M-POSSUM



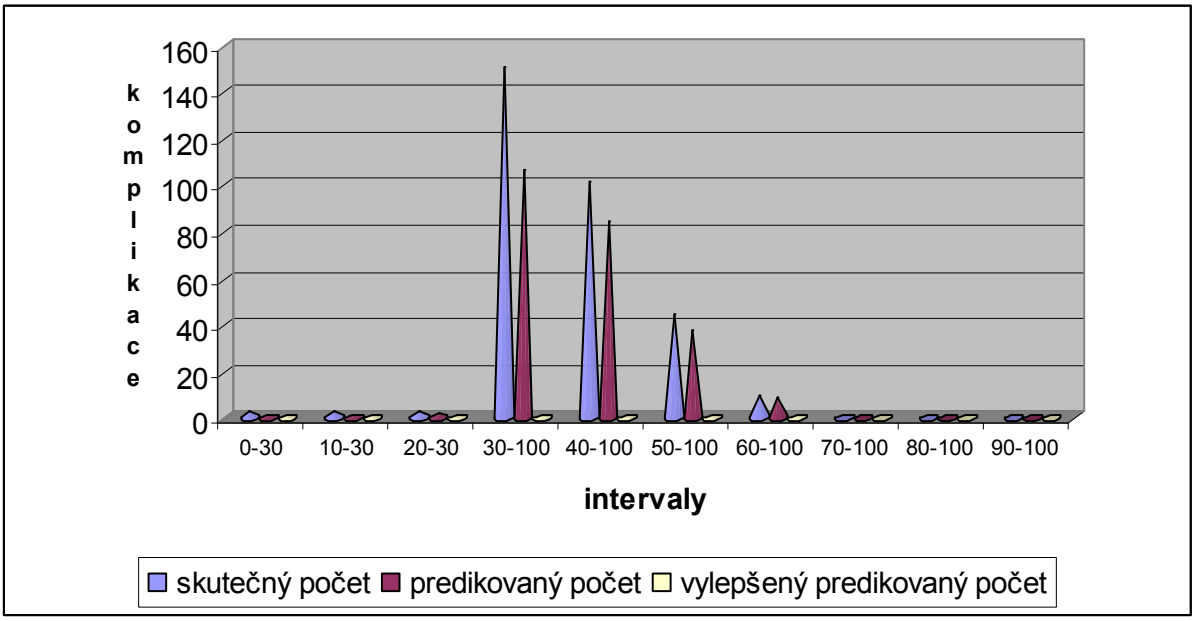
Obr. 6: Znázornění počtu komplikací v modelu M-POSSUM

2. Rovnice (1)

$$\ln\left(\frac{R}{1-R}\right) = -2{,}19973 + 0{,}0548604 * OS + 0{,}0725761 * PS$$

skupina (%)	počet pacientů	skutečný počet	předpovězený počet	poměr	předpovězený počet vylepšený	poměr vylepšený
0 – 30	11	3	1	3	3	1
10 - 30	11	3	1	3	3	1
20 - 30	11	3	2	1.5	3	1
30 - 100	353	150	106	1.4151	119	1.2605
40 - 100	209	101	84	1.2024	116	0.8707
50 - 100	76	44	38	1.1579	42	1.0476
60 - 100	15	10	9	1.1111	10	1
70 - 100	0	0	0	NaN	0	NaN
80 - 100	0	0	0	NaN	0	NaN
90 - 100	0	0	0	NaN	0	NaN

Tab 9: Exponenciální analýza a vylepšená exponenciální analýza s využitím rovnice z dat 2001- 2006



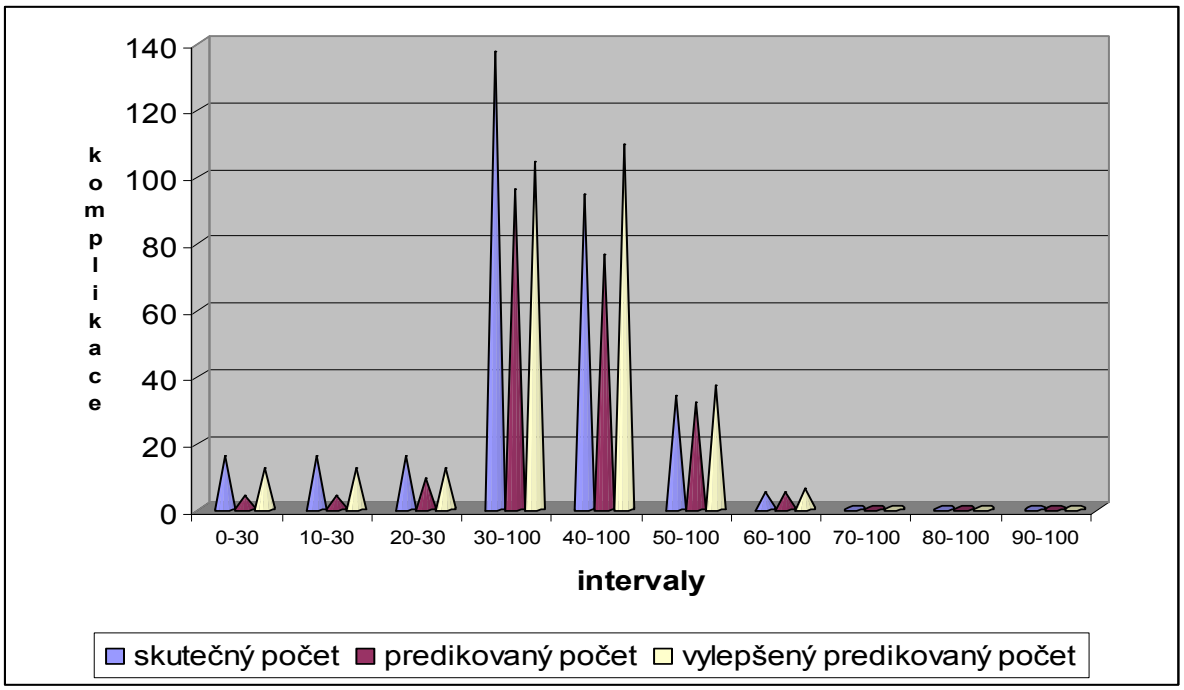
Obr. 7: Znázornění počtu komplikací v modelu z let 2001-2006

3. Rovnice (2)

$$\ln\left(\frac{R}{1-R}\right) = -2{,}33399 + 0{,}0702439 * OS + 0{,}0637289 * PS$$

skupina (%)	počet pacientů	skutečný počet	předpovězený počet	poměr	předpovězený počet vylepšený	poměr vylepšený
0 – 30	44	16	4	4	12	1.3333
10 - 30	44	16	4	4	12	1.3333
20 - 30	44	16	9	1.7778	12	1.3333
30 - 100	320	137	96	1.4271	104	1.3173
40 - 100	189	94	76	1.2368	109	0.8624
50 - 100	64	34	32	1.0625	37	0.9189
60 - 100	9	5	5	1	6	0.8333
70 - 100	0	0	0	NaN	0	NaN
80 - 100	0	0	0	NaN	0	NaN
90 - 100	0	0	0	NaN	0	NaN

Tab. 10: Exponenciální analýza a vylepšená exponenciální analýza s využitím rovnice z dat 2001- 2005



Obr. 8: Znázornění počtu komplikací v modelu z let 2001-2005

Z dosažených výsledků, reprezentovaných tabulkami, je viditelné, že vylepšená exponenciální analýza poskytuje reálnější výsledky než-li původní. Výsledné poměry jsou blízké číslu 1, což je pro nás ideální. Jsou-li poměry vyšší něž jedna, znamená to, že predikce je příliš optimistická, tedy počet predikovaných jedniček je výrazně menší než skutečnost. Naopak, jsou-li poměry výrazně nižší než jedna, predikce předpovídá více pacientů s komplikacemi než je skutečnost.

9 Závěr

Cílem práce bylo navrhnout algoritmus rozhodovacího procesu pro zařazení objektu, s neznámou příslušností, do jedné ze dvou skupin pomocí logistické regrese. Pomocí programu Statgraphics byly vytvořeny nové logistické modely a ty testovány.

Testování bylo prováděno pomocí logistické regrese jako nástroje pro diskriminaci a pomocí exponenciální analýzy a její modifikace.

Testováním na základě referenčního postupu zvaného exponenciální analýza bylo zjištěno, že v literatuře uváděný model POSSUM není nejvhodnější pro predikci morbidity při těchto operacích. Podobné výsledky prognózy přinesl model sestavený ze všech dat z let 2001-2006. Nejefektivnější výsledky prognózy s nejmenší chybou finálně přinesl až model sestavený z dat z let 2001-2005.

Z výsledků je patrné, že chirurgická data pocházející z roku 2006, byla zřejmě získána po aplikaci inovativních postupů. Model sebou nesl velké odlišnosti způsobené vývojem stanovení veličin OS, PS a morbidita. Nejlepším se ukázal model z dat z let 2001-2005 a jeho testování na datech z roku 2006. Díky dostatku nasbíraných dat se stal tento model vyhovujícím všem použitým standardním statistickým testům modelu. Navíc vyhověl i poněkud netradiční verifikaci dané exponenciální analýzou. Na datech z roku 2006 ukázal chybu pouhých 27%. Výsledný model má tvar:

$$\ln\left(\frac{R}{1-R}\right) = -2{,}33399 + 0{,}0702439 * OS + 0{,}0637289 * PS$$

Jednotlivé koeficienty byly nalezeny metodou vícerozměrné metody maximální věrohodnosti. Důležitou součástí modelu je také testování na nenulovost koeficientů. Toto bylo provedeno testem chí-kvadrát. Dle tohoto modelu můžeme s velkou pravděpodobností predikovat ohodnocení morbidity pacientů, pouze s využitím dostupných parametrů OS a PS. Tedy zda pacienti po provedení operace budou mít komplikace či nikoliv.

Exponenciální analýza dat ukazuje poměr mezi předpovězeným a reálným počtem pacientů s komplikacemi. Zde je vidět, že modifikovaná exponenciální analýza, která byla v rámci této práce zalgoritmizována, mnohem přesněji a reálněji popisuje skutečnost něž-li původní. Zatímco původní exponenciální analýza je při vyhodnocování počtu pacientů s komplikacemi velmi optimistická (tedy určí jich zpravidla méně než jich ve skutečnosti je). Modifikovaná exponenciální analýza dává výsledky mnohem reálnější (výsledný poměr je velmi blízký číslu 1).

Při porovnání modelů je patrné, že hodnoty jednotlivých koeficientů u námi vytvořených modelů se příliš neliší, ale model POSSUM má hodnoty velmi odlišné.Důvodem je odlišnost dat. Naše modely byly vytvořeny na základě konkrétního datového souboru z Fakultní nemocnice Ostrava-Poruba. Původní model byl vytvořen na základě dat získaných po operacích v anglických nemocnicích.

Pavlína Kuráňová

Seznam použité literatury

[1] L. Martínek: Aplikace specializovaných skórovacích systémů pro objektivizaci morbidity laparoskopických operací kolorekta, OSTRAVA 2006

[2] Copeland, G.P., Jones, D., Walters, M.: *POSSUM: a scoring system for surgical audit*. Br. J. Surg., 1991

[3] P.Hebák, J. Hustopecký, E. Jarošová, I. Pecáková: Vícerozměrné statistické metody (1), Praha 2004

[4] R.Briš, M.Litschmannová: STATISTIKA II., OSTRAVA 2007

[5] R.Briš: MODELOVÁNÍ RIZIKA MORBIDITY LAPAROSKOPICKÝCH OPERACÍ POMOCÍ LOGISTICKÉ REGRESE

[6] R.Briš, M.Litschmannová: STATISTIKA I. pro kombinované a distanční studium, OSTRAVA 2004

[7] J.Novovičová: Pravděpodobnost a matematická statistika, skriptum 2006

[8] Wijesinghe, L.D., Mahmood, T., Scott, D.J.A., Berridge, D.C., Kent, P.J., Kester, R.C.: Comparison of POSSUM and the Portsmouth predictor equation for predicting death following vascular surgery, 1998

[9] P.Jahoda: Zpřesněná exponenciální analýza, OSTRAVA

Použitý software: MatLab, Excel, Word, Statgraphics.

Seznam příloh

A Příloha
CD se zdrojovými kódy