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

Aplikace logistické regrese v chirurgii

The application of logistic regression in surgery

2012

Dušan Polek

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

V Ostravě 4. května 2012 . . . .

Rád bych poděkoval vedoucímu bakalářské práce RNDr. Pavlu Jahodovi, Ph.D. za čas, který věnoval k poskytování mi spousty cenných rad a informací.

Abstrakt

Práce se zabývá využitím logistické regrese jakožto nástroje pro zhodnocení a hlavně srovnání laparoskopické a otevřené metody operace střev. Otázkou je, která z těchto dvou metod je pro pacienta z hlediska pravděpodobnosti výskytu pooperačních komplikací tou lepší. Pomocí logistické regrese můžeme odhadnout pravděpodobnost výskytu pooperačních komplikací pro jednotlivé metody a následně vybrat optimální operační techniku. Tato volba pak závisí na konkrétním modelu. Pro jeho vytvoření jsou využita data shromážděná za období let 2001 - 2009 na chirurgické klinice Fakultní nemocnice Ostrava.

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

Abstract

This thesis is engaged in the use of logistic regression as a tool for evaluation and mainly comparison of laparoscopic and open way of bowel operations. The question is which way is better for the patient according to the likelihood of postoperative complications. Using logistic regression, we can estimate the probability of postoperative complications for each of these ways, which allows us to choose the optimal operational technique. This choise depends on the specific model. Data collected in the Surgery Clinic of the University Hospital Ostrava in years 2001 - 2009 were used to create these models.

Keywords: logistic regression, morbidity, medical data, maximum likelihood estimation.

Seznam použitých zkratek a symbolů

SE	–	Standard Error
χ^2	–	chí-kvadrát
K	–	Nezávislá proměnná Kardiální příznaky
L	–	Nezávislá proměnná Leukocyty
Z	–	Nezávislá proměnná Závažnost operace
POSSUM	–	Physiological and Operative Severity for enUmeration of Mortality and morbidity
PS	–	Fyziologické skóre
OS	–	Operační skóre

Obsah

Seznam tabulek

Seznam obrázků

1 Úvod

Logistická regrese je nástrojem sloužícím k popisování vztahů mezi výstupní proměnnou a jednou či více vstupními proměnnými. Cílem této modelující techniky je najít nejvhodnější model popisující vliv vstupních (nezávislých) proměnných (tzv. regresorů) na výstupní (závislou) proměnnou. Výstupní proměnná bývá často diskrétní, tj. nabývá dvou a více možných hodnot. Obor hodnot výstupních proměnných logistické regrese je dvouprvková množina, při výpočtech předpokládáme, že jde o proměnnou binární. To jest, že nabývá hodnot 0, nebo 1. Hodnota 0 říká, že pozorovaný jev nenastal, a hodnota 1 naopak, že nastal.

Dnes je logistická regrese v praxi často používána, například v bankovnictví, marketingu, či právě zdravotnictví, pro jehož potřebu je využita i v této práci. Implementována je pak v řadě statistických software, jako je například mnou pro tuto práci použitý *Statgraphics Plus 5.0.*

V této práci se zaměřím na využití logistické regrese jakožto nástroje pro zhodnocení a hlavně srovnání dvou operačních metod operace střev - laparoskopické a otevřené. Otázkou je, která z těchto dvou metod je pro pacienta z hlediska pravděpodobnosti výskytu pooperačních komplikací tou lepší. Pomocí logistické regrese můžeme odhadnout pravděpodobnost výskytu pooperačních komplikací pro jednotlivé metody a následně vybrat optimální operační techniku. Tato volba pak závisí na konkrétním modelu. Pro jeho vytvoření využiji data shromážděná za období let 2001 - 2009 na chirurgické klinice Fakultní nemocnice Ostrava (více informací o těchto datech viz kapitola 3.1).

Hlavním cílem pak není určit, zda je jedna či druhá operační technika lepší obecně, ale zda je lepší pro daného pacienta. Pomocí tohoto přístupu by se měla snížit pooperační morbidita u obou těchto metod.

2 Model logistické regrese

2.1 Popis modelu

Uvažujme proměnnou Y, jež nabývá pouze hodnoty 1 (jev nastal) a nebo 0 (jev nenastal). Zajímá nás, jak ovlivňují hodnoty proměnné X hodnoty proměnné Y. Proměnnou Y označme jako vysvětlovanou proměnnou a X jako vysvětlující proměnnou, nebo též regresor. Označme $\pi(x) = E(Y \mid x)$ střední hodnotu veličiny Y za předpokladu, že $X = x$. Díky tomu, že veličina Y nabývá pouze hodnot 1, nebo 0, má $\pi(x)$ též význam pravděpodobnosti, že věličina Y nabude hodnoty 1 (to jest pravděpodobnosti, že sledovaný jev nastane) při $X = x$. Plyne to okamžitě z následující rovnosti

$$\pi(x) = E(Y \mid x) = \sum_i y_i P(Y = y_i \mid x) = 1 \cdot P(Y = 1 \mid x) + 0 \cdot P(Y = 0 \mid x) = P(Y = 1 \mid x).$$

Závislost $\pi(x)$ na hodnotě x hledáme v případě logistické regrese ve tvaru

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}, \tag{1}$$

kde β_0 a β_1 jsou takzvané regresní koeficienty. Jejich číselnou hodnotu odhadneme na základě naměřených hodnot veličin X a Y metodou maximální věrohodnosti (viz sekce 2.3). Všimněme si, že při takto zvolené závislosti platí $\pi(x) \in (0, 1)$.

Logitovou transformací funkce $\pi(x)$ se nazývá funkce

$$g(x) = \beta_0 + \beta_1 x$$

a snadno se odvodí, že

$$g(x) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right]$$

Funkce $g(x)$ má mnohé žádoucí vlastnosti lineárního regresního modelu - jde o lineární funkci, může být spojitá a může nabývat hodnot na intervalu $(-\infty; \infty)$ v závislosti na x. Při výše uvedeném značení pak platí

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

2.2 Chyba modelu

Předpokládejme, že jsme naměřili hodnotu (x, y) náhodného vektoru (X, Y). Naměřenou hodnotu y proměnné Y můžeme vyjádřit jako

$$y = E(Y \mid x) + \varepsilon_x = \pi(x) + \varepsilon_x, \tag{2}$$

kde ε_x je chyba modelu při $X = x$. Může nabývat dvou hodnot:

- Pokud je $y = 1$, pak $\varepsilon_x = 1 - \pi(x)$. To nastane s pravděpodobností $\pi(x)$.

- A pokud je $y = 0$, pak $\varepsilon_x = -\pi(x)$. Tato možnost nastane s pravděpodobností $1 - \pi(x)$.

Náhodná proměnná ε_x má proto binomické rozdělení $B(1, \pi(x))$. Střední hodnota chyby modelu při daném $x \in R$ je rovna

$$E(\varepsilon_x) = (1 - \pi(x))\pi(x) + (-\pi(x))(1 - \pi(x)) = 0$$

a rozptyl je při daném $x \in R$ roven

$$D(\varepsilon_x) = \pi(x)[1 - \pi(x)].$$

2.3 Metoda maximální věrohodnosti

Obecně poskytuje metoda maximální věrohodnosti hodnoty pro neznámé parametry, které maximalizují pravděpodobnost obdržení pozorovaného souboru dat. Abychom použili tuto metodu, musíme nejdříve sestavit věrohodnostní funkci.

V našem případě hledáme hodnoty koeficientů β_0 a β_1.

Jak jsme viděli výše, pokud je Y kódováno hodnotami 0 nebo 1, pak výraz $\pi(x)$ vyjadřuje podmíněnou pravděpodobnost $P(Y = 1|x)$. Proto $P(Y = 0|x) = 1 - \pi(x)$. Předpokládejme, že jsme naměřili hodnoty (x_i, y_i) náhodného vektoru (X, Y), kde $i = 1, 2, \ldots, n$. Uvažujme, s jakou pravděpodobností naměříme právě hodnotu y_i při daném x_i.

- $y_i = 1$ s pravděpodobností $\pi(x_i)$
- $y_i = 0$ s pravděpodobností $1 - \pi(x_i)$

Proto výraz

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \tag{3}$$

vyjadřuje právě pravděpodobnost, že $Y = y_i$ při $X = x_i$. Za předpokladu, že jsou jednotlivá měření (pozorování) nezávislá, je pravděpodobnost, že jsme naměřili právě hodnoty y_i, kde $i = 1, 2, \ldots, n$ při daných x_i rovna součinu

$$\prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \tag{4}$$

Podle rovnice (1) je

$$\pi(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}.$$

Vidíme, že hodnota $\pi(x_i)$ závisí na hodnotách β_0 a β_1. Proto hodnota (4) také závisí na β_0 a β_1. Označme

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}, \tag{5}$$

kde $\boldsymbol{\beta} = (\beta_0, \beta_1)$ je vektor neznámých regresních koeficientů. Funkci l nazýváme věrohodnostní funkcí pro logistickou regresi.

Tato funkce vyjadřuje pravděpodobnost, že naměříme hodnoty y_i právě tak, jak jsme je naměřili. Proto je přirozené hledat β_0, β_1 takové, aby hodnota $l(\boldsymbol{\beta})$ byla maximální.

Matematicky je však jednodušší pracovat s logaritmem funkce l (uvažme, že l nabývá pouze kladných hodnot). Usnadní nám to derivování při hledání maxima funkce. Díky monotónii logaritmu nabudou funkce l a $\ln l$ maxima při stejných hodnotách β_0 a β_1.

$$L(\boldsymbol{\beta}) = \ln\left[l(\boldsymbol{\beta})\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1 - y_i) \ln\left[1 - \pi(x_i)\right] \right\} \tag{6}$$

Abychom nalezli hodnotu $\boldsymbol{\beta}$, která maximalizuje $L(\boldsymbol{\beta})$, do tohoto vztahu dosadíme $\pi(x_i)$ podle rovnice (1) a následně derivujeme $L(\boldsymbol{\beta})$ dle β_0:

$$\begin{aligned}\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} &= \sum_{i=1}^{n} \left[y_i \cdot \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} + (1 - y_i) \cdot \left(- \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) \right] = \\ &= \sum_{i=1}^{n} \left[y_i \cdot \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} + y_i \cdot \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right] = \\ &= \sum_{i=1}^{n} \left[y_i - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right],\end{aligned}$$

po zpětném dosazení $\pi(x_i)$ podle rovnice (1) tedy dostáváme, že

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} \left[y_i - \pi(x_i) \right], \tag{7}$$

a dle β_1:

$$\begin{aligned}\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} &= \sum_{i=1}^{n} \left[y_i \cdot x_i \cdot \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} + (1 - y_i) \cdot x_i \cdot \left(- \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) \right] = \\ &= \sum_{i=1}^{n} \left[y_i \cdot x_i \cdot \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} - x_i \cdot \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} + y_i \cdot x_i \cdot \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right] = \\ &= \sum_{i=1}^{n} x_i \cdot \left[y_i - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right],\end{aligned}$$

po zpětném dosazení $\pi(x_i)$ podle rovnice (1) tedy dostáváme, že

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1} = \sum_{i=1}^{n} x_i \left[y_i - \pi(x_i) \right]. \tag{8}$$

Výrazy v rovnicích (7) a (8) položíme rovny nule, čímž získáme soustavu tzv. věrohodnostních rovnic:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \tag{9}$$

$$\sum_{i=1}^{n} x_i \, [y_i - \pi(x_i)] = 0 \tag{10}$$

Jejich řešením, např. iteračními metodami, získáme $\hat{\boldsymbol{\beta}}$, tedy odhad maximální věrohodnosti parametru $\boldsymbol{\beta}$.

Zajímavým a užitečným důsledkem rovnice (9), kterého také budeme dále využívat, je, že:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i), \tag{11}$$

kde $\hat{\pi}(x_i)$ je odhad maximální věrohodnosti $\pi(x_i)$. To znamená, že součet pozorovaných hodnot $\boldsymbol{y}$ je roven součtu hodnot očekávaných.

2.4 Testování hypotéz

Po odhadnutí koeficientů modelu přichází obvykle na řadu posouzení významnosti jeho proměnných. To obvykle zahrnuje formulaci a testování statistických hypotéz k určení, zda nezávislé proměnné v modelu významně souvisí s výstupní proměnnou. Významnou pak můžeme označit takovou proměnnou, která svou přítomností v modelu nabízí lepší očekávané hodnoty, než model, ve kterém tato proměnná obsažená není.

2.4.1 Test poměru věrohodností

Pro tento účel se obvykle používá tzv. věrohodnostní poměr D (z angl. *deviance*), definovaný jako

$$D = -2 \sum_{i=1}^{n} \left[y_i \ln \left(\frac{\hat{\pi}(x_i)}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \hat{\pi}(x_i)}{1 - y_i} \right) \right], \tag{12}$$

kde $\hat{\pi}(x_i)$ je odhad modelu pro y_i. Součet na pravé straně rovnosti představuje poměr věrohodnosti modelu a tzv. *saturovaného modelu*, tj. hodnot vypočtených na základě naměřených dat.

$$D = -2 \sum_{i=1}^{n} \left[\ln \left(\frac{\hat{\pi}(x_i)}{y_i} \right)^{y_i} \cdot \left(\frac{1 - \hat{\pi}(x_i)}{1 - y_i} \right)^{(1-y_i)} \right] =$$
$$= -2 \ln \prod_{i=1}^{n} \frac{\hat{\pi}(x_i)^{y_i} [1 - \hat{\pi}(x_i)]^{1-y_i}}{y_i^{y_i} (1 - y_i)^{1-y_i}},$$

Označme věrohodnost modelu jako

$$l(model) = \prod_{i=1}^{n} \hat{\pi}(x_i)^{y_i}[1 - \hat{\pi}(x_i)]^{1-y_i}$$

a věrohodnost *saturovaného modelu* jako

$$l(sat.\ model) = \prod_{i=1}^{n} \hat{\pi}(x_i)^{y_i}[1 - \hat{\pi}(x_i)]^{1-y_i}.$$

Pak

$$D = -2 \ln \left[\frac{l(model)}{l(sat.\ model)} \right]$$

Vzhledem k tomu, že hodnoty výstupní proměnné nabývají pouze hodnot 0 nebo 1, pak je věrohodnost *saturovaného modelu* rovna 1,

$$l(sat.\ model) = 1.$$

Odtud

$$D = -2 \ln \left[l(model) \right],$$

tedy

$$D = -2 \ln \prod_{i=1}^{n} \hat{\pi}(x_i)^{y_i}[1 - \hat{\pi}(x_i)]^{1-y_i}.$$

Abychom ohodnotili významnost nezávislé proměnné, porovnáme hodnotu D vypočtenou pro model, který nezávislou proměnnou v rovnici obsahuje, a který ji neobsahuje (tj. kde $\beta_1 = 0$).

G = D(model bez proměnné) - D(model s proměnnou)

$$G = -2 \left[\ln l(model\ bez\ x) - \ln l(model\ s\ x) \right] \tag{13}$$

$$G = -2 \left[\ln \frac{l(model\ bez\ x)}{l(model\ s\ x)} \right] \tag{14}$$

kde $l(model\ bez\ x)$ je věrohodnost modelu, který proměnnou x obsahuje, a $l(model\ bez\ x)$ je věrohodnost modelu, který ji neobsahuje (tj. $\beta_1 = 0$).

Jestliže $\beta_1 = 0$, pak

$$\hat{\pi}(x_i) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}.$$

Při řešení věrohodnostní rovnice (9), resp. (11), obdržíme vztah

$$\sum_{i=1}^{n}\left(y_i - \frac{e^{\beta_0}}{1+e^{\beta_0}}\right) = 0$$

$$\sum_{i=1}^{n}\left[y_i - \left(\frac{1+e^{\beta_0}-1}{1+e^{\beta_0}}\right)\right] = 0$$

$$\sum_{i=1}^{n}\left(y_i - 1 + \frac{1}{1+e^{\beta_0}}\right) = 0$$

$$\sum_{i=1}^{n}(y_i - 1) + n \cdot \frac{1}{1+e^{\beta_0}} = 0$$

$$\sum_{i=1}^{n}(y_i - 1) = -n \cdot \frac{1}{1+e^{\beta_0}} \quad / \cdot (-1)$$

$$\sum_{i=1}^{n}(1 - y_i) = n \cdot \frac{1}{1+e^{\beta_0}}$$

$$1 + e^{\beta_0} = \frac{n}{\sum_{i=1}^{n}(1-y_i)}$$

$$e^{\beta_0} = \frac{n}{\sum_{i=1}^{n}(1-y_i)} - 1$$

$$e^{\beta_0} = \frac{n - \sum_{i=1}^{n}(1-y_i)}{\sum_{i=1}^{n}(1-y_i)}$$

$$e^{\beta_0} = \frac{n - n\sum_{i=1}^{n}(y_i)}{\sum_{i=1}^{n}(1-y_i)}$$

$$e^{\beta_0} = \frac{\sum_{i=1}^{n}(y_i)}{\sum_{i=1}^{n}(1-y_i)}$$

$$\beta_0 = \ln\frac{\sum_{i=1}^{n}(y_i)}{\sum_{i=1}^{n}(1-y_i)}$$

Označíme-li

$$n_1 = \sum_{i=1}^{n}(y_i)$$

počet případů, kdy sledovaný jev nastal, a

$$n_0 = \sum_{i=1}^{n}(1-y_i)$$

počet případů, kdy sledovaný jev nenastal, pak

$$\beta_0 = \ln \frac{n_1}{n_0}.$$

A tedy

$$\hat{\pi}(x_i) = \frac{e^{\beta_0}}{1+e^{\beta_0}} = \frac{\frac{n_1}{n_0}}{1+\frac{n_1}{n_0}} = \frac{n_1}{n_0+n_1} = \frac{n_1}{n}.$$

Pak je hodnota G dána vztahem

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}(x_i)^{y_i}[1-\hat{\pi}(x_i)]^{1-y_i}}\right] \tag{15}$$

nebo

$$G = 2\ln\left\{\sum_{i=1}^{n}[y_I \ln(\hat{\pi}(x_i)) + (1-y_i)\ln(1-\hat{\pi}(x_i))] - [n_1\ln(n_1) + n_0\ln(n_0) - n\ln(n)]\right\}, \tag{16}$$

náhodná veličina G má χ^2 rozdělení s 1 stupněm volnosti.

2.4.2 χ^2 test dobré shody

Naši populaci si roztřídíme do k disjunktních skupin (tzv. variant). χ^2 test dobré shody (angl. goodness-of-fit test) je test, který ověřuje, zda se empirické (pozorované) četnosti O_i jednotlivých variant náhodné veličiny shodují s očekávanými četnostmi E_i. Testujeme nulovou hypotézu

H_0: *Teoretické a empirické rozdělení se shoduje.*

při alternativní hypotéze

H_A: *Teoretické a empirické rozdělení se neshoduje.*

na hladině významnosti α. To, jak dobře se pozorované a očekávané četnosti shodují, nám udá rozdíl těchto četností. Je-li nulová hypotéza pravdivá, měly by být pozorované a očekávané četnosti jednotlivých variant přibližně stejné a tedy by měl tento rozdíl mít malou hodnotu.

To ověříme pomocí náhodné veličiny

$$G = \sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i}, \tag{17}$$

která má v případě platnosti nulové hypotézy a při splnění předpokladu, že provádíme dostatečně velký výběr (tj. takový výběr, kde všechny očekávané četnosti E_i jsou větší než 5), přibližně χ^2 rozdělení s $k-1$ stupni volnosti.

Je-li uvedený předpoklad splněn, pak

$$p - hodnota = 1 - F_0(x_{OBS}), \tag{18}$$

kde $F_0(x)$ je distribuční funkce χ^2 rozdělení s $k-1$ stupni volnosti.

2.5 Model vícerozměrné logistické regrese

Doposud jsme uvažovali pouze model s jednou vstupní proměnnou. Nyní si představíme (a modelem vícerozměrné logistické regrese budeme rozumět) takový model, který bude záviset na dvou a více vstupních proměnných.

2.5.1 Popis modelu vícerozměrné logistické regrese

Uvažujme vektor $\boldsymbol{x}' = (x_1, x_2, ..., x_p)$ představující soubor p nezávislých proměnných, tzv. regresorů. Podmíněnou pravděpodobnost, že výstupní proměnná $Y = 1$, pozorovaný jev nastal, tedy označíme $P(Y = 1|\boldsymbol{x}) = \pi(\boldsymbol{x})$. Logit pro model logistické regrese s více vstupními proměnnými je dán rovnicí

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{19}$$

a samotný model tedy bude mít tvar

$$\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1 + e^{g(\boldsymbol{x})}}. \tag{20}$$

2.5.2 Metoda maximální věrohodnosti pro model vícerozměrné logistické regrese

Předpokládejme, že máme vzorek n nezávislých pozorování $(\boldsymbol{x_i}, y_i)$, kde $i = 1, 2, ..., n$. Stejně jako v případě s jednou vstupní proměnnou, i tentokrát potřebujeme vypočíst odhad vektoru parametrů $\beta' = (\beta_0, \beta_1, ..., \beta_p)$. A i tentokrát k tomu použijeme metodu maximální věrohodnosti.

Věrohodnostní funkce je téměř shodná s věrohodnostní funkcí logistické regrese s jednou vstupní proměnnou

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}, \tag{21}$$

stejně tak logaritmická věrohodnostní funkce

$$L(\boldsymbol{\beta}) = \ln\left[l(\boldsymbol{\beta})\right] = \sum_{i=1}^{n} \{y_i \ln\left[\pi(x_i)\right] + (1 - y_i) \ln\left[1 - \pi(x_i)\right]\}. \tag{22}$$

Do tohoto vztahu dosadíme $\pi(x_i)$ podle rovnice (20) a derivujeme $L(\boldsymbol{\beta})$ tentokrát dle všech jeho $p + 1$ koeficientů $\beta_0, \beta_1, ..., \beta_p$:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \pi(\boldsymbol{x_i})], \tag{23}$$

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} \, [y_i - \pi(\boldsymbol{x_i})], \tag{24}$$

pro $j = 1, 2, ..., p$.

Tyto výrazy položíme rovny nule, čímž dostáváme soustavu $p+1$ věrohodnostních rovnic:

$$\sum_{i=1}^{n} [y_i - \pi(\boldsymbol{x_i})] = 0 \tag{25}$$

$$\sum_{i=1}^{n} x_{ij} [y_i - \pi(\boldsymbol{x_i})] = 0, \tag{26}$$

pro $j = 1, 2, ..., p$.

Jejich řešením získáme $\hat{\boldsymbol{\beta}}$, tedy odhad maximální věrohodnosti parametru $\boldsymbol{\beta}$.

2.5.3 Testování hypotéz pro model vícerozměrné logistické regrese

Analogicky jako v případě modelu s jednou vstupní proměnnou provedeme test poměru věrohodností, resp. χ^2 test dobré shody (viz kapitola 2.4.1, resp. 2.4.2).

3 Aplikace logistické regrese v chirurgii

Metodu logistické regrese můžeme v chirurgii využit například pro:

1. **Vytvoření modelu rizik pacientů,** kdy na základě dodaných dat můžeme vytvořit model určující např. pravděpodobnost výskytu pooperačních komplikací u konkrétního pacienta, který následně můžeme využít k předpovídání pooperačních komplikací pro budoucí pacienty se stejnými parametry.

2. V případě, že existují (alespoň) dvě možné techniky, jakými lze operaci provést, můžeme využít modelů pro jednotlivé techniky **k volbě optimální operační techniky** tak, že pro konkrétního pacienta zvolíme takovou techniku, aby se minimalizovalo riziko pooperačních komplikací.

3. **Srovnání kvality péče v různých zdravotnických ústavech.** Srovnávat však můžeme pouze pravděpodobnosti pooperačních komplikací u pacientů, kteří mají stejné parametry vstupních proměnných. Kdybychom srovnávali všechny pacienty bez rozdílu, nebylo by srovnání korektní. A to z toho důvodu, že v některých zdravotnických ústavech by mohli provést více náročných operací než v jiných, což by mělo vliv na procentuální výskyt pooperačních komplikací.

4. **Pozorovat a srovnávat vývoj kvality péče v čase.** Toho můžeme docílit tak, že si zvolíme za jednu ze vstupních proměnných například rok, ve kterém byla operace provedena. V případě, že by byl koeficient této proměnné kladný, rostla by úměrně s časem pravděpodobnost pooperačních komplikací. Záporný koeficient by pak naopak naznačoval zlepšování kvality péče v čase.

3.1 Lékařská data

Pro vytvoření modelu rizik pacientů mi byla poskytnuta data shromážděná za období let 2001 - 2009 na chirurgické klinice Fakultní nemocnice Ostrava. K dispozici mi tedy byla databáze více než tisícovky pacientů, kteří na této klinice a v tomto období podstoupili operaci střev. Pacienty jsem si nejprve rozdělil do dvou skupin a to podle zvolené operační techniky - tedy pacienty operované otevřeně (celkem 520 pacientů) a laparoskopicky (celkem 581 pacientů) zvlášť.

Každý pacient má přiřazeny vysvětlující proměnné *kardiální příznaky, leukocyty a závažnost operace* a nezávislou proměnnou *pooperační komplikace.* Oborem hodnot vysvětlujících proměnných je čtyřprvková množina $\{1,\ 2,\ 4,\ 8\}$ a oborem hodnot nezávislé proměnné je samozřejmě dvouprvková množina $\{0,\ 1\}$, kde hodnota 0 znamená, že pacient neměl žádné *pooperační komplikace* a hodnota 1 naopak, že u daného pacienta nějaké *pooperační komplikace* nastaly.

3.2 Vytvoření modelu rizik pacientů

Pro vytvoření modelu rizik pacientů jsem použil metodu logistické regrese (viz kapitola 2) a samotné modely jsem vygeneroval pomocí statistického software *Statgraphics Plus 5.0,* který má tuto metodu implementovánu.

3.2.1 Vytvoření modelu rizik pacientů operovaných otevřenou metodou

Nejprve vytvoříme model, kde je výstupní proměnná závislá na všech třech vstupních proměnných: *kardiální příznaky, leukocyty a závažnost operace*. Model má tvar

$$\pi_o(\boldsymbol{x}) = \frac{e^{-1,32102+0,143822\cdot K+0,150608\cdot L+0,109341\cdot Z}}{1+e^{-1,32102+0,143822\cdot K+0,150608\cdot L+0,109341\cdot Z}}, \tag{27}$$

kde $\boldsymbol{x}' = (K, L, Z)$, kde K značí proměnnou *Kardiální příznaky*, L - *Leukocyty* a Z - *Závažnost operace*.

Nyní použijeme test poměru věrohodností, jehož výstup je zobrazen v tabulce 1.

Proměnná	χ^2	St. volnosti	P-hodnota
K	7,48168	1	0,0062
L	0,894883	1	0,3442
Z	5,93592	1	0,0148

Tabulka 1: Test poměru věrohodností modelu rizik pacientů operovaných otevřenou metodou

Věrohodnostní poměr modelu obsahujícího všechny tři vstupní proměnné a modelu, který neobsahuje proměnnou *Leukocyty*, je roven $0,894883$ a *p-hodnota* je rovna $0,3442$, což znamená, že na hladině významnosti $0,10$ nezamítáme nulovou hypotézu, tedy že proměnná *Leukocyty* není pro tento model statisticky významnou a měli bychom zvážit její odstranění z modelu.

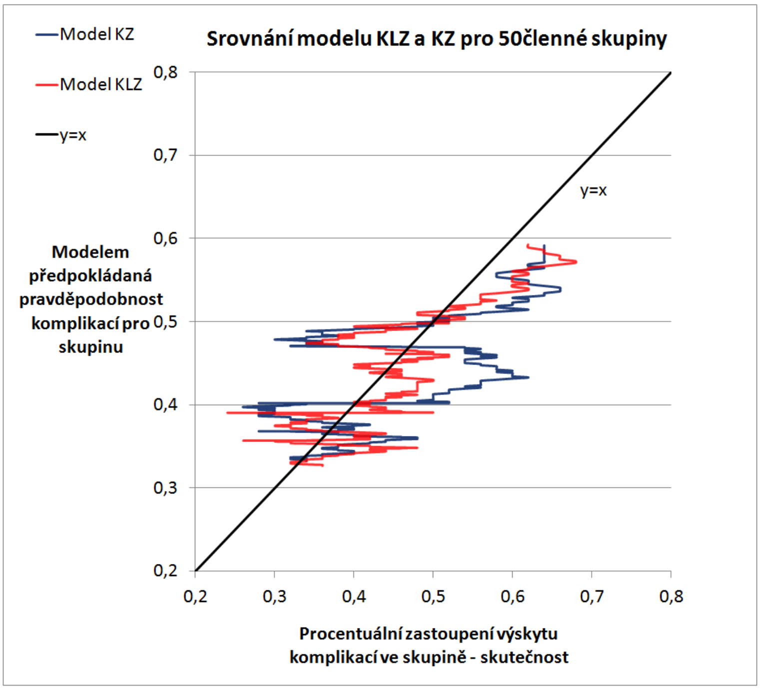
Sestavíme tedy model, kde je naše proměnná závislá pouze na proměnných *Kardiální příznaky* a *Závažnost operace*, které se ukázaly být pro model otevřené metody statisticky významné.

$$\pi(\boldsymbol{x}) = \frac{e^{-1,10138+0,143073\cdot K+0,104175\cdot Z}}{1+e^{-1,10138+0,143073\cdot K+0,104175\cdot Z}} \tag{28}$$

Koeficienty jednotlivých proměnných a konstanta v tomto modelu jsou si docela podobné s původním modelem daným v rovnici (27). To, že se oba modely příliš neliší, ukazuje i graf na obrázku 1. Výraz v rovnici (28) tedy budeme považovat za náš konečný model rizik pacientů operovaných otevřenou metodou.

3.2.2 Vytvoření modelu rizik pacientů operovaných laparoskopickou metodou

I v tomto případě nejprve vytvoříme model, ve kterém je výstupní proměnná závislá na všech třech vstupních proměnných: *kardiální příznaky, leukocyty a závažnost operace*. Model má tvar



Obrázek 1: Srovnání modelů KLZ a KZ pro otevřenou metodu

$$\pi(\boldsymbol{x}) = \frac{e^{-2{,}3921+0{,}107797\cdot K+0{,}292899\cdot L+0{,}281453\cdot Z}}{1+e^{-2{,}3921+0{,}107797\cdot K+0{,}292899\cdot L+0{,}281453\cdot Z}} \tag{29}$$

kde $\boldsymbol{x}' = (K, L, Z)$, kde K značí proměnnou *Kardiální příznaky*, L - *Leukocyty* a Z - *Závažnost operace*.

Nyní přichází na řadu opět test poměru věrohodností, jehož výstup je znázorněn v tabulce 2.

Proměnná	χ^2	St. volnosti	P-hodnota
K	3,67771	1	0,0551
L	2,4124	1	0,1204
Z	37,2847	1	0,0000

Tabulka 2: Test poměru věrohodností modelu rizik pacientů operovaných laparoskopickou metodou

Nejnižší věrohodnostní poměr modelu obsahujícího všechny tři vstupní proměnné a modelu, který jednu z nich neobsahuje, je opět u proměnné *Leukocyty*, $2,4124$. Zároveň má také nejvyšší *p-hodnotu*, ta je rovna $0,1204$, což znamená, že na hladině významnosti $0,10$ nezamítáme nulovou hypotézu, tedy že proměnná *Leukocyty* není i pro tento model statisticky významnou a měli bychom zvážit její odstranění z modelu.

Poměrně malý rozdíl vyšel také pro proměnnou *Kardiální příznaky*. *P-hodnota* je sice vyšší než $0,05$, ale zase nižší než $0,10$. Rozhodnutí, zda je proměnná *Kardiální příznaky* statisticky významná nebo ne, necháme tedy až na test poměru věrohodností pro model, který neobsahuje proměnnou *Leukocyty*.

Model obsahující pouze proměnné *Kardiální příznaky* a *Závažnost operace*, má tvar

$$\pi(\boldsymbol{x}) = \frac{e^{-2,00799+0,107512\cdot K+0,274828\cdot Z}}{1+e^{-2,00799+0,107512\cdot K+0,274828\cdot Z}} \tag{30}$$

Nyní zvážíme, zda je i proměnná *Kardiální příznaky* pro náš model statisticky významná. K tomu opět použije test poměru věrohodností (viz tabulka 3).

Proměnná	χ^2	St. volnosti	P-hodnota
K	3,66597	1	0,0555
Z	35,919	1	0,0000

Tabulka 3: Test poměru věrohodností zjednodušeného modelu rizik pacientů operovaných laparoskopickou metodou

Opět nám vyšlo, že rozhodnutí, zda proměnnou z modelu vyřadit nebo ne, záleží na tom, na jaké hladině významnosti budeme soudit. Protože je *p-hodnota* nižší než $0,10$, je proměnná *Kardiální příznaky* statisticky významná se spolehlivostí 90%.

Koeficienty jednotlivých proměnných a konstanta v tomto modelu jsou si docela podobné s původním modelem daným v rovnici (29). To, jak se oba modely liší, je graficky znázorněno na obrázku 2. Do grafu byl pro ukázku zařazen také model obsahující pouze proměnnou *Závažnost operace*, jehož logistická regresní rovnice má tvar

$$\pi(\boldsymbol{x}) = \frac{e^{-1,76684+0,271886\cdot Z}}{1+e^{-1,76684+0,271886\cdot Z}} \tag{31}$$

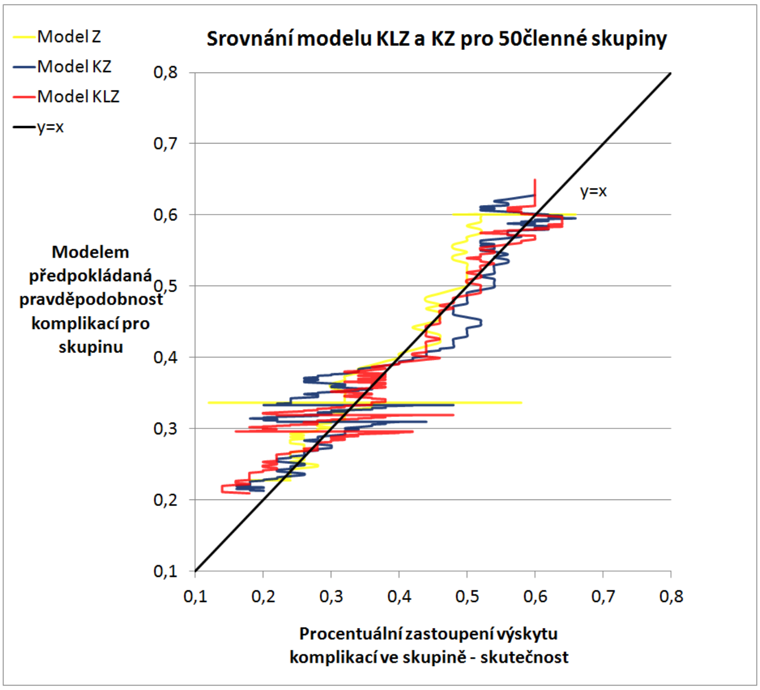
Jako náš konečný model rizik pacientů operovaných laparoskopickou metodou budeme stejně tak jako u modelu metody otevřené operace považovat model obsahující proměnné *Kardinální přiznaky* a *Závažnost operace*, tedy výraz v rovnici (30).

3.3 χ^2 test dobré shody

χ^2 test dobré shody ověřuje hypotézu, zda jsou námi vytvořené logistické regresní funkce [pro metodu otevřenou (28) a laparoskopickou (30)] adekvátní vzhledem k pozorovaným údajům.

Pro tuto kapitolu si označme

- n počet pacientů ($n = n_1 + n_0$),
- n_1 počet pacientů s výskytem pooperačních komplikací,
- n_0 počet pacientů bez pooperačních komplikací



Obrázek 2: Srovnání modelů KLZ, KZ a Z pro laparoskopickou metodu

- $\hat{\pi}_i$ modelem odhadnutou pravděpodobnost, že se u i-tého pacienta objeví pooperační komplikace
- $\hat{n}_1 = \sum_{i=1}^{n} \hat{\pi}_i$ modelem odhadnutý počet komplikací u n pacientů
- $\hat{n}_0 = \sum_{i=1}^{n} (1 - \hat{\pi}_i)$ modelem odhadnutý počet pacientů bez komplikací z n pacientů

3.3.1 χ^2 test dobré shody modelu rizik pacientů operovaných otevřenou metodou

Nyní provedeme χ^2 test dobré shody pro model rizik pacientů operovaných otevřenou metodou.

Nejprve si pacienty rozdělíme do skupin (variant) podle toho, do kterého logitového intervalu přísluší (viz tabulka 4).

Varianta	Logitový interval	n	n_1	$\hat{n}_1$	n_0	$\hat{n}_0$
1	$<$ -0,541527	187	67	66,7704	120	120,23
2	-0,541527 až -0,398454	141	55	56,6374	86	84,3626
3	-0,398454 až -0,112308	95	44	44,3111	51	50,6889
4	$>$ -0,112308	97	55	53,2811	42	43,7189
Celkem		520	221		299	

Tabulka 4: χ^2 test dobré shody modelu rizik pacientů operovaných otevřenou metodou

Při tomto rozdělení do variant vyšel $\chi^2 = 0,207481$ se 2 stupni volnosti. *P-hodnota* $= 0,901459$.

Nyní si pacienty rozdělíme do skupin (variant) podle jejich různých kombinací parametrů, které mají (viz tabulka 5).

Varianta	Kombinace (K,Z)	n	n_1	$\hat{n}_1$
1	(1,1),(1,2),(2,1)	32	8	10,25285
2	(2,2)	33	16	11,64142
3	(1,4)	122	43	44,87329
4	(2,4),(4,2)	148	59	59,57818
5	(1,8)	53	28	24,84713
6	(4,4)	35	12	16,51764
7	(2,8)	52	26	26,23615
8	(8,2)	7	6	3,937913
9	(4,8)	15	10	8,632444
10	(8,4),(8,8)	23	13	14,47131
Celkem		520	221	

Tabulka 5: χ^2 test dobré shody modelu rizik pacientů operovaných otevřenou metodou (2)

V tomto případě vyšel $\chi^2 = 0,301412$ s 9 stupni volnosti. *P-hodnota* $= 0,807915$.

Oba předchozí testy určily, že logistická regresní funkce pro model rizik pacientů operovaných otevřenou metodou sedí na pozorovaná data. Protože je *p-hodnota* vyšší než $0,10$, nemáme žádný důvod zamítnout adekvátnost tohoto modelu (se spolehlivostí 90%).

3.3.2 χ^2 test dobré shody modelu rizik pacientů operovaných laparoskopickou metodou

Nyní provedeme χ^2 test dobré shody pro model rizik pacientů operovaných laparoskopickou metodou.

Nejprve si pacienty rozdělíme do skupin (variant) podle toho, do kterého logitového intervalu přísluší (viz tabulka 6).

Varianta	Logitový interval	n	n_1	$\hat{n}_1$	n_0	$\hat{n}_0$
1	$<$ -0,801165	227	61	63,1395	166	163,86
2	-0,801165 až -0,693653	174	62	57,9804	112	116,02
3	-0,693653 až -0,0485785	63	25	25,5145	38	37,4855
4	$>$ -0,0485785	117	69	70,3655	48	46,6345
Celkem		581	217		364	

Tabulka 6: χ^2 test dobré shody modelu rizik pacientů operovaných laparoskopickou metodou

Při rozdělení do těchto variant vyšel $\chi^2 = 0,602275$ se 2 stupni volnosti. *P-hodnota* $= 0,739976$.

Nyní si pacienty rozdělíme do skupin (variant) podle jejich různých kombinací parametrů, které mají (viz tabulka 7).

Varianta	Kombinace (K,Z)	n	n_1	$\hat{n}_1$
1	(1,2)	30	6	6,172081
2	(2,2),(4,2)	50	10	11,43035
3	(1,4)	147	45	45,53709
4	(2,4)	174	62	57,9804
5	(8,2),(4,4)	48	16	18,19666
6	(8,4)	15	9	7,317853
7	(1,8)	46	26	26,4035
8	(2,8)	53	32	31,80245
9	(4,8),(8,8)	18	11	12,15943
Celkem		581	217	

Tabulka 7: χ^2 test dobré shody modelu rizik pacientů operovaných laparoskopickou metodou (2)

V tomto případě vyšel $\chi^2 = 0,127752$ s 8 stupni volnosti. *P-hodnota* $= 0,996246$.

Oba předchozí testy určují, že logistická regresní funkce pro model rizik pacientů operovaných otevřenou metodou sedí na pozorovaná data. Protože je *p-hodnota* vyšší než $0,10$, nemáme žádný důvod zamítnout adekvátnost tohoto modelu (se spolehlivostí 90%).

3.4 Intervaly spolehlivosti pro pravděpodobnost pooperačních komplikací

Máme skupinu pacientů o N členech, kde se u n pacientů vyskytly pooperační komplikace. Nechť $\hat{\pi}_i$ udává modelem odhadnutou pravděpodobnost, že se u i-tého pacienta objeví pooperační komplikace. Pak číslo

$$\hat{n} = \sum_{i=1}^{N} \hat{\pi}_i$$

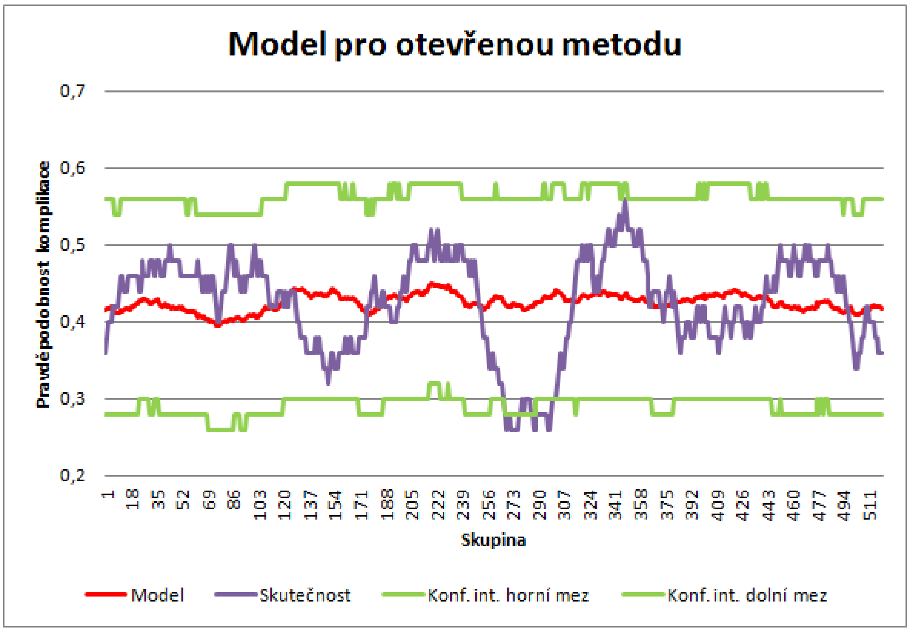
udává modelem odhadnutý počet komplikací v této skupině. Pak je odhad pravděpodobnosti výskytu komplikací ve skupině

$$p = \frac{\hat{n}}{N}$$

Nechť X je náhodná veličina udávající počet komplikací ve skupině. Pak

$$X \to Bi(N, p).$$

Hledáme takový interval, pro který platí



Obrázek 3: Grafické znázornění modelu rizik pacientů operovaných otevřenou metodou

$$P(x_{\frac{\alpha}{2}} \leq X \leq x_{1-\frac{\alpha}{2}}) = 1 - \alpha,$$

tedy takový, že $100(1-\alpha)\%$ hodnot leží v tomto intervalu.

Grafické znázornění námi vytvořených modelů rizik pacientů včetně jejich 95% intervalů spolehlivosti můžete nalézt na obrázcích 3 a 4.

3.5 Využití modelu pro volbu optimální operační techniky

Pomocí našich modelů můžeme pro daného pacienta zvolit operační techniku tak, aby se minimalizovalo riziko pooperačních komplikací.

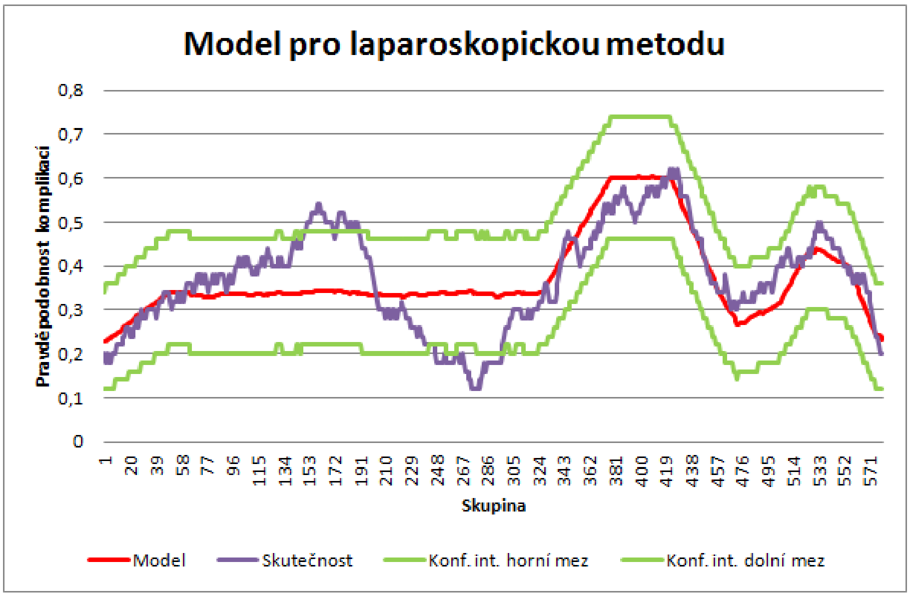
Pravděpodobnost pooperačních komplikací je odhadována vztahem

$$\pi(K, Z) = \frac{e^{\phi(K,Z)}}{1 + e^{\phi(K,Z)}},$$

kde

- $\phi(K, Z) = -1,10138 + 0,143073 \cdot K + 0,104175 \cdot Z$ pro otevřenou metodu (dále značeno jako $\phi_o(K, Z)$),
- $\phi(K, Z) = -2,00799 + 0,107512 \cdot K + 0,274828 \cdot Z$ pro otevřenou metodu (dále značeno jako $\phi_l(K, Z)$, resp. $\pi_l(K, Z)$).

Dále označme $\pi_o(K, Z)$ pravděpodobnost komplikací po otevřené operaci



Obrázek 4: Grafické znázornění modelu rizik pacientů operovaných laparoskopickou metodou

$$\pi_o(K, Z) = \frac{e^{\phi_o(K,Z)}}{1 + e^{\phi_o(K,Z)}}$$

a $\pi_l(K, Z)$ pravděpodobnost komplikací po laparoskopické operaci

$$\pi_l(K, Z) = \frac{e^{\phi_l(K,Z)}}{1 + e^{\phi_l(K,Z)}}.$$

Nyní nás bude zajímat, pro jaké hodnoty K, Z je splněna nerovnost.

$$\pi_l(K, Z) < \pi_o(K, Z)$$

Řešíme tedy nerovnici

$$\frac{e^{\phi_l(K,Z)}}{1 + e^{\phi_l(K,Z)}} < \frac{e^{\phi_o(K,Z)}}{1 + e^{\phi_o(K,Z)}}.$$

Tato nerovnost je ekvivalentní s nerovností

$$\phi_l(K, Z) < \phi_o(K, Z).$$

Po dosazení dostáváme

$$-2,00799 + 0,107512 \cdot K + 0,274828 \cdot Z < -1,10138 + 0,143073 \cdot K + 0,104175 \cdot Z$$

a po úpravě dostáváme

$$-0,90661 - 0,035561 \cdot K + 0,170653 \cdot Z < 0$$

Odtud je zřejmé, že zvyšující se hodnoty proměnné *Závažnost operace* nás nabádají k použití otevřené operace, kdežto zvyšující se hodnoty proměnné *Kardiální příznaky* naopak k zvolení laparoskopické metody. Konkrétní doporučení, která z tohoto vztahu plynou, jsou uvedena v tabulce 8, která obsahuje kombinace vstupních proměnných, pro které podle modelu můžeme doporučit laparoskopickou metodu, a tabulce 9, která naopak obsahuje ty kombinace, pro které by bylo lepší využít metody otevřené.

K	Z	$\phi_l(K,Z)$	$\phi_o(K,Z)$	$\phi_l(K,Z) - \phi_o(K,Z)$	Doporučení
1	1	-1,62565	-0,85413	-0,77152	laparoskopická
1	2	-1,35082	-0,74996	-0,60087	laparoskopická
1	4	-0,80117	-0,54161	-0,25956	laparoskopická
2	1	-1,51814	-0,71106	-0,80708	laparoskopická
2	2	-1,24331	-0,60688	-0,63643	laparoskopická
2	4	-0,69365	-0,39853	-0,29512	laparoskopická
4	1	-1,30311	-0,42491	-0,8782	laparoskopická
4	2	-1,02829	-0,32074	-0,70755	laparoskopická
4	4	-0,47863	-0,11239	-0,36624	laparoskopická
8	1	-0,87307	0,147379	-1,02045	laparoskopická
8	2	-0,59824	0,251554	-0,84979	laparoskopická
8	4	-0,04858	0,459904	-0,50849	laparoskopická

Tabulka 8: Doporučení laparoskopické operační techniky pro dané hodnoty proměnných Kardiální příznaky (K) a Závažnost operace (Z).

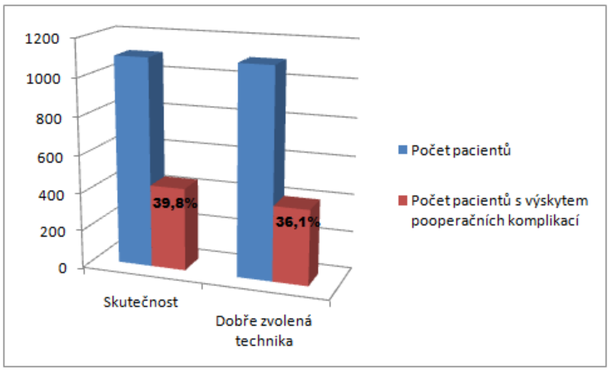
K	Z	$\phi_l(K,Z)$	$\phi_o(K,Z)$	$\phi_l(K,Z) - \phi_o(K,Z)$	Doporučení
1	8	0,298146	-0,12491	0,423053	otevřená
2	8	0,405658	0,018166	0,387492	otevřená
4	8	0,620682	0,304312	0,31637	otevřená
8	8	1,05073	0,876604	0,174126	otevřená

Tabulka 9: Doporučení otevřené operační techniky pro dané hodnoty proměnných Kardiální příznaky (K) a Závažnost operace (Z).

Vzhledem k tomu, že pacienti, kteří jsou obsahem našich dat, byli operovaní náhodně zvolenou operační technikou, můžeme za pomoci doporučení z těchto tabulek srovnat procentuální zastoupení výskytu jejich komplikací oproti tomu, kdyby byli operováni optimální technikou.

Z celkového počtu 1101 pacientů se vyskytly pooperační komplikace u 438 z nich. Komplikace nastaly tedy přibližně u $39,78\%$ pacientů.

Pacientů, kteří byli operováni technikou doporučenou z předchozích tabulek, je celkem 588, pooperační komplikace pak nastaly u 213 z nich. Komplikace nastaly tedy u



Obrázek 5: Porovnání výskytu komplikací ve skutečnosti a v případě, kdyby byli všichni pacienti operováni doporučenou technikou

$36,22\%$ pacientů operovaných takovou technikou, která by jim podle našich modelů také byla doporučena.

Zbylí pacienti pak byli operováni právě opačnou technikou, než kterou náš model doporučuje. Na tyto pacienty tedy aplikujeme doporučený model (tedy ten, kterým operováni nebyli). Takových pacientů bylo 513, a na základě obou našich modelů by pooperační komplikace nastaly u $184,9852$ pacientů, tedy u $36,0595\%$ z nich.

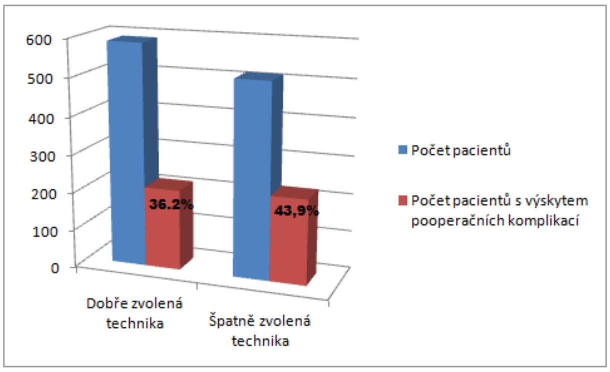
Pokud tedy spojíme pacienty operované správnou technikou s pacienty operovanými špatnou, ale u kterých jsme pravděpodobnost výskytu komplikací určili příslušným modelem techniky pro ně doporučené, získáme vzorek pacientů operovaných doporučenou technikou a vyjde nám, že komplikace by nastaly u 398 z 1101 pacientů, tedy pouze u $36,15\%$ z nich, což je o více než $3,5\%$ pacientů méně než než u těch, kterým byla náhodně vybrána operační technika (viz obrázek 5).

Zajímavější, avšak možná ne až tak důležitý, je rozdíl, kdy porovnáme poměr výskytu pooperačních komplikací ve skutečnosti u pacientů operovaných špatně či dobře zvolenou technikou. Jak již bylo výše zmíněno, pacientů, kteří byli operováni technikou doporučenou, tedy dobře zvolenou, je celkem 588, pooperační komplikace pak nastaly u 213 ($36,22\%$) z nich. Ostatní pacienti byli operováni špatně zvolenou technikou oproti doporučení, tedy 513 pacientů, kdy u 225 ($43,86\%$) z nich nastaly komplikace. Rozdíl je tedy více než $7,5\%$ ve prospěch správně zvolené techniky operace (viz obrázek 6).

3.6 Nedostatky modelu

Při sestavování modelu rizik pacientů operovaných laparoskopickou metodou jsem narazil na zajímavou skupinu pacientů s podobnými parametry, pro které model nemodeluje příliš dobře.

Pro mnou vybrané padesátičlenné skupiny pacientů, u kterých se modelem předpovězená průměrná pravděpodobnost výskytu komplikací držela mezi $0,28$ a $0,34$, se model příliš nedržel skutečnosti, což se projevuje také na grafickém znázornění tohoto modelu na obrázku 2 (detailněji viz také na obrázku 7). Při bližším prozkoumání jsem



Obrázek 6: Porovnání výskytu komplikací u pacientů operovaných dobrou a špatnou operační technikou

zjistil, že velký vliv na tento nedostatek modelu má velký počet pacientů s modelem odhadnutou pravděpodobností výskytu pooperačních komplikací $0,333221$ a $0,309776$, což jsou pacienti s proměnnými *kardiální přiznaky* ohodnocenými jako 2 a 1 a *závažnost operace* ohodnocenou jako 4.

Podrobnější zhodnocení tohoto problému by však vyžadovalo odborné posouzení lékařů.

3.7 Nejdůležitější výsledky modelování rizik touto metodou dosažené ve světě

POSSUM (z angl. *Physiological and Operative Severity for enUmeration of Mortality and morbidity*) je skórovací systém postavený na metodě logistické regrese, který byl vyvinut počátkem 90. let Copelandem a kol., sloužící k modelování rizika výskytu pooperačních komplikací u pacienta.

K odhadnutí pravděpodobnosti výskytu pooperačních komplikací využívá těchto dvou nezávislých proměnných:

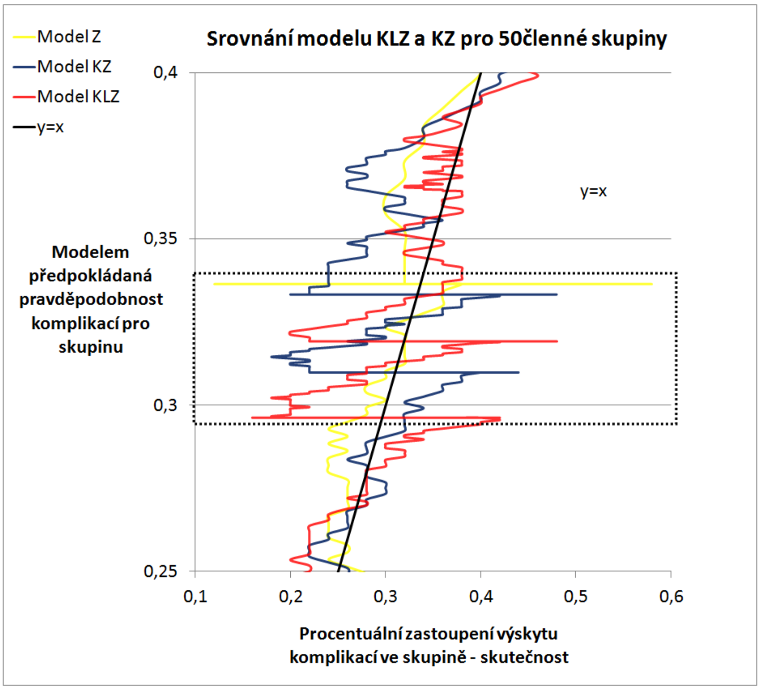
- fyziologické skóre - PS - číslo charakterizující celkový zdravotní stav pacienta, jeho hodnota je dána součtem hodnot proměnných *kardiální příznaky, respirační příznaky, systolický krevní tlak, pulz, Glasgow coma score, hemoglobin, leukocyty, urea v séru, natrium v séru, kalium v séru, EKG,*

- operační skóre - OS - číslo charakterizující operační výkon, jeho hodnota je dána součtem hodnot proměnných *závažnost operace, počet operačních postupů, celková ztráta krve, peritoneální znečištění, rakovina, operační režim,*

kde oborem hodnot těchto proměnných je čtyřprvková množina $\{1, 2, 4, 8\}$.

Pravděpodobnost výskytu pooperačních komplikací v tomto modelu udává vztah

$$\pi(PS, OS) = \frac{e^{\phi(PS,OS)}}{1 + e^{\phi(PS,OS)}},$$



Obrázek 7: Nedostatek modelu pro laparoskopickou techniku operace

kde

$$\phi(PS, OS) = -5,91 + 0,16 \cdot PS + 0,19 \cdot OS.$$

Při aplikaci tohoto modelu na naše data jsme však zjistili, že je nemodeluje dostatečně dobře.

4 Závěr

Cílem práce bylo vytvořit modely logistické regrese, k čemuž byl použit program *Statgraphics Plus 5.0,* pro dvě techniky operace střev, techniky laparoskopické a otevřené, a pomocí nich tyto techniky porovnat. Tyto modely jsem vytvořil na základě dat shromážděných za období let 2001 - 2009 na chirurgické klinice Fakultní nemocnice Ostrava.

Při testování jednotlivých modelů jsem zjistil, že nejvhodnější model, ať už pro metodu laparoskopické či otevřené operace, je model obsahující pouze vstupní proměnné *kardiální příznaky (K)* a *závažnost operace (Z)*. Model pravděpodobnosti výskytu pooperačních komplikací pro pacienty operované otevřenou technikou má tvar

$$\pi(\boldsymbol{x}) = \frac{e^{-1{,}10138+0{,}143073\cdot K+0{,}104175\cdot Z}}{1+e^{-1{,}10138+0{,}143073\cdot K+0{,}104175\cdot Z}},$$

a pro pacienty operované technikou laparoskopickou

$$\pi(\boldsymbol{x}) = \frac{e^{-2{,}00799+0{,}107512\cdot K+0{,}274828\cdot Z}}{1+e^{-2{,}00799+0{,}107512\cdot K+0{,}274828\cdot Z}}.$$

Na základě obou těchto modelů jsem byl schopen určit, která operační technika je pro konkrétního pacienta vhodnější, tedy taková, aby minimalizovala pravděpodobnost výskytu jeho pooperačních komplikací. Pro pacienty s proměnnou *závažnost operace* ohodnocenou číslem 8 vyšla jako vhodnější metoda otevřené operace, pro všechny ostatní je pak na základě modelu vhodnější metoda laparoskopická.

Výsledky této práce by v případě aplikace v praxi měly mít za následek snížení výskytu pooperačních komplikací operace střev u všech pacientů.

5 Reference

[1] Hosmer D. W., Lemeshow S., *Applied Logistic Regression,* Wiley 2000, ISBN 0-471-35632-8.

[2] Litschmannová M., *Úvod do statistiky* [online], Ostrava 2011. Dostupné z: http://mi21.vsb.cz/modul/uvod-do-statistiky [cit. 2010-04-18].

[3] Jahoda P., Martínek, L., Briš, R. *Laparoscopic versus open surgery.* In *Proceedings of SMRLO 2010,* Beer Sheva: Shamoon College of Engineering, 2010, 474-483.

[4] Jahoda P., Martínek, L., Briš, R. *Risk of postoperative complications after surgeries: Laparoscopic versus open surgery.* In *Proceedings of ESREL 2011,* Taylor & Francis Group, 2012, p. 1502-1507.

[5] Copeland G. P., Jones D., Wakters M., *POSSUM: a scoring system for surgical audit.* Br. J. Surg., 1991, 78, p. 356-360.