MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV MATEMATIKY A STATISTIKY

Bakalářská práce

BRNO 2013 **FILIP ZLÁMAL**

MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV MATEMATIKY A STATISTIKY

Logistická regrese v R

Bakalářská práce

Filip Zlámal

Vedoucí práce: doc. RNDr. Zdeněk Pospíšil, Dr. **Brno 2013**

Bibliografický záznam

Autor:	Filip Zlámal Přírodovědecká fakulta, Masarykova univerzita Ústav matematiky a statistiky
Název práce:	Logistická regrese v R
Studijní program:	Matematika
Studijní obor:	Statistika a analýza dat
Vedoucí práce:	doc. RNDr. Zdeněk Pospíšil, Dr.
Akademický rok:	2012/2013
Počet stran:	x + 46
Klíčová slova:	logistický regresní model, metoda maximální věrohodnosti, Newtonova-Raphsonova metoda, ROC křivka

Bibliographic Entry

Author:	Filip Zlámal Faculty of Science, Masaryk University Department of Mathematics and Statistics

Title of Thesis:	Logistic Regression in R
Degree Programme:	Mathematics
Field of Study:	Statistics and Data Analysis
Supervisor:	doc. RNDr. Zdeněk Pospíšil, Dr.
Academic Year:	2012/2013
Number of Pages:	x + 46
Keywords:	logistic regression model, maximum likelihood method, Newton-Raphson method, ROC curve

Abstrakt

Cílem této práce je vytvořit souhrn základních metodik užívaných k vytvoření a popisu logistického regresního modelu s interpretací jeho výstupu společně s praktickou ukázkou postupu při vytváření takového modelu ve statistickém programovém prostředí R. Tato práce je členěna do šesti kapitol. Úvod je věnován vysvětlení pojmu logistické regrese společně s několika příklady praktického užití a motivací pro její studium. V první kapitole je matematicky odvozen vztah pro logistický regresní model. Ve druhé kapitole je popsána metoda maximální věrohodnosti, pomocí které se odhadují regresní koeficienty modelu. Třetí kapitola se věnuje Newtonově-Raphsonově metodě, která slouží k praktickému výpočtu těchto odhadů. Čtvrtá kapitola popisuje testování hypotéz o koeficientech logistického regresního modelu a testování podmodelu. V páté kapitole je uvedeno několik způsobů, jakými lze posoudit kvalitu proložení dat modelem. V závěrečné šesté kapitole je praktická ukázka vytvoření logistického regresního modelu na reálných medicínských datech s použitím statistického softwaru R, verze 2.15.2. V přílohách jsou pak uvedeny vytvořené funkce v R použité k diagnostice výsledního logistického regresního modelu.

Abstract

The aim of this work is to summarize basic methods used for description and interpretation of the logistic regression model with an example of application of the methods for designing such a model using statistical software R. This work is organized into seven chapters. In chapter one, the term of logistic regression is explained with several examples of its application. In chapter two, the logistic regression model is characterized from a mathematical point of view. In chapter three, the maximum likelihood estimation method is described, which is used for logistic regression coefficient estimation. The fourth chapter is devoted to Newton-Raphson method, which is used for calculation of the estimations. The fifth chapter describes hypothesis testing of logistic regression coefficients and testing of the submodel. In chapter six, several ways to evaluate the quality of a model are presented. The final seventh chapter shows how to find a logistic regression model using statistical software R, version 2.15.2. on the example of real medical data. The appendices include R functions developed for testing of the final model.

Masarykova univerzita

Přírodovědecká fakulta

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Student: **Filip Zlámal**

Studijní program - obor: **Matematika - Statistika a analýza dat**

Ředitel Ústavu matematiky a statistiky PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje bakalářskou práci s tématem:

Logistická regrese

Logistic regression

Oficiální zadání: V teoretické části práce vysvětlete princip metody, její předpoklady a interpretaci; uvažujte binární i kategorické proměnné. Popište metody odhadu parametrů a statistické testy významnosti koeficientů a těsnosti proložení dat logistickým modelem. V praktické části použijte vybudovanou teorii na analýzu reálných dat z oblasti medicíny. Výpočty provádějte v prostředí R-language.

Literatura: Doporučená literatura

MELOUN, Milan a Jiří MILITKÝ. Statistická analýza experimentálních dat. Vyd. 2., upr. a rozš. Praha: Academia, 2004. 953 s. ISBN 80-200-1254-0., PEKÁR, Stanislav a Marek BRABEC. Moderní analýza biologických dat. 1. Zobecněné lineární modely v prostředí R. 1. vyd. Praha: Scientia, 2009. 226 s. Biologie dnes. ISBN 978-80-86960-44-9., ZVÁRA, Karel. Regrese. Praha, 2008. 253 s. ISBN 978-80-7378-041-8.

Vedoucí bakalářské práce: doc. RNDr. Zdeněk Pospíšil, Dr.

Datum zadání bakalářské práce: květen 2012

Datum odevzdání bakalářské práce: dle harmonogramu ak. roku 2012/2013

V Brně dne 31.10.2012

prof. RNDr. Jiří Rosický, DrSc.
Ředitel Ústavu matematiky a statistiky

Zadání bakalářské práce převzal dne: 11. 1. 2013

Podpis studenta

Poděkování

Na tomto místě bych chtěl velmi poděkovat doc. RNDr. Zdeňkovi Pospíšilovi, Dr. za cenné připomínky a čas strávený při konzultacích. Můj dík též patří prof. MUDr. Anně Vašků, Csc. a MUDr. Janu Máchalovi za přínosné informace k praktické části této práce.

Prohlášení

Prohlašuji, že jsem svoji bakalářskou práci vypracoval samostatně s využitím informačních zdrojů, které jsou v práci citovány.

Brno 19. května 2013 .....

Filip Zlámal

Obsah

Úvod

Logistická regrese (též logistický regresní model nebo model logistické regrese) je název pro regresní model s binární (dichotomickou) závisle proměnnou. Byla navržena v 60. letech 20. století jako alternativa k metodě nejmenších čtverců. Dříve se týkala většina úloh aplikace logistické regrese zejména oblasti medicíny a epidemiologie. Vysvětlovaná (závislá) proměnná představuje např. přítomnost nebo nepřítomnost choroby. Logistická regrese pak umožňuje modelovat např. riziko vzniku srdeční choroby jako funkci řady antropometrických a biochemických parametrů (pohlaví, věk, BMI, krevní tlak, hladina cholesterolu, kouření apod.). V průmyslu zase můžeme sledovat úspěšnost nebo neúspěšnost nějakého výrobku a logistickou regresí lze určit, které veličiny se na úspěšnosti významně podílejí. V bankovnictví se používá k vytvoření modelů, které dokáží odhadnout na základě řady parametrů o klientovi banky (např. věk, pohlaví, nejvyšší dosažené vzdělání) žádajícím o úvěr, jestli bude tento úvěr splácet řádně či nikoliv. Metoda logistické regrese je též alternativou k diskriminanční analýze a analýze směsi normálních rozložení [3]. Výsledný model tak lze užít ke klasifikování objektů.

Logistická regrese se od lineární regrese liší v tom, že predikuje pravděpodobnost toho, zda-li se nějaká událost stane nebo nestane. K vytvoření vazební podmínky mezi touto pravděpodobností a lineárním prediktorem tvořeným nezávisle proměnnými $X_1, \ldots, X_m$ se používá *logitová transformace.* Rozdíl mezi logistickou a lineární regresí spočívá v tom, že logistická regrese používá kategorickou vysvětlovanou (závislou) proměnou, kdežto u lineární regrese je vysvětlovaná proměnná spojitá.

Podle typu závislé proměnné rozlišujeme

a) *binární logistickou regresi* - týká se binární závislé proměnné, která nabývá pouze dvou možných hodnot, např. absence a přítomnost jevu, muž a žena.

b) *ordinální logistickou regresi* - závislou proměnnou je veličina ordinální typu, nabývající více možných stavů, mezi nimiž existuje přirozené uspořádání, např. stadium závažnosti nějakého onemocnění, typ odpovědi v dotazníku s možnými odpověďmi vůbec ne, spíše ne, spíše ano, určitě ano.

c) *(multi)nominální logistickou regresi* - týká se nominální závislé proměnné o více než dvou úrovních stavů, mezi nimiž existuje pouze odlišnost, např. barva očí, rasa.

Obdobně jako u lineární regrese, vektor vysvětlujících proměnných u všech třech druhů logistické regrese může obsahovat více proměnných, a to jak spojitých, zvaných *prediktory*, tak kategoriálních, zvaných *faktory*.

V této práci se dále pod pojmem logistická regrese rozumí binární logistická regrese.

Kapitola 1

Odvození vztahu pro model logistické regrese

U lineární regrese se modeluje střední hodnota spojité normálně rozdělené náhodné veličiny Y pomocí souboru nezávislých náhodných veličin $X_1, \ldots, X_m$. Při výpočtech se používá metod známých z lineární algebry.

V případě, kdy závislou proměnou Y je dichotomická náhodná veličina (tj. veličina, která může nabývat pouze dvou hodnot), není již možné vzhledem k charakteru této proměnné použít lineární regresi k predikci její střední hodnoty.

V této kapitole je proto odvozen vztah umožňující modelovat pravděpodobnost pro Y pomocí nezávislých náhodných veličin $X_1, \ldots, X_n$.

Začněme z definicí dichotomické náhodné veličiny. Nechť náhodná veličina $Y \sim A(\vartheta)$, $0 < \vartheta < 1$. To znamená, že

$$P(Y = y) = \begin{cases} \vartheta & , y = 1 \\ 1 - \vartheta & , y = 0 \\ 0 & , \text{jinak} \end{cases}$$

Tuto pravděpodobnost lze též přepsat do kompaktnější podoby

$$P(Y = y) = \vartheta^y (1 - \vartheta)^{1-y} \qquad \text{pro } y = 0, 1. \tag{1.1}$$

Přímým výpočtem lze spočítat střední hodnotu a rozptyl veličiny Y:

$$\mathrm{E}Y = \vartheta \qquad \mathrm{D}Y = \vartheta(1 - \vartheta).$$

Je zřejmé, že použití modelu ve tvaru $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$ není korektní, jelikož na levé straně vystupuje diskrétní veličina a na straně pravé mohou být veličiny spojité (veličiny $X_1, \ldots, X_m$ mohou být kategoriální i spojité a pro koeficienty v modelu platí $-\infty < \beta_i < \infty$). Proto nebudeme modelovat přímo hodnotu veličiny Y, ale *pravděpodobnost* toho, že Y nabude určité hodnoty. Jelikož Y je dichotomická, platí, že $P(Y = 1) = 1 - P(Y = 0)$. To znamená, že si můžeme vybrat, zda modelovat $P(Y = 1)$ nebo $P(Y = 0)$. Zvolíme např. $P(Y = 1)$.

Tato pravděpodobnost nabývá hodnot v intervalu $(0, 1)$. Kdybychom nyní vytvořili model $P(Y = 1) = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$, mohlo by se stát, že získáme takového odhady koeficientů $\beta_0, \beta_1, \ldots, \beta_m$, že pro určité realizace $x_1, \ldots, x_m$ veličin $X_1, \ldots, X_m$ dostaneme predikované hodnoty pravděpodobnosti ležící mimo interval $(0, 1)$, jelikož lineární prediktor může teoreticky nabývat všech hodnot z $\mathbb{R}$.

1.1 Šance

Tento problém lze částečně vyřešit zavedením pojmu *šance* (angl. *odds*) jako podílu

$$\text{odds}(P(Y=1)) = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)}. \tag{1.2}$$

Šance vyjadřuje, kolikrát je vyšší pravděpodobnost toho, že Y nabude hodnoty 1, než pravděpodobnost, že nabude hodnoty 0. Hodnoty šance leží v intervalu $(0,\infty)$.

1.2 Logitová transformace

Nyní je ještě třeba vzájemně jednoznačně transformovat interval $(0,\infty)$ na $(-\infty,\infty)$. K tomuto účelu se hodí použití funkce přirozeného logaritmu, čímž zavedeme *logitovou funkci*

$$\text{logit}(P(Y=1)) = \ln(\text{odds}(P(Y=1))) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right). \tag{1.3}$$

Takto transformovanou pravděpodobnost již můžeme modelovat obdobně jako je tomu u lineární regrese

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m, \tag{1.4}$$

odkud vyjádříme pravděpodobnost

$$P(Y=1) = \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 X_1+\cdots+\beta_m X_m)}}. \tag{1.5}$$

Při označení $\beta = (\beta_0,\beta_1,\ldots,\beta_m)'$ a $\mathbf{X} = (1,X_1,\ldots,X_m)'$ tento vztah můžeme přepsat jako

$$P(Y=1) = \frac{1}{1+\mathrm{e}^{-\mathbf{X}'\beta}}. \tag{1.6}$$

1.3 Model logistické regrese

Jelikož pro různé realizace $\mathbf{x}$ náhodného vektoru $\mathbf{X}$ nabývá pravděpodobnost (1.6) různých hodnot, je proto tato pravděpodobnost podmíněná, tudíž

$$\boxed{P(Y=1|\mathbf{X}=\mathbf{x}) = \frac{1}{1+\mathrm{e}^{-\mathbf{x}'\beta}}}. \tag{1.7}$$

Můžeme si všimnout, že na rozdělení náhodného vektoru $\mathbf{X}$ nebyly během odvozování kladeny žádné podmínky. Tím se stává logistická regrese poměrně univerzálním nástrojem.

Pro úplnost lze odvodit vztah pro model doplňkové pravděpodobnosti, tj.

$$P(Y=0|\mathbf{X}=\mathbf{x}) = 1-P(Y=1|\mathbf{X}=\mathbf{x}) = \frac{\mathrm{e}^{-\mathbf{x}'\beta}}{1+\mathrm{e}^{-\mathbf{x}'\beta}} = \frac{1}{1+\mathrm{e}^{\mathbf{x}'\beta}}. \tag{1.8}$$

Můžeme si všimnout, že $P(Y=1)=\vartheta=\mathrm{E}Y$, tzn. že modelem (1.7) predikujeme střední hodnotu náhodné veličiny Y v závislosti na realizacích $\mathbf{x}$.

K praktickému vytvoření logistického regresního modelu je potřeba mít dostatečný počet měření. Mějme soubor o rozsahu n. Nechť $Y_1,\ldots,Y_n \sim A(\vartheta), 0<\vartheta<1$. Označme y_i realizaci náhodné veličiny Y_i. Nechť dále ke každé Y_i přísluší soubor m náhodných veličin $X_{i1},\ldots,X_{im}$ s realizacemi $x_{i1},\ldots,x_{im}$. Označme $\mathbf{x}=(1,x_{i1},\ldots,x_{im})'$. Potom model logistické regrese pro i-tou proměnnou je ve tvaru

$$P(Y_i=1|\mathbf{X}_i=\mathbf{x}_i)=\frac{1}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}. \tag{1.9}$$

Nyní je cílem získat odhady koeficientů modelu $\beta=(\beta_0,\beta_1,\ldots,\beta_m)'$. K tomuto účelu se používá *metoda maximální věrohodnosti*. O jejím popisu a vlastnostech takto získaných odhadů pojednává další kapitola.

1.4 Poměr šancí

Podívejme se ještě na význam regresních koeficientů. Z (1.7) vyplývá, že pokud je nějaké $\beta_i>0$, tak pokud x_i poroste, poroste tím i pravděpodobnost $P(Y=1)$. Čím větší je β_i, tím tento nárůst bude rychlejší. Zavedeme pojem *poměr šancí* (angl. odds ratio) vztahem

$$\mathrm{OR}(X_i)=\frac{\mathrm{odds}(P(Y=1|X_1=x_1,\ldots,X_{i-1}=x_{i-1},X_i=x_i+1,X_{i+1}=x_{i+1},\ldots,X_m=x_m))}{\mathrm{odds}(P(Y=1|X_1=x_1,\ldots,X_{i-1}=x_{i-1},X_i=x_i,X_{i+1}=x_{i+1},\ldots,X_m=x_m))}. \tag{1.10}$$

Po dosazení z (1.7) dostaneme

$$\mathrm{OR}(X_i)=\mathrm{e}^{\beta_i}. \tag{1.11}$$

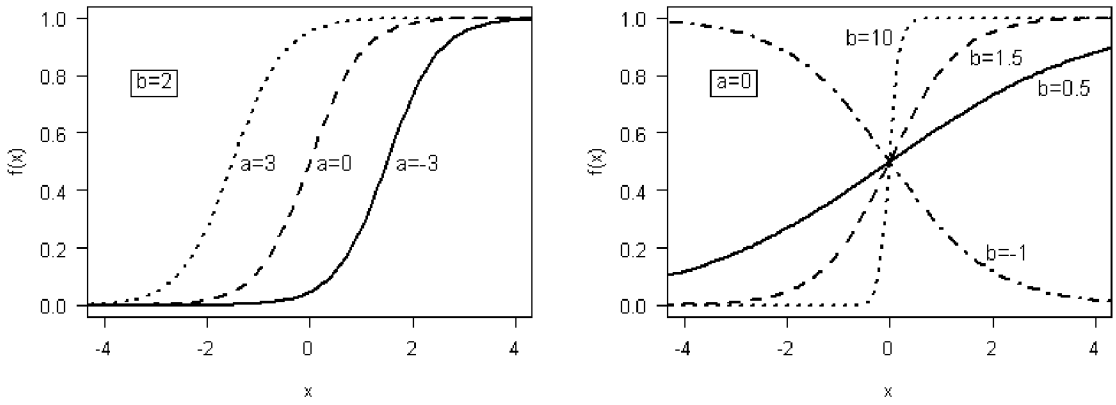
Poměr šancí $\mathrm{OR}(X_i)$ udává, kolikrát se zvětší šance na to, aby se Y realizovala hodnotou 1, když se hodnota nezávislé proměnné zvýši o 1, jestliže zbývajících $n-1$ veličiny je fixovaných. Tento poměr je jednoznačně určen koeficientem β_i. Je–li $\beta_i>0$, je $\mathrm{OR}(X_i)>1$. Je–li $\beta_i<0$, je $\mathrm{OR}(X_i)<1$.

1.5 Logistická funkce

Z popisu logistické funkce vyplyne kvalitativní význam koeficientů v modelu logistické regrese. Definujeme ji jako funkci

$$f(x;a,b)=\frac{1}{1+\mathrm{e}^{-(a+bx)}}, \tag{1.12}$$

kde parametry $a,b\in\mathbb{R}$. Jejím definičním oborem je $\mathbb{R}$, obor hodnot tvoří interval $(0,1)$, je lichou funkcí vzhledem k bodu $\left[-\frac{a}{b},\frac{1}{2}\right]$ (pro $b\neq 0$), pro $b=0$ je tato funkce konstatní, je rostoucí na celém $\mathbb{R}$, $\lim_{x\to\infty}f(x)=1$, $\lim_{x\to-\infty}f(x)=0$, je třídy $C^\infty(\mathbb{R})$. Umožnuje popsat logistický regresní model $P(Y=1|X=x)=\frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 x)}}$ s jedním prediktorem. Následující obrázky ukazují několik grafů logistické funkce pro různé hodnoty koeficientů a a b.



Obrázek 1.1: Grafy logistických funkcí pro různé hodnoty parametrů a a b.

Parametr a (u logistického regresního modelu se jedná o absolutní člen β_0) určuje posunutí logistické křivky podél osy x, kdežto parametr b (u logistického regresního modelu to bude parametr β_1) pak „strmost“ křivky v okolí bodu $\left[-\frac{a}{b};\frac{1}{2}\right]$.

Kapitola 2

Metoda maximální věrohodnosti

K bodovému odhadu koeficientů u lineárního regresního modelu $(\mathbf{Y}, \mathbf{X}\beta, \sigma^2\mathbf{I})$ se používá *metoda nejmenších čtverců*, která spočívá v nalezení takových hodnot koeficientů modelu, které minimalizují tzv. součet čtverců $(\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta)$. Jediným požadavkem pro to, aby tento odhad existoval, je, aby matice $\mathbf{X}$ měla lineárně nezávislé sloupce.

U modelu logistické regrese tuto metodu ovšem nelze korektně použít, a to kvůli charakteru závislé proměnné. Z toho důvodu se používá k odhadům koeficientů modelu *metoda maximální věrohodnosti* (angl. *maximum likelihood estimation*, zkr. MLE), která má širší použití (např. k bodových odhadům parametrů rozdělení, u dalších typů zobecněných lineárních modelů [4], u beta regrese [5]).

V této kapitole uvedeme její popis, předpoklady pro použití, jejich vlastnosti a odvodíme tzv. systém věrohodnostních rovnic pro případ logistické regrese.

2.1 Popis metody maximální věrohodnosti

Mějme $\mathbf{X} = (X_1, \ldots, X_n)'$ náhodný vektor, jehož složky tvoří náhodný výběr pocházející z rozdělení s hustotou $f(\mathbf{x}|\theta)$, kde $\mathbf{x} \in \mathbb{R}^n$, $\theta = (\theta_1, \ldots, \theta_m)' \in \Omega$ je vektorem parametrů charakterizující toto rozdělení. O $f(\mathbf{x}|\theta)$ předpokládáme, že pochází z nějakého systému hustot $\{f(\mathbf{x}|\theta); \theta \in \Omega\}$, který je vektorem parametrů $\theta \in \Omega \subseteq \mathbb{R}^m$ jednoznačně určen.

Jelikož $X_1, \ldots, X_n$ tvoří náhodný výběr, je sdružená hustota pravděpodobnosti vektoru $\mathbf{X} = (X_1, \ldots, X_n)'$ rovna

$$f(\mathbf{x}|\theta) = f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{2.1}$$

Jedná se o funkci proměnných $x_1, \ldots, x_n$, přičemž θ zde vystupuje jako vektor parametrů, které jsou fixované.

Definujeme *věrohodnostní funkci* formálně stejným vztahem jako (2.1), tj.

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta), \tag{2.2}$$

ovšem s tím rozdílem, že proměnnou je nyní θ a $\mathbf{x}$ stojí v roli parametru.

Cílem metody je nalézt pro dané realizace $\mathbf{x}$ vektoru $\mathbf{X}$ (tj. prakticky pro n nezávislých naměřených hodnot téže proměnné X) takový odhad vektoru parametrů θ, který bude *maximalizovat* věrohodnostní funkci (2.2).

Definice 2.1. *Odhad $\hat{\theta}_{ML} \in \Omega$ nazveme* maximálně věrohodným odhadem (zkráceně ML odhad) *právě tehdy, když pro libovolné $\mathbf{x} \in \mathbb{R}^n$ a pro všechna $\theta \in \Omega$ platí $L(\hat{\theta}_{ML}|\mathbf{x}) \geq L(\theta|\mathbf{x})$.*

Dále budeme místo $L(\theta|\mathbf{x})$ psát pro jednoduchost $L(\theta)$.

Často je ovšem výhodnější místo s věrohodností funkcí pracovat s jejím přirozeným logaritmem, proto se zavádí *logaritmická věrohodnostní funkce*

$$l(\theta) = \ln L(\theta), \tag{2.3}$$

kterou lze zapsat s využitím (2.2) jako

$$l(\theta) = \ln\left(\prod_{i=1}^{n} f(x_i|\theta)\right) = \sum_{i=1}^{n} \ln f(x_i|\theta). \tag{2.4}$$

Jelikož je logaritmická funkce monotónní, nemění polohu ML odhadu, tj. má–li funkce $L(\theta)$ maximum v bodě $\hat{\theta}_{ML}$, má v tomtéž bodě maximum i funkce $\ln L(\theta)$, a obráceně.

K nalezení maxima funkce $L(\theta)$ použijeme metod známých z matematické analýzy sloužících k hledání extrémů funkcí více proměnných. Předpokládejme, že $L(\theta)$ má parciální derivace alespoň druhého řádu na Ω. Nejprve vyřešíme *systém věrohodnostních rovnic*

$$\frac{\partial L(\theta)}{\partial \theta_j} = 0, \qquad j = 1, \ldots, m \tag{2.5}$$

s řešením $\theta = \hat{\theta}$. Dále je třeba ověřit, že $L(\theta)$ nabývá v bodě $\hat{\theta}$ svého maxima, musí tedy platit

$$H\left(\hat{\theta}\right) = \left(\frac{\partial^2 L(\theta)}{\partial \theta_i \partial \theta_j}\right)_{i,j=1}^{m}\Bigg|_{\theta=\hat{\theta}} < 0, \tag{2.6}$$

tj. že Hessova matice $H\left(\hat{\theta}\right)$ je negativně definitní. Je–li množina Ω ohraničená, je třeba navíc ještě vyšetřit hodnoty $L(\theta)$ na hranici Ω.

2.2 Vlastnosti ML odhadu

Zde shrneme základní vlastnosti ML odhadu.

Definice 2.2. *Systém hustot $\{f(\mathbf{x}|\theta); \theta \in \Omega\}$ nazveme* regulární *právě tehdy, když platí:*

1. *Ω je neprázdná otevřená množina.*
2. *$M = \{\mathbf{x} \in \mathbb{R}^n; f(\mathbf{x}|\theta) > 0\}$ nezávisí na θ.*
3. *$\forall \theta \in \Omega$ a pro skoro všechna $\mathbf{x} \in M$ existuje konečná parciální derivace $f_i'(\mathbf{x}|\theta) = \frac{\partial f(\mathbf{x}|\theta)}{\partial \theta_i}$, $i = 1, \ldots, m$.*

4. $\forall\theta \in \Omega$ *platí* $\mathrm{E}\left[\frac{\partial}{\theta_i}\ln f(\mathbf{X}|\theta)\right] = \int_M \frac{f_i'(\mathbf{x}|\theta)}{f(\mathbf{x}|\theta)}\,\mathrm{d}F(\mathbf{x}|\theta) = 0$, $i = 1,\ldots,m$.

5. $\forall i,j = 1,\ldots,m$ *platí, že existuje konečný integrál*

$$J_{ij}(\theta) = \int_M \frac{f_i'(\mathbf{x}|\theta)f_j'(\mathbf{x}|\theta)}{f^2(\mathbf{x}|\theta)}\,\mathrm{d}F(\mathbf{x}|\theta).$$

6. *Matice* $\mathbf{J}_n(\theta) = (J_{ij}(\theta))_{i,j=1}^m$ *je* $\forall\theta \in \Omega$ *pozitivně definitní.*

Matici $\mathbf{J}_n(\theta)$ nazýváme *Fisherovou informační maticí*. Tu lze též vyjádřit pomocí ekvivalentních definic (viz [1], [8]):

$$J_{ij}(\theta) = \mathrm{cov}\left(\frac{\partial\ln f(\mathbf{X}|\theta)}{\partial\theta_i}, \frac{\partial\ln f(\mathbf{X}|\theta)}{\partial\theta_j}\right) = \mathrm{E}\left(\frac{\partial\ln f(\mathbf{X}|\theta)}{\partial\theta_i}\cdot\frac{\partial\ln f(\mathbf{X}|\theta)}{\partial\theta_j}\right) \tag{2.7}$$

nebo

$$J_{ij}(\theta) = -\mathrm{E}\left(\frac{\partial^2\ln f(\mathbf{X}|\theta)}{\partial\theta_i\partial\theta_j}\right). \tag{2.8}$$

Její význam mj. vyplývá z následující věty.

Věta 2.1. *Nechť systém hustot* $\{f(x|\theta);\theta\in\Omega\}$ *je regulární a má Fisherovou informační maticí* $\mathbf{J}_n(\theta)$. *Nechť jsou dále splněny následující předpoklady:*

P1: Ω *je parametrický prostor, který obsahuje takový neprázdný otevřený interval* ω, *že skutečná hodnota parametru* $\theta_0 \in \omega$.

P2: $\mathbf{X} = (X_1,\ldots,X_n)'$, *kde* X_i *jsou nezávislé stejně rozdělené náhodné veličiny s hustotou* $f(x|\theta)$.

P3: Nechť $M = \{x; f(x|\theta) > 0\}$ *nezávisí na* θ.

P4: $\forall\theta_1,\theta_2 \in \Omega$ *platí* $f(x|\theta_1) = f(x|\theta_2)$ *právě tehdy, když* $\theta_1 = \theta_2$.

P5: Pro skoro všechna $x \in \mathbb{R}$, *pro všechna* $\theta \in \omega$ *a pro všechna* $r,s,t = 1,\ldots,m$ *existuje derivace* $\frac{\partial^3 f(x|\theta)}{\partial\theta_r\partial\theta_s\partial\theta_t}$.

P6: Pro všechna $\theta \in \omega$ *platí*

$$\int_M \frac{\partial^2 f(x|\theta)}{\partial\theta_r\partial\theta_s}\,\mathrm{d}F(x|\theta) = 0, \quad r,s = 1,\ldots,m.$$

P7: Pro všechna $r,s,t = 1,\ldots,m$ *existují funkce* $M_{rst}(x) \geq 0$ *takové, že*

$$m_{rst} = \mathbf{E}_{\theta_0} M_{rst}(X) < \infty$$

a

$$\left|\frac{\partial^3\ln f(x|\theta)}{\partial\theta_r\partial\theta_s\partial\theta_t}\right| \leq M_{rst}(x).$$

Potom platí:

1. *(konzistence) Jestliže $n \to \infty$, pak ke každému $\varepsilon > 0$ existuje s pravděpodobností blížící se jedné takové řešení $\hat{\theta}_n$ systému věrohodnostních rovnic, že $||\hat{\theta}_n - \theta_0|| < \varepsilon$.*

2. *(asymptotická normalita) Položme*

$$\mathbf{U}(\theta) = \begin{pmatrix} \frac{\partial L(\theta)}{\theta_1} \\ \vdots \\ \frac{\partial L(\theta)}{\theta_m} \end{pmatrix}.$$

   Pak pro $n \to \infty$ platí

$$\frac{1}{\sqrt{n}}\mathbf{U}(\theta_0) \xrightarrow{\mathrm{d}} \mathbf{N}(\mathbf{0}, \mathbf{J}(\theta_0)).$$

3. *(eficience) Existuje–li pro každé dostatečně velké n a pro každou hodnotu $\mathbf{X}$ takový kořen $\hat{\theta}_n$ systému věrohodnostních rovnic, že $\hat{\theta}_n$ je konzistentním odhadem parametru θ_0, pak*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathrm{d}} \mathbf{N}\left(\mathbf{0}, [\mathbf{J}(\theta_0)]^{-1}\right).$$

Přednosti maximálně věrohodného odhadu vycházejí z jeho asymptotických vlastností. Je-li tedy rozsah souboru malý, nelze použití této metody obecně považovat za „bezpečné“.

2.3 ML odhad koeficientů u logistického regresního modelu

V kapitole 1 byl odvozen vztah (1.7) pro model pravděpodobnosti dichotomické náhodné veličiny pomocí nezávislých veličin $X_1, \ldots, X_m$. V této podkapitole odvodíme systém věrohodnostních rovnic vedoucí k získání odhadu koeficientů modelu.

2.3.1 Odvození věrohodnostních rovnic

Mějme náhodný výběr $Y_1, \ldots, Y_n \sim A(\vartheta)$, $0 < \vartheta < 1$, s realizacemi $y_1, \ldots, y_n$. To znamená, že podle (1.1) je pravděpodobnost

$$P(Y_i = y_i) = \vartheta^{y_i}(1-\vartheta)^{1-y_i}. \tag{2.9}$$

Pro střední hodnotu a rozptyl veličiny tak Y_i platí

$$\mathrm{EY_i} = \vartheta \qquad \mathrm{DY_i} = \vartheta(1-\vartheta).$$

Každé realizaci y_i přísluší realizace $x_{i1}, \ldots, x_{im}$ veličin $X_{i1}, \ldots, X_{im}$. Označme $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{im})'$, $\beta = (\beta_0, \beta_1, \ldots, \beta_m)'$ a $\mathbf{x}_i'\beta = \sum_{j=0}^{m} x_{ij}\beta_j$ (přičemž $x_{i0} = 1$). Podle (1.9) modelujeme pravděpodobnost (2.9) jako

$$P(Y_i = y_i | \mathbf{X}_i = \mathbf{x}_i) = \left(\frac{1}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}\right)^{y_i} \left(\frac{\mathrm{e}^{-\mathbf{x}_i'\beta}}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}\right)^{1-y_i} = \frac{\left(\mathrm{e}^{-\mathbf{x}_i'\beta}\right)^{1-y_i}}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}} \tag{2.10}$$

Věrohodnostní funkce je potom ve tvaru

$$L(\beta) = \prod_{i=1}^{n} P(Y_i = y_i) = \prod_{i=1}^{n} \frac{\left(\mathrm{e}^{-\mathbf{x}_i'\beta}\right)^{1-y_i}}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}} \tag{2.11}$$

Je výhodné dále pracovat s logaritmickou věrohodnostní funkcí:

$$\begin{aligned} l(\beta) = \ln L(\beta) &= \ln\left(\prod_{i=1}^{n} \frac{\left(\mathrm{e}^{-\mathbf{x}_i'\beta}\right)^{1-y_i}}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}\right) = \sum_{i=1}^{n} \ln\left(\frac{\left(\mathrm{e}^{-\mathbf{x}_i'\beta}\right)^{1-y_i}}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}\right) = \\ &= \sum_{i=1}^{n} \left[(y_i-1)\mathbf{x}_i'\beta - \ln\left(1+\mathrm{e}^{-\mathbf{x}_i'\beta}\right)\right]. \end{aligned} \tag{2.12}$$

Jako poznámku lze uvést, že vhodnou úpravou vztahu pro $l(\beta)$ dostaneme

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \left[y_i \ln\left(\frac{1}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}\right) + (1-y_i)\ln\left(\frac{1}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}\right)\right] = \\ &= \sum_{i=1}^{n} \left[y_i \ln P(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i) + (1-y_i)\ln(1-P(Y_i = y_i|\mathbf{X}_i = \mathbf{x}_i))\right]. \end{aligned} \tag{2.13}$$

Věrohodnostní rovnice pro $l(\beta)$ jsou ve tvaru

$$\begin{aligned} 0 = \frac{\partial l(\beta)}{\partial \beta_r} &= \sum_{i=1}^{n} \left[(y_i-1)x_{ir} - \frac{\mathrm{e}^{-\mathbf{x}_i'\beta}}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}(-x_{ir})\right] \\ 0 &= \sum_{i=1}^{n} \left[y_i - 1 + \frac{\mathrm{e}^{-\mathbf{x}_i'\beta}}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}\right] x_{ir}, \qquad r = 0,1,\ldots,m. \end{aligned} \tag{2.14}$$

Označme dále $n_0 = \mathrm{card}\{y_i; y_i = 0\}$ a $n_1 = \mathrm{card}\{y_i; y_i = 1\}$, tj. n_0 je počet realizací y_i rovných 0 a n_1 počet realizací y_i rovných 1. Zřejmě je $n = n_0 + n_1$.

Soustavu $m+1$ rovnic (2.14) můžeme ještě upravit:

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \left[y_i - 1 + \frac{\mathrm{e}^{-\mathbf{x}_i'\beta}}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}\right] x_{ir} \\ 0 &= \sum_{i=1}^{n} \left[y_i - \left(1 - \frac{\mathrm{e}^{-\mathbf{x}_i'\beta}}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}\right)\right] x_{ir} \\ 0 &= \sum_{i=1}^{n} \left[y_i - \frac{1}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}\right] x_{ir} \\ 0 &= \sum_{i=1}^{n} y_i x_{ir} - \sum_{i=1}^{n} \frac{1}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}} x_{ir} \end{aligned} \tag{2.15}$$

$$\sum_{\{i;y_i=1\}} x_{ir} = \sum_{i=1}^{n} \frac{1}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}} x_{ir} \qquad r = 0,1,\ldots,m \tag{2.16}$$

Pro $r = 0$ dostaneme první z věrohodnostních rovnic:

$$\sum_{\{i;y_i=1\}} 1 = \sum_{i=1}^{n} \frac{1}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}$$

$$\boxed{n_1 = \sum_{i=1}^{n} \frac{1}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}} \tag{2.17}$$

a zbývajících m věrohodnostních rovnic je ve stejném tvaru jako (2.16):

$$\boxed{\sum_{\{i;y_i=1\}} x_{ir} = \sum_{i=1}^{n} \frac{1}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}} x_{ir}} \qquad r = 1,\ldots,m. \tag{2.18}$$

Jedná se o nelineární soustavu $m+1$ rovnic o $m+1$ neznámých $\beta_0, \beta_1, \ldots, \beta_m$. S použitím (2.15) lze tuto soustavu zapsat v maticovém tvaru

$$\boxed{\mathbf{X}'(\mathbf{Y} - \vartheta(\beta)) = \mathbf{0}}, \tag{2.19}$$

kde $(\mathbf{X})_{ij} = x_{ij}$, $(\mathbf{Y})_i = y_i$, $(\vartheta(\beta))_i = \frac{1}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}$.

Označme řešení těchto rovnic jako $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m)'$. Toto řešení ovšem nelze obecně nalézt v algebraickém tvaru, proto se hledá numericky, a to pomocí *Newtonovy-Raphsonovy metody*, o jejíž aplikaci u logistické regrese pojednává další kapitola.

2.3.2 Fisherova informační matice

Z věty 2.2 vyplývá význam Fisherovy informační matice: její inverze je asymptotickou kovarianční maticí konzistentního odhadu vektoru parametrů β. Označíme-li $\mathbf{T}(\mathbf{X})$ nestranný odhad parametru θ_0, potom z tvrzení 3 ve větě 2.2 s použitím Raovy-Cramérovy věty (viz [1]) plyne, že matice $\mathrm{Var}[\mathbf{T}(\mathbf{X})] - [\mathbf{J}(\theta_0)]^{-1}$ je asymptoticky pozitivně semidefinitní.

Pro praktické získání odhadu $\mathbf{J}(\beta)$ u logistické regrese postupujeme tak, že nejprve nalezneme mínus matici druhých parciálních derivací logaritmické věrohodnostní funkce $-\left(\frac{\partial^2 l(\beta)}{\partial\beta_r\partial\beta_s}\right)_{r,s=0}^{m}$, kterou dále vyčíslíme pro $\beta = \hat{\beta}$. Stačí tedy parciálně zderivovat výraz na pravé straně rovnice (2.15) podle β_s:

$$\frac{\partial^2 l(\beta)}{\partial\beta_r\partial\beta_s} = \frac{\partial}{\beta_s}\left(\sum_{i=1}^{n}\left[y_i - \frac{1}{1+\mathrm{e}^{-\mathbf{x}_i'\beta}}\right]x_{ir}\right) = -\sum_{i=1}^{n}\frac{\mathrm{e}^{-\mathbf{x}_i'\beta}}{\left(1+\mathrm{e}^{-\mathbf{x}_i'\beta}\right)^2}x_{ir}x_{is}, \tag{2.20}$$

tudíž

$$\left(\mathbf{J}(\tilde{\beta})\right)_{rs} = -\left.\frac{\partial^2 l(\beta)}{\partial\beta_r\partial\beta_s}\right|_{\beta=\tilde{\beta}} = \sum_{i=1}^{n}\frac{\mathrm{e}^{-\mathbf{x}_i'\tilde{\beta}}}{\left(1+\mathrm{e}^{-\mathbf{x}_i'\tilde{\beta}}\right)^2}x_{ir}x_{is}. \tag{2.21}$$

Všimněme si, že $\frac{\mathrm{e}^{-\mathbf{x}_i'\tilde{\beta}}}{\left(1+\mathrm{e}^{-\mathbf{x}_i'\tilde{\beta}}\right)^2} = \vartheta_i(\tilde{\beta})(1-\vartheta_i(\tilde{\beta}))$. Pokud označíme $\mathbf{V}(\tilde{\beta}) = \mathrm{diag}(\vartheta_1(\tilde{\beta})(1-\vartheta_1(\tilde{\beta})),\ldots,\vartheta_n(\tilde{\beta})(1-\vartheta_n(\tilde{\beta}))$, můžeme odhad Fisherovy informační matice zapsat ve tvaru

$$\boxed{\mathbf{J}(\tilde{\beta}) = \mathbf{X}'\mathbf{V}(\tilde{\beta})\mathbf{X}}. \tag{2.22}$$

Tato matice je symetrická, diagonální prvky $\mathbf{V}(\tilde{\beta})$ jsou kladné, tudíž $\mathbf{J}(\tilde{\beta})$ je přinejmenším pozitivně semidefinitní. Má-li $\mathbf{X}$ plnou sloupcovou hodnost, je $\mathbf{J}(\tilde{\beta})$ dokonce pozitivně definitní.

Lze ukázat, že hustota pro logistický regresní model je regulární a že splňuje předpoklady věty 2.2 (viz např. [9]). To znamená, že tvrzení této věty platí pro maximálně věrohodné odhady koeficientů logistického regresního modelu.

2.3.3 Interpretace věrohodnostních rovnic

Můžeme si všimnout, že v rovnicích (2.17) a (2.18) vystupuje na pravé straně člen $\frac{1}{1+e^{-\mathbf{x}_i'\beta}}$, který je dle předpokladu modelu roven pravděpodobnosti, že náhodná veličina Y_i nabude hodnoty 1, tj. $\frac{1}{1+e^{-\mathbf{x}_i'\beta}} = P(Y_i = 1)$. Dosazením do uvedených rovnic dostaneme, že řešení soustavy věrohodnostních rovnic $\hat{\beta}$ musí být takové, aby

1) $n_1 = \sum_{i=1}^{n} P(Y_i = 1)$, tj. aby počet všech realizací rovných 1 byl roven součtu všech modelem přidělených pravděpodobností jednotlivým případům,

2) $\sum_{\{i;y_i=1\}} x_{ir} = \sum_{i=1}^{n} P(Y_i = 1)x_{ir}$, což lze formálně upravit na $\sum_{\{i;y_i=1\}} x_{ir} = = n_1 \sum_{i=1}^{n} \frac{P(Y_i=1)}{n_1} x_{ir}$, tj. aby součet realizací veličiny X_r, pro něž je $y_i = 1$, byl úměrný váženému průměru těchto realizací, kde váhy jsou dány modelem přiřazenými pravděpodobnostmi jednotlivým případům, a konstantou úměrnosti je n_1.

Kapitola 3

Newtonova-Raphsonova metoda

Jedná se o známou iterační metodu používanou k numerickému řešení soustavy nelineárních rovnic. V této kapitole ukážeme její použití pro nalezení řešení systému věrohodnostních rovnic 2.19.

3.1 Definice a popis metody

Mějme systém nelineárních rovnic

$$\mathbf{F}(\mathbf{x}) = \mathbf{0} \tag{3.1}$$

přičemž $\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))'$ je vektorová funkce proměnné $\mathbf{x} = (x_1, \ldots, x_m)' \in \mathbb{R}^m$. Řešením této soustavy rozumíme vektor $\xi = (\xi_1, \ldots, \xi_m)' \in \mathbb{R}^m$, pro který platí $\mathbf{F}(\xi) = \mathbf{0}$.

Definujeme *Jacobiovu matici* $\mathbf{J_F}(\mathbf{x})$ funkce $\mathbf{F}$ jako matici parciálních derivací

$$\mathbf{J_F}(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_m} \end{pmatrix}.$$

Nechť je $\mathbf{J_F}(\mathbf{x})$ regulární se spojitými prvky v okolí bodu ξ. *Newtonovou-Raphsonovou (iterační) metodou* nazýváme metodu

$$\boxed{\mathbf{x}^{k+1} = \mathbf{x}^k - \mathbf{J_F}^{-1}\left(\mathbf{x}^k\right)\mathbf{F}\left(\mathbf{x}^k\right), \qquad k = 0, 1, 2, \ldots.} \tag{3.2}$$

O konvergenci posloupnosti $\{\beta^k\}_{k=0}^{\infty}$ vedoucí k řešení ξ pojednává následující věta.

Věta 3.1. *Nechť ξ je řešením rovnice $\mathbf{F}(\mathbf{x}) = \mathbf{0}$. Nechť $\mathbf{J_F}(\mathbf{x})$ je regulární matice se spojitými prvky v okolí $O(\xi)$ bodu ξ, přičemž $||\mathbf{J_F}^{-1}(\mathbf{x})||_\infty \leq K$, kde $K = \text{konst.}$, $\forall \mathbf{x} \in O(\xi)$. Nechť funkce $f_i(\mathbf{x})$, $i = 1, \ldots, m$, mají spojité druhé parciální derivace v $O(\xi)$.*

Posloupnost $\{\mathbf{x}^k\}_{k=0}^{\infty}$ určená Newtonovou-Raphsonovou iterační metodou konverguje ke kořenu ξ za předpokladu, že počáteční aproximace $\mathbf{x}^0$ leží dostatečně blízko ξ.

Tato věta předpokládá volbu „dobré“ počáteční aproximace, která bude dostatečně blízko řešení ξ. Ovšem pro systém nelineárních rovnic není snadné takovou aproximaci získat, jelikož pro soustavu $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ neexistují jednoduché a vždy konvergentní metody.

3.2 Newtonova-Raphsonova metoda pro systém věrohodnostních rovnic

V předchozí kapitole byl získán systém věrohodnostních rovnic (2.19) pro model logistické regrese

$$\mathbf{X}'(\mathbf{Y} - \vartheta(\beta)) = \mathbf{0}, \tag{3.3}$$

kde $(\mathbf{X})_{ij} = x_{ij}$, $(\mathbf{Y})_i = y_i$, $(\vartheta(\beta))_i = \vartheta_i = \frac{1}{1+e^{-\mathbf{x}_i'\beta}}$. Jedná se o soustavu $m+1$ rovnic o $m+1$ neznámých reprezentovaných vektorem $\beta = (\beta_0, \beta_1, \ldots, \beta_m)'$, přičemž vektorovou funkcí je $\mathbf{F}(\beta) = \mathbf{X}'(\mathbf{Y} - \vartheta(\beta))$. Význam této funkce je takový, že se jedná o gradient logaritmu věrohodnostní funkce.

Jacobiho maticí $\mathbf{J}_\mathbf{F}(\beta)$ je zde mínus Fisherova informační matice (viz (2.22))

$$\mathbf{J}_\mathbf{F}(\beta) = -\mathbf{J}(\beta) = -\mathbf{X}'\mathbf{V}(\beta)\mathbf{X},$$

kde $\mathbf{V}(\beta) = \operatorname{diag}(\vartheta_0(1-\vartheta_0), \ldots, \vartheta_m(1-\vartheta_m))$, a která, když $\mathbf{X}$ má lineárně nezávislé sloupce, je negativně definitní na $\mathbb{R}^{m+1}$ (jelikož $\mathbf{J}(\beta)$ je pozitivně definitní), tudíž je regulární, a tedy existuje inverzní matice $\mathbf{J}_\mathbf{F}^{-1}(\beta)$, $\forall \beta \in \mathbb{R}^{m+1}$.

Jelikož jsou splněny předpoklady věty 3.1 ($\mathbf{J}_\mathbf{F}(\beta)$ je regulární matice se spojitými prvky, prvky vektorové funkce $\mathbf{F}(\beta) = \mathbf{X}'(\mathbf{Y} - \vartheta(\beta))$ jsou spojité, a to dokonce na celém $\mathbb{R}^{m+1}$), lze k řešení systému (3.3) použít Newtonovu-Raphsonovu metodu.

Označme $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m)'$ řešení soustavy (3.3) a β^0 počáteční aproximace řešení pocházející z okolí $O(\hat{\beta})$. Potom Newtonova-Raphsonova metoda pro systém věrohodnostních rovnic je

$$\boxed{\beta^{k+1} = \beta^k + \left(\mathbf{X}'\mathbf{V}\left(\beta^k\right)\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{Y} - \vartheta\left(\beta^k\right)\right).} \tag{3.4}$$

Podle věty 3.3 konverguje posloupnost $\{\beta^k\}_{k=0}^{\infty}$, kde $\beta^k = (\beta_0^k, \beta_1^k, \ldots, \beta_m^k)'$, k řešení $\hat{\beta}$.

Tato iterativní procedura probíhá tak dlouho (tj. řekněme, že proběhne celkem N iterací), dokud „vzdálenost“ mezi dvěma řešeními daná vektorovou normou $\left|\beta^N - \beta^{N-1}\right|$ není menší než předem zadaná konstanta ε (např. $\varepsilon = 10^{-8}$). Potom β^N prohlásíme numerickým řešením systému (3.3).

Newtonova-Raphsonova metoda pro řešení soustavy věrohodnostních rovnic se označuje jako *Fisherova skórovací metoda* (angl. *Fisher's scoring method* nebo *Fisher's scoring algorithm*).

3.3 Možné obtíže při hledání numerického řešení

Při numerickém řešení věrohodnostních rovnic se můžeme, jako téměř u všech numerických metod, setkat s určitými překážkami, které znemožňují nalezení numerického řešení.

- Hodnota alespoň jednoho parametru se bude blížit k nekonečnu. Obvykle je to známka toho, že model je nevhodně specifikován nebo když numerické řešení není

stabilní kvůli malému rozsahu souboru. Parametr, který se blíží k nekonečnu, jistě nebude konvergovat. Někdy ovšem může být vhodné nechat model konvergovat, i když obsahuje tyto parametry. Identifikujeme-li pomocí nějaké podmínky, že hodnota parametru má tendenci divergovat (např. když standardní chyba odhadu je více než třikrát větší než je velikost samotného odhadu), ponecháme pro další iterace jeho hodnotu konstantní.

- Další problém se může objevit kvůli limitacím použití Newtonovy-Raphsonovy metody. Po určité, řekněme po r-té, iteraci může dojít k tomu, že hodnotu, k níž posloupnost dosud konvergovala, „přestřelí“ a tím může dojít k zacyklení, čímž proces nebude konvergovat. Takovou situaci lze ošetřit zadáním podmínky, aby hodnota logaritmické věrohodnostní funkce byla u následující iterace vyšší než u té předcházející, tj. $l\left(\beta^{i+1}\right) > l\left(\beta^{i}\right)$. Dojde-li k poklesu hodnoty logaritmické věrohodnostní funkce, existuje tak nebezpečí, že proces nebude konvergovat k lokálnímu maximu nebo dokonce nebude konvergovat vůbec. Jedním ze způsobů, jak se vypořádat s tímto problémem, je používat po získání odhadů β^{r} a β^{r+1} tzv. step-halving spočívající v pokračování posloupnosti danou předpisem $\beta^{r+k+2} = \beta^{r+k} + \frac{1}{2}\left(\beta^{r+k+1} - \beta^{r+k}\right)$ pro $k = 0,1,2,\ldots$, tzn. pokračujeme s body „ležícími mezi“ β^{r} a β^{r+1}. Jestliže hodnota logaritmu věrohodnostní funkce bude v každé následující iteraci. Pokračujeme tak dlouho, dokud je splněna podmínka $l(\beta^{r+k}) < l(\beta^{r+k+1})$. Celkem to znamená, že je třeba kontrolovat jak posloupnost $\beta^{0},\beta^{1},\beta^{2},\ldots$, tak i posloupnost logaritmů věrohodnostní funkce $l(\beta^{0}), l(\beta^{1}), l(\beta^{2}),\ldots$.

3.4 Iterativní vážená metoda nejmenších čtverců

Rovnici (3.4) lze formálně upravit do tvaru

$$\begin{aligned}\beta^{k+1} &= \beta^{k} + \left[\mathbf{X}'\mathbf{V}\left(\beta^{k}\right)\mathbf{X}\right]^{-1}\mathbf{X}'\left[\mathbf{Y} - \vartheta\left(\beta^{k}\right)\right] = \\ &= \left[\mathbf{X}'\mathbf{V}\left(\beta^{k}\right)\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{V}\left(\beta^{k}\right)\left[\mathbf{X}\beta^{k} + \mathbf{V}^{-1}\left(\beta^{k}\right)\left(\mathbf{Y} - \vartheta\left(\beta^{k}\right)\right)\right] = \\ &= \left[\mathbf{X}'\mathbf{V}\left(\beta^{k}\right)\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{V}\left(\beta^{k}\right)\mathbf{Z}\left(\beta^{k}\right), \qquad (3.5)\end{aligned}$$

kde $\mathbf{Z}\left(\beta^{k}\right) = \mathbf{X}\beta^{k} + \mathbf{V}^{-1}\left(\beta^{k}\right)\left(\mathbf{Y} - \vartheta\left(\beta^{k}\right)\right)$. Rovnice (3.5) je formálně shodná se vztahem pro výpočet nejlepšího nestranného odhadu u obecného lineárního modelu. Proto se této modifikaci říká *iterativní (pře)vážená metoda nejmenších čtverců* (angl. *iteratively (re)weighted least squares*). $\mathbf{X}$ je maticí plánu (stejně jako u lineární regrese), $\mathbf{V}\left(\beta^{k}\right)$ je matice vah a $\mathbf{Z}\left(\beta^{k}\right)$ stojí v roli vektoru pozorování.

Existuje řada dalších metod a modifikací, např. Böhningova metoda, iterativní vážení (iterative scaling), modifikované iterativní vážení (modified iterative scaling), metoda line search. Jejich použití může snížit potřebnou dobu pro nalezení řešení systému věrohodnostních rovnic (viz [11]).

Kapitola 4

Testování hypotéz u logistického regresního modelu

U lineárního regresního modelu se používá řada testů k ověření různých hypotéz týkajících se parametrů modelu. Jedním z důležitých předpokladů je předpoklad normality.

S výjimkou nezávislosti se u logistického regresního modelu nesetkáváme s nějakými speciálními typy předpokladů. Namísto „přesných“ testů se proto používají k ověřování hypotéz u logistické regrese testy *asymptotické*, které lze použít tehdy, je-li rozsah výběru dostatečně velký. Tyto testy jsou založeny na věrohodnostní funkci a jejich použití pro jednorozměrný i vícerozměrný parametr shrnují následující věty.

4.1 Testy o ML odhadech

Věta 4.1. *Nechť jsou splněny předpoklady věty 2.2 pro jednorozměrný parametrický prostor* Ω. *Označme*

$$LM(\theta_0) = \frac{[L'(\theta_0)]^2}{nJ(\theta_0)}, \tag{4.1}$$

$$U_{LM} = \frac{L'(\theta_0)}{\sqrt{nJ(\theta_0)}}, \tag{4.2}$$

$$W(\theta_0) = n(\hat{\theta}_n - \theta_0)^2 J(\hat{\theta}_n), \tag{4.3}$$

$$U_W = \sqrt{nJ(\hat{\theta}_n)}(\hat{\theta}_n - \theta_0), \tag{4.4}$$

$$LR(\theta_0) = 2[L(\hat{\theta}_n - L(\theta_0)]. \tag{4.5}$$

Potom $LM(\theta_0)$ *má asymptoticky* $\chi^2(1)$ *a* U_{LM} *má asymptoticky rozdělení* $N(0,1)$. *Je-li navíc funkce* $J(\theta)$ *spojitá v bodě* θ_0, *pak* U_W *má asymptoticky rozdělení* $N(0,1)$ *a* $W(\theta_0)$ *a* $LR(\theta_0)$ *mají asymptoticky rozdělení* $\chi^2(1)$.

Věta 4.2. *Nechť jsou splněny předpoklady věty 2.2 a nechť Fisherova informační matice*

$\mathbf{J}(\theta)$ *je spojitá v bodě* θ_0. *Označme*

$$LM(\theta_0) = \frac{1}{n}[\mathbf{U}(\theta_0)]'[\mathbf{J}(\theta_0)]^{-1}[\mathbf{U}(\theta_0)], \tag{4.6}$$

$$W(\theta_0) = n(\hat{\theta}_n - \theta_0)'\mathbf{J}(\hat{\theta}_n)(\hat{\theta}_n - \theta_0), \tag{4.7}$$

$$LR(\theta_0) = 2[L(\hat{\theta}_n) - L(\theta_0)]. \tag{4.8}$$

Potom každá z náhodných veličin $LM(\theta_0)$, $W(\theta_0)$, $LR(\theta_0)$ *má asymptoticky* $\chi^2(m)$ *rozdělení.*

Tato věta zůstává v platnosti, i když se místo matice $\mathbf{J}(\theta_0)$ použije nějaký její konzistentní odhad $\tilde{\mathbf{J}}(\theta_0)$. Též tvrzení založené na $W(\theta_0)$ platí, když se matice $\mathbf{J}(\hat{\theta}_n)$ nahradí nějakým konzistentním odhadem $\tilde{\mathbf{J}}_n$ matice $\mathbf{J}(\theta_0)$.

Věta 4.1 umožňuje testování hypotézy

$$H_0 : \theta = \theta_0 \qquad \times \qquad H_1 : \theta \neq \theta_0.$$

Používají se tři typy testů:

1. *Skórový (Raův) test*[1] je založený na $LM(\theta_0)$. Nulovou hypotézu zamítáme, když $LM(\theta_0) \geq \chi^2_{1-\alpha}(m)$. Skórový test má tu výhodu, že $LM(\theta_0)$ neobsahuje maximálně věrohodný odhad $\hat{\theta}_n$. Inverze k $\mathbf{J}(\theta_0)$ nebývá problém, jelikož dimenze m nebývá obvykle příliš velká.

2. *Waldův test* je založený na $W(\theta_0)$. Nulovou hypotézu zamítáme, když $W(\theta_0) \geq \chi^2_{1-\alpha}(m)$. U Waldova testu není třeba počítat inverzi k Fisherově informační matici, ovšem výpočet $\hat{\theta}_n$ i $\mathbf{J}(\hat{\theta}_n)$ může být náročný.

3. *Test založený na věrohodnostním poměru* je založený na $LR(\theta_0)$. Nulovou hypotézu zamítáme, když $LR(\theta_0) \geq \chi^2_{1-\alpha}(m)$. Tento test nevyžaduje znalost Fisherovy informační matice.

V případě jednorozměrného parametru se někdy k testování hypotézy $H_0 : \theta = \theta_0$ používá statistika U_{LM} místo $LM(\theta_0)$ a U_W místo $W(\theta_0)$. Je tomu tak proto, že testy založené na U_{LM} a U_W umožňují testovat H_0 i proti jednostranné alternativě $H_1 : \theta > \theta_0$ nebo $H_1 : \theta < \theta_0$.

4.2 Testování podmodelu

Uvažujme logistický regresní model M s odhady koeficientů $\mathbf{b}$ a podmodel $\tilde{M}$ s odhady koeficientů $\tilde{\mathbf{b}}$. Podmodel například dostaneme vyloučením některých regresorů. Zajímá nás, zda-li se M a M' významně liší. K tomuto účelu lze použít *test poměrem věrohodností*, který se provádí prostřednictvím tzv. deviancí, které si zavedeme.

Mějme nejbohatší možný model, který má tolik parametrů, kolik je různých hodnot vektorů $\mathbf{x}_i$ (tj. když $n = m$). Model, který by lépe prokládal data (tj. s větší hodnotou

[1] Dříve se mu říkalo *test založený na Lagrangeových multiplikátorech*.

věrohodnostní funkce), neexistuje. Tento model označme jako *saturovaný* a hodnotu jeho věrohodnostní funkce označme l_{max}. Každý další model je jeho podmodelem. Přiléhavost tohoto podmodelu lze posoudit pomocí *deviance*

$$D(\mathbf{b}) = 2(l_{max} - l(\mathbf{b})). \tag{4.9}$$

Čím je hodnota deviance větší, tím je přiléhavost podmodelu menší. Deviance je analogií reziduálního součtu čtverců u lineárního regresního modelu.

Odhadem středních hodnot u saturovaného modelu jsou přímo naměřené hodnoty y_i. Dle vztahu (2.13) lze přímo spočítat jeho hodnotu věrohodnostní funkce[2]:

$$l_{max} = \sum_{i=1}^{n} (y_i \ln y_i + (1-y_i)\ln(1-y_i)) = 0. \tag{4.10}$$

Devianci modelu pro podmodel saturovaného modelu tak bude

$$D(\mathbf{b}) = -2l(\mathbf{b}) = -2\sum_{i=1}^{n} \left[(y_i - 1)\mathbf{x}_i'\beta - \ln\left(1 + \mathrm{e}^{-\mathbf{x}_i'\beta}\right) \right]. \tag{4.11}$$

Chceme–li porovnat nějaký obecný model M a jeho podmodel $\tilde{M}$, použijeme test poměrem věrohodností, a to pomocí deviancí modelu a podmodelu. Testovou statistikou je

$$\begin{aligned} 2(l(\mathbf{b}) - l(\tilde{\mathbf{b}})) &= \left(2\left(l_{max} - l(\tilde{\mathbf{b}})\right)\right) - (2(l_{max} - l(\mathbf{b}))) = \\ &= D(\tilde{\mathbf{b}}) - D(\mathbf{b}). \end{aligned} \tag{4.12}$$

Tato testová statistika má za platnosti testovaného podmodelu asymptoticky rozdělení $\chi^2(q)$, kde q je rozdíl počtu nezávislých parametrů v porovnávaných modelech. Nulovou hypotézu H_0 : *$\tilde{M}$ je podmodelem modelu M* zamítáme na hladině významnosti α, když $D(\tilde{\mathbf{b}}) - D(\mathbf{b}) \geq \chi^2_{1-\alpha}(q)$.

Testování podmodelu se prakticky používá

1. pro testování významnosti celého modelu porovnáním daného modelu s tzv. nulovým modelem $P(Y=1|\mathbf{X}=\mathbf{x}) = \frac{1}{1+\mathrm{e}^{-b_0}}$, tj. testuje se hypotéza

$$H_0 : \beta_1 = \cdots = \beta_m = 0 \qquad \times \qquad H_1 : \text{alespoň jedno } \beta_i \neq 0,$$

2. pro ověření, zda nějaký podsoubor regresorů $\beta_{i_1}, \ldots, \beta_{i_k}$ (s výjimkou absolutního členu β_0) významně přispívá k vysvětlení variability závislé binární proměnné testováním hypotézy

$$H_0 : \beta_{i_1} = \cdots = \beta_{i_k} = 0 \qquad \times \qquad H_1 : \text{alespoň jedno } \beta_{i_k} \neq 0.$$

Hypotézu $H_0 : \beta_i = 0$ o nulovosti jednotlivých koeficientů modelu lze tedy testovat nejen pomocí Waldova testu, ale i testem poměrem věrohodností.

[2] $0.\ln(0) = 0$

4.3 Intervaly spolehlivosti

Waldův test založený na U_W umožňuje konstrukci intervalů spolehlivosti pro koeficienty modelu. Označíme-li $SE_i = \frac{1}{\sqrt{\{\mathbf{J}(\mathbf{b})\}_{ii}}}$ standardní chybu i-tého koeficientu modelu, potom $100(1-\alpha)\%$ *interval spolehlivosti pro* β_i je

$$\left(b_i - u_{1-\frac{\alpha}{2}} SE_i; b_i + u_{1-\frac{\alpha}{2}} SE_i\right). \tag{4.13}$$

Díky vztahu (1.11) mezi poměry šancí a koeficienty modelu je $100(1-\alpha)\%$ *interval spolehlivosti pro poměr šancí*

$$\left(\mathrm{e}^{b_i - u_{1-\frac{\alpha}{2}} SE_i}; \mathrm{e}^{b_i + u_{1-\frac{\alpha}{2}} SE_i}\right), \tag{4.14}$$

kde $u_{1-\frac{\alpha}{2}}$ je $1-\frac{\alpha}{2}$ kvantil standardizovaného normálního rozdělení.

Kapitola 5

Diagnostika

Máme-li již vytvořený logistický regresní model, potřebujeme posoudit, jak kvalitně prokládá data. K tomuto účelu se používá řady koeficientů, testů dobré shody i grafických pomůcek, jimž je věnována tato kapitola. Z grafických nástrojů je zmíněna ta nejpoužívanější, tzv. ROC křivka.

5.1 Skóre, prahový bod

Začněme s logistickým regresním modelem M se závislou proměnnou Y a nezávisle proměnnými $X_1, \ldots, X_m$. Tento model každému i-tému případu s realizací y_i přiřadí na základě jemu příslušných naměřených hodnot nezávislých proměnných $x_{i1}, \ldots, x_{im}$ pravděpodobnost, že tato realizace nabude hodnoty 1. Tato predikovaná pravděpodobnost se nazývá *skóre (logistické skóre)*

$$s_i = P(Y_i = 1 | \mathbf{X}_i = \mathbf{x}_i).$$

Skóre tak ohodnocuje každé měření zvlášť a umožňuje tak posoudit kvalitu modelu. Za ideální lze považovat takový model, který naměřeným hodnotám $y_i = 1$ přisoudí $s_i = 1$ a těm s $y_i = 0$ přisoudí $s_i = 0$. S takovou dokonalou direfenciací se ovšem v praxi nelze setkat, jelikož vždy mezi oběma skupinami, které se snažíme modelem odlišit, existuje prakticky téměř vždy určitý překryv (a navíc s_i může ležet pouze v intervalu $(0,1)$). Cílem je tak vytvořit model, který by se alespoň blížil k ideálnímu. Hodnoty skóre by tak pro většinu $y_i = 1$ měly být blízké 1 a většina hodnot s $y_i = 0$ by zase měla být blízká 0. U praktických modelů se lze pochopitelně setkat se situací, kdy pro nějaké měření bude skóre např. $s_i \approx 0,5$. Pak vzniká otázka, do které skupiny takový případ přiřadit. Z toho důvodu se zavádí pojem *prahového bodu* P_C (angl. *cutoff point* nebo též *cut-off point*). Pokud hodnota skóre s_i bude větší než P_C, zařadí se tento případ do skupiny „jedničkové“, pokud bude menší nebo roven P_C, dostane se do skupiny „nulové“. Hodnotu P_C lze pak volit na základě několika kritérií zmíněných dále.

5.2 Koeficienty

Zde uvedeme výčet několika koeficientů (též indexů), které číselně vyjadřují kvalitu proložení dat modelem. Budeme potřebovat pouze skutečné realizace hodnot $y_1, \ldots, y_n$ a jim

příslušné modelem přiřazená skóre $s_1, \ldots, s_n$. Nechť se v tomto souboru nachází n_0 hodnot s $y_i = 0$ a n_1 hodnot s $y_i = 1$.

5.2.1 Koeficienty založené na distribučních funkcích

Kolmogorovova-Smirnovova statistika

Zavedeme distribuční funkce skóre

$$F_0(x) = \frac{1}{n} \sum_{i=1} I(s_i \leq x, y_i = 0), \tag{5.1}$$

$$F_1(x) = \frac{1}{n} \sum_{i=1} I(s_i \leq x, y_i = 1), \tag{5.2}$$

kde funkce $I(A) = \begin{cases} 1 & ,A \text{ platí} \\ 0 & ,A \text{ neplatí} \end{cases}$.
$F_0(x)$ je distribuční funkcí „nulových“ případů, $F_1(x)$ těch „jedničkových“. Významnost jejich odlišnosti zjistíme pomocí Kolmogorova-Smirnovova testu založeného na statistice

$$KS = \sup_{x \in (0,1)} |F_0(x) - F_1(x)|, \tag{5.3}$$

která se prakticky spočítá jako $KS = \max_{x \in (0,1)} |F_0(x) - F_1(x)|$. Zjištěná hodnota KS se porovná s tabelovanou hodnotou $KS_{1-\alpha}(n_0, n_1)$. Je–li $KS \geq KS_{1-\alpha}(n_0, n_1)$, zamítneme hypotézu o shodě distribučních funkcí $F_0(x)$ a $F_1(x)$.

Kendallovo τ, Somersovo D, Goodmanovo-Kruskalovo γ

Řekneme, že dvě dvojice (y_i, s_i) a (y_j, s_j) tvoří

- *konkordantní pár*, když $\text{sign}(y_i - y_j) = \text{sign}(s_i - s_j)$
- *diskordantní pár*, když $\text{sign}(y_i - y_j) = -\text{sign}(s_i - s_j)$
- *vázaný pár*, když $y_i = y_j$ nebo $s_i = s_j$.

Pro logistickou regresi je u konkordantního páru hodnota přiřazeného skóre „jedničkovému“ případu vyšší než „nulovému“, u diskordantního páru je tomu naopak, u vázaného páru jsou hodnoty závislé proměnné nebo hodnoty skóre totožné.

Označme jako S_k počet konkordantních páru v souboru (o rozsahu n), S_d počet diskordantních párů, $S_{v,1}$ počet vázaných párů v prvním souboru $(y_1, \ldots, y_n)$, $S_{v,2}$ počet vázaných párů v druhém souboru $(s_1, \ldots, s_n)$ a $S_{v,12}$ počet vázaných párů v prvním i druhém souboru zároveň. Je zřejmé, že počet všech párů je $\frac{1}{2}n(n-1) = S_k + S_d + S_{v,1} + S_{v,2} + S_{v,12}$.

Následující koeficienty mají společný čitatel, ale liší se ve jmenovateli ([4], [13]):

$$\textbf{Kendallovo } \tau_a: \qquad \tau_a = \frac{S_k - S_d}{S_k + S_d + S_{v,1} + S_{v,2} + S_{v,12}}$$

$$\textbf{Kendallovo } \tau_b: \qquad \tau_b = \frac{S_k - S_d}{\sqrt{(S_k + S_d + S_{v,1})(S_k + S_d + S_{v,2})}}$$

$$\textbf{Somersovo } D_{asym}: \qquad D_{asym} = \frac{S_k - S_d}{S_k + S_d + S_{v,2}}$$

$$\textbf{Goodmanovo-Kruskalovo } \gamma: \qquad \gamma = \frac{S_k - S_d}{S_k + S_d}$$

Hodnoty těchto koeficientů leží mezi -1 a 1, přičemž jsou-li všechny pár konkordantní, mají všechny koeficienty hodnotu 1, jsou-li naopak všechny diskordantní, mají hodnotu -1. Budou-li koeficienty blízko 0, bude počet konkordantních a diskordantních párů přibližně stejný, tzn. model rozlišuje špatně nebo vůbec nerozlišuje mezi „nulovými" a „jedničkovými" případy.

Kendallovo τ_a oproti τ_b nezohledňuje vázané páry. Somersův koeficient D_{asym}[1] je modifikací τ_b a provádí korekci s ohledem na vazby pouze mezi hodnotami $s_1, \ldots, s_n$. Naproti tomu Goodmanovo-Kruskalovo γ vázané páry nebere v úvahu a nabývá hodnoty 1, když v souboru nejsou diskordantní páry, i když počet vazeb bude nenulový.

U logistické regrese lze očekávat, že počet $S_{v,1}$ bude mnohem větší než $S_{v,2}$ a $S_{v,12}$, jelikož první soubor je tvořen veličinami obsahujícími pouze 0 a 1, kdežto druhý nabývá hodnot z intervalu $(0,1)$. Proto by mělo platit, že $D_{asym} \approx \gamma$, $\tau_a \approx \tau_b$ a $D_{asym}, \gamma > \tau_a, \tau_b$. „Relativně velký" rozdíl mezi D_{asym} a γ může indikovat výskyt duplicit v souboru.

5.2.2 Koeficienty determinace

Vzhledem k podobnosti mezi deviancí u logistického regresního modelu a reziduálních součtem čtverců (RSS) u lineární regrese byly snahy rozšířit pojem koeficientu determinace také na logistickou regresi.

Uvažujme *nulový model*, tj. model, který predikuje vždy stejnou pravděpodobnost $P(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{1}{1+e^{-b_0}}$. Označme devianci tohoto modelu jako D_0. Porovnejme nulový model s nějakým jiným modelem, např. $P(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{1}{1+e^{-\mathbf{x}'\mathbf{b}}}$, s deviancí $D(\mathbf{b})$. Potom lze zavést (viz [2])

$$\textbf{McFaddenův koeficient determinace}: \qquad R^2_L = 1 - \frac{D(\mathbf{b})}{D_0}$$

$$\textbf{Coxův-Snellův koeficient determinace}: \qquad R^2_{CS} = 1 - e^{(D(\mathbf{b}) - D_0)/n}$$

$$\textbf{Nagelkerkův koeficient determinace}: \qquad R^2_N = \frac{1 - e^{(D(\mathbf{b}) - D_0)/n}}{1 - e^{-D_0/n}}.$$

Čím je model M více „vzdálen" od nulového modelu, tím jsou R^2_L, R^2_{CS} a R^2_N dále od nuly.

[1] Existuje ještě symetrická varianta $D_{sym} = \frac{S_k - S_d}{S_k + S_d + \frac{1}{2}(S_{v,1} + S_{v,2})}$, která se používá pro měření míry asociace mezi dvěma ordinálními veličinami, pokud nerozlišujeme, která z nich je závislá a která nezávislá.

McFaddenův koeficient se získá prostým dosazením deviance na místo příslušných součtu čtverců ve vztahu pro koeficient determinace u lineární regrese. Coxův-Snellův koeficient se dostane užitím vztahu pro logaritmickou věrohodnostní funkci u lineární regrese

$$l(\beta) = -\frac{n}{2}(1+\ln(2\pi)-\ln n)-\frac{n}{2}\ln RSS,$$

z nějž se vyjádří reziduální součet čtverců a dosadí do vztahu pro koeficient determinace. Jeho nevýhodou, že nemůže překročit hodnotu $1-\mathrm{e}^{-D_0/n}$, které je menší než jedna. Vydělení Coxova-Snellova koeficientu touto horní hranicí jej tak normujeme na jedničku, a tím dostaneme Nagelkerkův koeficient determinace.

5.2.3 Testy dobré shody

V této části uvedené testy jsou založeny na porovnání naměřených hodnot y_i a jim modelem přiřazených skóre s_i pomocí známého Pearsonova testu dobré shody. Všechny testové statistiky mají asymptoticky χ^2 rozdělení [14].

Pearsonův χ^2 test : $$\chi^2 = \sum_{i=1}^{n}\frac{(y_i-s_i)^2}{s_i(1-s_i)}$$

Rozdílový test deviance : $$D = -2\sum_{i=1}^{n}\left[s_i\ln\left(\frac{s_i}{1-s_i}\right)+\ln(1-s_i)\right]$$

Hosmerův-Lemeshowův test : $$C_g = \sum_{k=0}^{1}\sum_{g=1}^{G}\frac{(o_{kg}-e_{kg})^2}{e_{kg}}$$

Hosmerův-Lemeshowův test : $$H_g = \sum_{k=0}^{1}\sum_{g=1}^{G}\frac{(o'_{kg}-e'_{kg})^2}{e'_{kg}}$$

Pearsonův test je určen Pearsonovými rezidui $\frac{y_i-s_i}{\sqrt{s_i(1-s_i)}}$. Testové statistiky χ^2 a D mají asymptoticky rozdělení $\chi^2(n-m-1)$. Hodnoty realizací těchto dvou statistik se obvykle budou lišit. Nicméně větší pozornost by měla být věnována situaci, kdy tyto rozdíly budou velké. To může naznačovat, že použití těchto metod není vyhovující.

Hosmerův-Lemeshowův test založený na C_g spočívá ve vytvoření G skupin získaných skóre $s_1,\ldots,s_n$. Nejprve se všechna skóre uspořádají podle velikosti. Do první skupiny se vloží přibližně $\frac{n}{G}$ nejmenších hodnot s_i, do druhé též přibližně $\frac{n}{G}$ nejmenších hodnot ze zbývající skupiny skórů atd. Zpravidla se volí 10 skupin. Test je založen na očekávaní, že „nulové“ případy by se měly nacházet v nižších skupinách, kdežto „jedničkové“ případy v těch horních. Spočítají se četnosti obou typů případů v jednotlivých skupinách

$$o_{1g} = \sum_{i=1}^{n_g} y_i \qquad o_{0g} = \sum_{i=1}^{n_g}(1-y_i)$$

a zjistí se očekávané četnosti

$$e_{1g} = \sum_{i=1}^{n_g} s_i \qquad e_{0g} = \sum_{i=1}^{n_g}(1-s_i),$$

kde $g = 1,\ldots,G$, které se porovnají pomocí testové statistiky C_g. V případě „dobrého“ proložení dat modelem má $C_g \approx \chi^2(G-2)$.

Hosmerův-Lemeshowův test založený na H_g se liší od C_g pouze způsobem vytvoření skupin. U H_g se postupuje tak, že se tvoří $G = 10$ skupin[2] podle skóre $I_1 = [0;0,1)$, $I_2 = [0,1;0,2)$, …, $I_{10} = [0,9;1]$ a spočítají se hodnoty

$$o'_{1g} = \sum_{\{i:s_i \in I_g\}} y_i \qquad o'_{0g} = \sum_{\{i:s_i \in I_g\}} (1-y_i)$$

a

$$e'_{1g} = \sum_{\{i:s_i \in I_g\}} s_i \qquad e'_{0g} = \sum_{\{i:s_i \in I_g\}} (1-s_i),$$

$g = 1,\ldots,10$. Obdobně jako u C_g má v případě „dobrého“ proložení dat modelem $H_g \approx \chi^2(G-2) = \chi^2(8)$.

Testy zde uvedené patří k nejpoužívanějším. Vedle nich existuje ještě celá řada dalších (např. dalších 5 typů Hosmerových-Lemeshowových testů, Brownův test, Stukelův test, Tsiatisův test), o nichž se lze více dočíst např. v [14].

5.3 Informační kritéria

S rostoucím počtem regresorů (nejen) u logistického regresního modelu roste též hodnota logaritmické věrohodnostní funkce, což je na jednu stranu výhodné, jelikož tak „důvěryhodnost“ modelu roste. Na druhou ovšem velký počet regresorů znamená potřebu kontrolovat více proměnných, což nemusí být vždy výhodné (např. z ekonomického hlediska). Proto byla vyvinuta řada tzv. informačních kritérií, které hodnotu logaritmu věrohodnostní funkce pro daný model *penalizují* s ohledem na počet použitých regresorů pro výstavbu modelu a tím umožní usnadnit výběr modelu. Za takto vhodně „vyvážený“ model se považuje ten, který při porovnání s ostatními dosahuje nejnižší hodnoty informačního kritéria.

Označme l hodnotu logaritmické věrohodností funkce nějakého modelu M, který má m regresorů a je vytvořen ze souboru o rozsahu n. Zde jsou uvedeny nejznámější typy informačních kritérií pro takový model:

Akaikeovo informační kritérium:	$AIC = -2l + 2m$
Bayesovo informační kritérium:	$BIC = -2l + m\ln n$
Hannanovo-Quinnovo informační kritérium:	$HQ = -2l + m\ln(\ln n)$.

5.4 ROC křivka

Zkratka ROC pochází z anglického názvu *receiver operating characteristic*. Pojem ROC křivka lze přeložit jako *křivka (graf) operační prahové charakteristiky*. V logistické regresi

[2] Lze zobecnit i pro jiný počet skupin, nejčastěji se používá právě 10. Jiný počet se využije zejména tehdy, bude–li některá ze skupin méně obsazena.

se používá k hodnocení kvality vytvořeného modelu. Používá se mj. k detekci signálu, který nebylo možné vždy správně přijmout.

5.4.1 Senzitivita a specificita

S konstrukcí ROC křivky jsou úzce spjaty pojmy *senzitivita* a *specificita*. Mějme celkem N objektů, které lze rozdělit do dvou skupin A a B (např. pacienti s ischemickou chorobou srdeční a bez ní, klienti banky řádně splácející úvěr a klienti banky, kteří se opožďují se splátkou). Označme N_A počet objektů ve skupině A a N_B objektů náležející skupině B. Dále mějme klasifikační proces, který umožňuje zařadit daný objekt na základě jeho určitých charakteristik (např. u pacientů pohlaví, věk, BMI, rodinná anamnéza, u klientů banky pohlaví, věk, zaměstnání, socioekonomický status) buď do skupiny A, nebo do skupiny B. Zajímá nás, jak „přesný“ je tento rozhodovací algoritmus, tzn. pokud je objekt zařazen do jedné ze skupin, jaká je pravděpodobnost, že do něj skutečně patří? Proto se definují

- *senzitivita (citlivost)* jako pravděpodobnost, že objekt, který byl zařazen do skupiny A, do skupiny A skutečně patří,
- *specificita* jako pravděpodobnost, že objekt, který byl zařazen do skupiny B (tj. nebyl zařazen do A), do skupiny B skutečně patří (tj. nepatří do A).

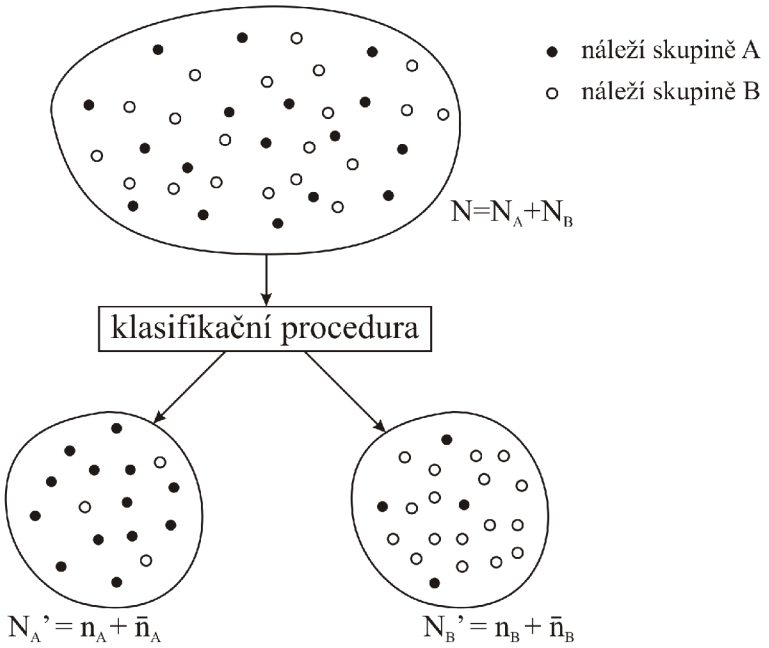
Za ideální lze považovat takový klasifikační algoritmus, pro který budou senzitivita a specificita rovny 1.

Dejme tomu, že klasifikační procedura celkový soubor N objektů rozdělí do obou skupin tak, že do skupiny A zařadí N'_A objektů a do skupiny B jich řadí N'_B. Z celkového počtu N'_A jich ve skutečnosti n_A patří do A a $\bar{n}_A$ je chybně zařazených (patří do B). Analogicky pro N'_B (tj. z nich n_B patří ve skutečnosti do B a $\bar{n}_B$ patří do A). Pro přehled zapišme vztahy mezi těmito počty

$$N = N_A + N_B = N'_A + N'_B$$

$$N'_A = n_A + \bar{n}_A \qquad N_A = n_A + \bar{n}_B$$

$$N'_B = n_B + \bar{n}_B \qquad N_B = n_B + \bar{n}_A.$$



Obrázek 5.1: Schematické zobrazení klasifikace objektů.

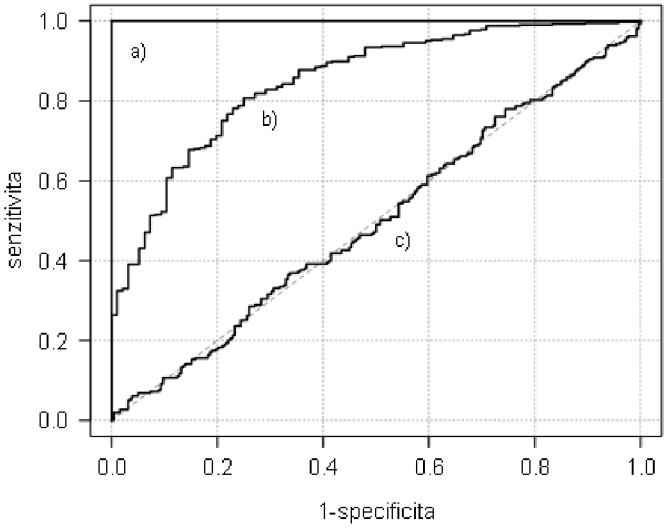
Senzitivitu a specificitu pak odhadneme jako podíly

$$\text{sens} = \frac{n_A}{N_A} \qquad \text{spec} = \frac{n_B}{N_B}, \tag{5.4}$$

tj. odhad senzitivity je poměr mezi počtem správně klasifikovaných objektů skupiny A vůči počtu všech objektů z A, odhad specificity je pak analogicky poměr mezi počtem správně zařazených objektů do skupiny B vůči počtu všech objektů z B.

Logistickou regresí lze tímto způsobem provést zpětnou klasifikaci objektů pro ohodnocení její kvality. K tomu účelu je podstatná volba prahového bodu P_C. Pro různé hodnoty prahového bodu dosáhneme různých hodnot senzitivity a specificity. Graficky se jejich vztah reprezentuje právě pomocí ROC křivky. Jedná se o *graf závislosti senzitivity na* $1 -$ *specificitě*. Použijeme-li stejné označení pro hustoty skóre „nulových“ a „jedničkových“ případů jako je (5.1) a (5.2), lze ROC křivku definovat jako množinu $\{(1-F_0(x), F_1(x)); x \in [0,1]\}$.

Pro ideální model má ROC křivka tvar lomené čáry procházející body $[0;0]$, $[0;1]$ a $[1;1]$. Pro náhodný model se ROC křivka nachází kolem úsečky spojující body $[0;0]$ a $[1;1]$. ROC křivka pro reálný model by měla být pokud možno co nejblíže ke křivce pro ideální model.



Obrázek 5.2: ROC křivky pro a) ideální model, b) reálný model, c) náhodný model.

Prahový bod lze zvolit tak, aby byla splněna některá z následujících podmínek:

- dosažení požadované senzitivity testu
- dosažení požadované specificity testu
- maximalizace součtu senzitivity a specificity
- aby euklidovská vzdálenost mezi levým horním bodem $[0;1]$ a ROC křivkou byla co nejmenší.

Rozlišovací schopnost modelu se nejčastěji ohodnocuje pomocí C statistiky, která vyjadřuje *plochu pod ROC křivkou* (označuje se též jako AUC, zkratka z angl. *area under curve*). Pro ideální model je $AUC = 1$, pro náhodný model je $AUC \approx 0,5$ (viz Obrázek).

Kapitola 6

Model pro aterosklerózu

V této kapitole ukážeme na příkladu klinických dat použití metody logistické regrese, a to s využitím statistického softwaru *R*, verze 2.15.2.

6.1 Popis souboru

6.1.1 Úvod

Ateroskleróza (CAD, z angl. Coronary Artery Disease) je název pro onemocnění projevující se kornatěním tepen, které vzniká v důsledku ukládání tukových látek do stěn tepny. Mezi známé faktory ovlivňující tento proces patří zvýšená hladina celkového cholesterolu a LDL, kouření, hypertenze, diabetes mellitus, fyzická inaktivita, obezita, dědičné predispozice, VLDL částice, homocystein, C-reaktivní protein atd.
Matrix metalloproteinázy (zkr. MMP) je skupina 28 důležitých enzymů podílejících se na remodelaci základních buněčných membrán a složek extracelulární matrix. Tyto enzymy jsou zapojeny do mnoha fyziologicko-patologických procesů jako např. růst a remodelace kostí, hojení, nádorové bujení, artritida. Hladiny MMP jsou v dospělých tkáních téměř nedetekovatelné, ale při zranění, nemoci nebo těhotenství je jejich exprese zvýšená. U MMP 13 (též kolagenáza 3) v patologických stavech, kdy je potřeba oprava nebo remodelace tkáně, dochází ke zvýšení její exprese, čímž se MMP 13 podílí na změnách struktury a složení kolagenové matrix, což vede k remodelaci levé komory srdce. Zvýšené hladiny MMP 13 byly též pozorovány u pacientů s aterosklerózou.

6.1.2 Cíl práce

Cílem je na základě získaných dat vytvořit vhodný logistický regresní model pro predikci aterosklerózy u pacientů, u nichž je podezření na toto onemocnění, a dále zjistit, zda polymorfismus na lokusu rs640198 genu kódujícího tvorbu MMP 13 je významným faktorem ovlivňujícím vznik aterosklerózy.

6.1.3 Popis dat

Data pocházejí ze studie, kterou realizoval Ústav patologické fyziologie Masarykovy univerzity a 1. interní kardioangiologická klinika Fakultní nemocnice u sv. Anny v Brně. Studie se zúčastnilo celkem 1071 pacientů s podezřením na přítomnost aterosklerózy. Pacienti byli krátkodobě hospizalizováni na 1. interní kardioangiologické klinice Fakultní nemocnice u sv. Anny v Brně v období mezi říjnem 2005 a únorem 2007 a podstoupili kompletní kardiologické vyšetření. Pomocí koronární angiografie byla prokázána nebo vyvrácena přítomnost aterosklerózy[1].
Toto vyšetření potvrdilo u 845 pacientů přítomnost aterosklerózy a u 226 pacientů ji vyloučilo.
Od každého pacienta je k dispozici 27 veličin, z toho 16 spojitých nezáporných, 1 diskrétní a 10 kategoriálních, z nichž 8 je dichotomických, 1 ordinální a 1 multinomická. Hodnoty veličin byli získány od pacientů při nástupu na vyšetření. Jejich názvy, kódování v této práci, jednotky a vysvětlení shrnuje následující tabulka:

Spojité proměnné

Kód v *R*	Název	Jednotka	Vysvětlení
`vek`	věk	l	věk pacienta
`vyska`	výška	cm	výška pacienta
`vaha`	váha	kg	váha pacienta
`BMI`	BMI	kg/m^2	index tělesné hmotnosti, slouží jako měřítko obezity, BMI=(váha v kg)/(výška v m)^2
`EF`	ejekční frakce	bez rozměru	podíl systolického objemu a end–diastolického objemu
`TF`	tepová frekvence	min^{-1}	počet tepů srdce za minutu
`TKs`	systolický krevní tlak	mm Hg	hodnota systolického krevního tlaku
`TKd`	diastolický krevní tlak	mm Hg	hodnota diastolického krevního tlaku
`leu`	leukocyty	μl^{-1}	počet bílých krvinek v tisících v 1 μl krve (norma je 4–10 μl^{-1})
`trombo`	trombocyty	μl^{-1}	počet krevních destiček v tisících v 1 μl krve (norma je 150–400 μl^{-1})
`fibrinogen`	fibrinogen	g/l	glykoprotein nezbytný při srážení krve, jeho koncentrace v krvi stoupá při zánětu nebo poškození tkání, je mediátorem agregace krevních destiček (norma 1,5–3,0 g/l)

[1] Za klinicky významné postižení tepny se považuje zúžení jejího průsvitu ve dvourozměrném zobrazení o více než 50 %.

Kód v *R*	Název	Jednotka	Vysvětlení
`chol`	celkový cholesterol	mmol/l	steroidní látka potřebná v lidském těle pro tvorbu hormonů a vitamínu D, pomáhá zpracovávat tuky, je důležitý při tvorbě buněčných membrán, jeho vysoká koncentace s sebou nese zdravotní rizika, zejména onemocnění srdce (norma je 0,00–5,00 mmol/l, zvýšená hladina je 5,01–6,50 mmol/l)
`HDL`	HDL	mmol/l	vysokodenzitní lipoprotein, „hodný“ cholesterol, jeho vysoký podíl (vzhledem k LDL) je známkou dobré schopnosti vyloučit nadbytečný cholesterol z organismu
`LDL`	LDL	mmol/l	nízkodenzitní lipoprotein, „zlý“ cholesterol, jeho zvýšená hladina (nad 3 mmol/l) způsobuje usazování nadbytečného cholesterolu v cévních stěnách
`glykemie`	glykémie	mmol/l	hladina glukózy v krvi (norma je 3,3–5,5 mmol/l, u diabetiků je horní hranice 6,0 mmol/l na lačno a 7,5 mmol/l po jídle
`CRP`	C–reaktivní protein	mg/l	protein syntetizovaný v játrech, slouží jako indikátor infekčních onemocnění, u zdravého člověka jsou jeho hladiny nízké (do 6 mg/l), hodnota CRP v rozmezí 6–40 mg/l odpovídá spíše virové infekci, hodnota CRP nad 40 mg/l spíše bakteriální infekci, mírně zvýšená hladina CRP patří mezi známky vysokého kardiovaskulárního rizika

Tabulka 6.1: Spojité proměnné v souboru pacientů.

Diskrétní proměnné

Kód	Název	Vysvětlení
`no_stenoz`	počet stenóz	počet zúžených částí (stenóz) věnčitých tepen

Tabulka 6.2: Diskrétní proměnné v souboru pacientů.

Kategoriální proměnné

Kód	Název	Typ proměnné	Vysvětlení
`CAD`	ateroskleróza	dichotomická	viz informace v úvodu u popisu dat
`sex`	pohlaví	dichotomická	zda je pacient žena nebo muž
`koureni`	kouření	dichotomická	zda je pacient kuřák (i bývalý) nebo ne
`hypertenze`	hypertenze	dichotomická	vysoká hladina krevního tlaku (opakovaně nad 140/90 mm Hg)
`diabetes`	diabetes mellitus	dichotomická	chronické onemocnění projevující se poruchou metabolizmu sacharidů
`renal.ins`	renální insuficience	dichotomická	selhání ledvin, onemocnění spočívající ve ztrátě schopnosti ledvin vylučovat z těla odpadní látky
`ACEi`	ACE inhibitory	dichotomická	léky používané k léčbě hypertenze, jako prevence při onemocnění ledvin u diabetiků
`beta.blok`	beta blokátory	dichotomická	léky používané k léčbě kardiovaskulárních onemocnění (hypertenze, ischemická choroba srdeční, při srdečním selhání)
`statin`	statiny	dichotomická	léky snižující hladinu některých lipidů v krvi
`tepny`	tepny	ordinální	počet věnčitých tepen (tepny zásobující srdce) postižených aterosklerózou
`MMP13`	MMP 13	multinomická	polymorfismus na lokusu rs640198 genu pro MMP 13

Tabulka 6.3: Kategoriální proměnné v souboru pacientů.

Dichotomické veličiny nabývají hodnot 0 a 1, kde 0 značí logickou hodnotu FALSE a 1 logickou hodnotu TRUE (např. pacient s hodnotou hypertenze 0 hypertenzí netrpí a pacient s hodnotou hypertenze 1 ano).

Veličiny `no_stenoz` a `tepny` jednoznačně určují veličinu `CAD` (pacienti s nulovým počtem stenóz mají nulový počet postižených částí tepen, a tudíž se jedná o pacienty bez aterosklerózy).

6.2 Exploratorní analýza

Před samotným vytvořením logistického regresního modelu byla provedena exploratorní analýza dat s cílem detekovat chybějící a extrémní hodnoty, ověřit normalitu, prozkoumat závislosti mezi spojitými proměnnými pomocí korelační analýzy a možné asociace

mezi kategoriálními proměnnými pomocí testů nezávislosti v kontingenčních tabulkách. Hladina významnosti v celé následující analýze činí $\alpha = 0,05$.

Detekce extrémních hodnot

Pomocí *krabicových grafů* (funkce `boxplot`) byly zjištěny následující extrémní hodnoty:

Veličina	Extrémní hodnoty
`vek`	1 5 6 12
`fibrinogen`	218
`leu`	127 136
`trombo`	9,4

Tabulka 6.4: Zjištěné extrémní hodnoty v souboru.

Jelikož se s největší pravděpodobností jedná o chybné údaje, byly tyto hodnoty z data odstraněny (tj. v R nahrazeny hodnotou NA).

Chybějící hodnoty

V datovém souboru se též nachází chybějící hodnoty, jejichž počet pro jednotlivé proměnné shrnuje následující tabulka:

Proměnná	Počet chybějících hodnot	Proměnná	Počet chybějících hodnot	Proměnná	Počet chybějících hodnot
`vek`	4	`fibrinogen`	34	`hypertenze`	0
`vyska`	7	`chol`	11	`diabetes`	0
`vaha`	6	`HDL`	12	`renal.ins`	9
`BMI`	7	`LDL`	38	`ACEi`	0
`EF`	0	`glykemie`	21	`beta.blok`	0
`TF`	0	`ln.CRP`	38	`statin`	0
`TKs`	0	`CAD`	0	`no.stenoz`	0
`TKd`	0	`MMP13`	0	`tepny`	0
`leu`	8	`sex`	0		
`trombo`	8	`koureni`	0		

Tabulka 6.5: Zjištěné chybějící hodnoty v souboru.

Ověření normality

Pomocí *histogramů* (funkce `hist`) byl prozkoumán přibližný tvar rozdělení jednotlivých proměnných. Veličina CRP vykazovala velké zešikmení směrem doleva, ovšem logaritmickou transformací bylo dosaženo symetričtějšího rozdělení, proto se v další analýze pracuje s přirozeným logaritmem CRP místo s CRP, pro který je použito označení `ln.CRP`.

Normalita byla ověřena jednak pro všechny pacienty, jednak pro každou skupinu pacientů (bez CAD a s CAD) zvlášť, a to pomocí *kvantil-kvantilového grafu* (funkce `qqnorm`

a `qqline`) a *Pearsonových* χ^2 *testem* (funkce `pearson.test` z knihovny `nortest`). Grafickým zhodnocením lze dospět k závěru, že rozdělení následujících veličin se příliš neliší od normálního:

u všech pacientů: `vek`, `vyska`, `vaha`, `BMI`, `TKs`, `fibrinogen`, `ln.CRP`
u pacientů bez CAD: `vek`, `vyska`, `vaha`, `BMI`, `TKs`, `trombo`, `fibrinogen`, `ln.CRP`
u pacientů s CAD: `vek`, `vyska`, `vaha`, `BMI`, `TKs`, `fibrinogen`, `ln.CRP`

Co se týče proměnných neuvedených v tabulce, na základě tvaru jejich histogramů a hodnot výběrových šikmostí vyplývá, že vykazují rozdělení mírně zešikmené směrem doprava (s výjimkou EF se zešikmením směrem doleva).

Hypotéza o normalitě pomocí Pearsonova χ^2 testu nebyla zamítnuta u veličin

u všech pacientů: –
u pacientů bez CAD: `vek`, `vaha`, `BMI`, `trombo`, `fibrinogen`, `chol`, `LDL`, `ln.CRP`
u pacientů s CAD: –

Korelace

Pomocí *Spearmanova korelačního koeficientu* (funkce `cor.test` s argumentem `method="spearman"`) byla vyšetřena lineární závislost mezi proměnnými. Celkem bylo zjištěno 65 významných korelací (ze 105 možných). Dvojice s nejvyššími hodnotami velikosti korelačního koeficientu jsou uvedeny v následující tabulce.

Korelace mezi	R	p
`chol`–`LDL`	$0,9314$	$< 0,001$
`vaha`–`BMI`	$0,7712$	$< 0,001$
`fibrinogen`–`ln.CRP`	$0,6154$	$< 0,001$
`TKs`–`TKd`	$0,5965$	$< 0,001$
`vyska`–`vaha`	$0,5840$	$< 0,001$

Tabulka 6.6: Nejvyšší hodnoty korelací mezi spojitými proměnnými v souboru.

Velikosti ostatních korelačních koeficientů byly menší než $0,37$, a tudíž se slabou nebo zanedbatelnou mírou lineární závislosti.

Meziskupinová porovnání

K porovnání rozdílů skupin pacientů s CAD a bez CAD byl použit *Mannův–Whitneyův test*. Zjištěné statisticky významné rozdíly jsou uvedeny v následující tabulce (Δ je rozdíl mediánů mezi skupinami pacientů s CAD a bez CAD).

Významné rozdíly u veličiny	Δ	p
vek	3 l	$< 0{,}001$
vyska	2 cm	$< 0{,}001$
EF	-5	$< 0{,}001$
TKs	10 mm Hg	$< 0{,}001$
leu	$1{,}0\,\mu\text{l}^{-1}$	$< 0{,}001$
fibrinogen	0,3 g/l	$< 0{,}001$
chol	$-0{,}32$ mmol/l	$< 0{,}001$
HDL	$-0{,}18$ mmol/l	$< 0{,}001$
LDL	$-0{,}21$ mmol/l	$< 0{,}001$
glykemie	0,40 mmol/l	0,002
ln.CRP	0,37	$< 0{,}001$

Tabulka 6.7: Významné rozdíly mediánů veličin mezi pacienty s CAD a bez CAD.

Testování nezávislosti v kontingenčních tabulkách

Pearsonovým χ^2 testem nezávislosti byla zjišťována významnost vztahu mezi 9 kategorickými proměnnými CAD, MMP13, sex, koureni, hypertenze, diabetes, renal_ins, ACEi, beta_blok a statin. Cramérovým koeficientem kontingence byla měřena míra asociace. Bylo nalezeno 26 statisticky významných vztahů (z 36 možných). Nejvyšší hodnota Cramérova koeficientu kontingence je u dvojice CAD a statin, druhá nejvyšší pak u dvojice CAD a beta_blok. Zbývající hodnoty byly menší než $0,3$, a tudíž tyto míry asociace jsou slabé nebo zanedbatelné. Uveďme ještě výsledky pro vztah mezi CAD a zbývajícími 8 veličinami:

Vztah mezi CAD a	χ^2	df	p	Cramérovo V
MMP13	5,84	2	0,0540	$0{,}0738$
sex	40,34	1	$< 0{,}001$	$0{,}1941$
koureni	0,22	1	0,6423	$0{,}0142$
hypertenze	2,69	1	0,1009	$0{,}0501$
diabetes	20,00	1	$< 0{,}001$	$0{,}1367$
renal_ins	7,21	1	0,0073	$0{,}0824$
ACEi	44,85	1	$< 0{,}001$	$0{,}2046$
beta_blok	156,25	1	$< 0{,}001$	$0{,}3820$
statin	370,19	1	$< 0{,}001$	$0{,}5879$

Tabulka 6.8: Významné asociace mezi CAD a kategoriálními proměnnými.

6.3 Logistický regresní model

Díky silné korelaci mezi chol a LDL, vaha a BMI byly veličiny LDL a vaha z modelování vyloučeny. Nebude-li uvedeno jinak, pochází uvedené funkce ze základních knihoven sta-

tistického softwaru R (např. `base`). Jako výchozí jsme zvolili model

$$\begin{aligned}\text{logit}(P(CAD_{ijklmnopqr}=1)) = \beta_0 &+ sex_j + koureni_k + hypertenze_l + diabetes_m + \\ &+ renal.ins_n + ACEi_o + beta.blok_p + statin_q + MMP13_r + \\ &+ \beta_1 vek_i + \beta_2 vyska_i + \beta_3 BMI_i + \beta_4 EF_i + \beta_5 TF_i + \\ &+ \beta_6 TKs_i + \beta_7 leu_i + \beta_8 trombo_i + \beta_9 fibrinogen_i + \\ &+ \beta_{10} chol_i + \beta_{11} HDL_i + \beta_{12} glykemie_i + \beta_{13} ln.CRP_i,\end{aligned}$$

který realizujeme v R pomocí funkce `glm` (funkce pro zobecněné lineární modely) syntaxí

```
> model.cad <- glm(CAD~sex+koureni+hypertenze+diabetes+renal.ins+
                     ACEi+beta.blok+statin+MMP13+vek+vyska+BMI+EF+TF+
                     TKs+leu+trombo+fibrinogen+chol+HDL+glykemie+
                     ln.CRP,data=data.cad,family=binomial)
```

a který je uložen v proměnné `model.cad`. V důsledku chybějících hodnot je model vystavěn ze souboru 941 pacientů majících úplné informace o hodnotách veličin vytvářejících model.

```
> summary(model.cad)

Call:
glm(formula = CAD ~ sex + koureni + hypertenze + diabetes +
    renal.ins + ACEi + beta.blok + statin + MMP13 + vek + vyska +
    BMI + EF + TF + TKs + leu + trombo + fibrinogen + chol + HDL +
    glykemie + ln.CRP, family = binomial, data = data.cad)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9632   0.1425   0.2861   0.4490   2.5534

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.377945   4.042083   0.341 0.733179
sexžena     -1.114553   0.352783  -3.159 0.001581 **
koureni1    -0.019158   0.344035  -0.056 0.955591
hypertenze1 -0.308742   0.311750  -0.990 0.322003
diabetes1    0.206734   0.327602   0.631 0.528006
renal.ins1  -0.246129   0.304521  -0.808 0.418946
ACEi1        0.235869   0.237285   0.994 0.320208
beta.blok1   1.869730   0.286618   6.523 6.87e-11 ***
statin1      3.032489   0.310808   9.757  < 2e-16 ***
MMP13GT      0.165393   0.238060   0.695 0.487209
MMP13TT      0.241751   0.424603   0.569 0.569114
vek          0.037570   0.013626   2.757 0.005828 **
vyska       -0.040939   0.018202  -2.249 0.024502 *
```

```
BMI          -0.011617   0.031084  -0.374 0.708611
EF           -0.028315   0.010600  -2.671 0.007555 **
TF            0.002034   0.011502   0.177 0.859648
TKs           0.017312   0.007370   2.349 0.018825 *
leu           0.232534   0.069927   3.325 0.000883 ***
trombo       -0.001763   0.002249  -0.784 0.433100
fibrinogen    0.076858   0.209647   0.367 0.713913
chol          0.073464   0.120739   0.608 0.542885
HDL          -1.513454   0.422454  -3.583 0.000340 ***
glykemie      0.037346   0.086280   0.433 0.665124
ln.CRP        0.044557   0.124514   0.358 0.720459
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 960.31  on 940  degrees of freedom
Residual deviance: 548.72  on 917  degrees of freedom
AIC: 596.72

Number of Fisher Scoring iterations: 6
```

Tento model byl redukován zpětnou eliminací pomocí funkce `step`, založené na hodnotě Akaikeova informačního kritéria AIC, a výstup uložen do proměnné `model.cad.step`:

```
> model.cad.step <- step(model.cad)
> summary(model.cad.step)

Call:
glm(formula = CAD ~ sex + beta.blok + statin + vek + vyska +
    EF + TKs + leu + HDL, family = binomial, data = data.cad)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9398   0.1522   0.2964   0.4401   2.4469

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.29933    3.50854   0.370 0.711134
sexžena     -1.11721    0.33907  -3.295 0.000984 ***
beta.blok1   1.90324    0.28257   6.735 1.64e-11 ***
statin1      3.07017    0.29998  10.235  < 2e-16 ***
vek          0.03804    0.01193   3.190 0.001423 **
vyska       -0.03877    0.01773  -2.187 0.028736 *
EF          -0.03302    0.01006  -3.281 0.001034 **
TKs          0.01779    0.00712   2.499 0.012468 *
```

```
leu          0.23825    0.06265   3.803 0.000143 ***
HDL         -1.43588    0.34889  -4.116 3.86e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 960.31  on 940  degrees of freedom
Residual deviance: 554.14  on 931  degrees of freedom
AIC: 574.14

Number of Fisher Scoring iterations: 6
```

Původní model, obsahující 22 proměnných, byl redukován na model tvořený 9 proměnnými. Testem poměrem věrohodností ověříme jednak významnost tohoto modelu a jednak jej porovnáme s původním modelem, a to pomocí funkce `lrtest` z knihovny `lmtest`:

```
> lrtest(model.cad.step)
Likelihood ratio test

Model 1: CAD ~ sex + beta.blok + statin + vek + vyska + EF + TKs +
    leu + HDL
Model 2: CAD ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  10 -277.07
2   1 -480.15 -9 406.17  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> lrtest(model.cad,model.cad.step)
Likelihood ratio test

Model 1: CAD ~ sex + koureni + hypertenze + diabetes + renal.ins +
    ACEi + beta.blok + statin + MMP13 + vek + vyska + BMI + EF + TF +
    TKs + leu + trombo + fibrinogen + chol + HDL + glykemie +
    ln.CRP
Model 2: CAD ~ sex + beta.blok + statin + vek + vyska + EF + TKs +
    leu + HDL
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  24 -274.36
2  10 -277.07 -14 5.4236      0.979
```

Tedy redukovaný model je významně odlišný od nulového modelu a významně se neliší od modelu původního.

O všech proměnných tvořících redukovaný model lze očekávat asociaci s CAD, jak ostatně ukazují výsledky exploratorní analýzy. Z praktického hlediska se ovšem veličina `vyska` zdá být „podezřelou“. Proto nejprve vytvoříme 9 modelů pro `CAD`, kde v roli nezávislé proměnné budou vystupovat postupně všechny veličiny z redukovaného modelu:

```
> data.cad2 <- na.omit(data.cad)
> model.sex <- glm(CAD~sex,data=data.cad2,family=binomial)
> model.bb <- glm(CAD~beta.blok,data=data.cad2,family=binomial)
> model.stat <- glm(CAD~statin,data=data.cad2,family=binomial)
> model.vek <- glm(CAD~vek,data=data.cad2,family=binomial)
> model.vys <- glm(CAD~vyska,data=data.cad2,family=binomial)
> model.EF <- glm(CAD~EF,data=data.cad2,family=binomial)
> model.TKs <- glm(CAD~TKs,data=data.cad2,family=binomial)
> model.leu <- glm(CAD~leu,data=data.cad2,family=binomial)
> model.HDL <- glm(CAD~HDL,data=data.cad2,family=binomial)
```

Porovnáme proto hodnoty odhadů regresních koeficientů u redukovaného modelu a modelů samostatných:

Veličina	Model samostatný	Model redukovaný	Relativní změna
sex	−1,1862	−1,1172	0,0582
beta.blok	2,2562	1,9032	0,1565
statin	3,4336	3,0702	0,1058
vek	0,0332	0,0380	−0,1446
vyska	0,0242	−0,0388	2,6033
EF	−0,0518	−0,0330	0,3629
TKs	0,0171	0,0178	−0,0409
leu	0,2643	0,2383	0,0984
HDL	−2,0393	−1,4359	0,2959

Tabulka 6.9: Odhady koeficientů u samostatných modelů a u modelu redukovaného a jejich relativní změna.

Poslední sloupec určuje relativní změnu koeficientu mezi jeho hodnotou v samostatném a v redukovaném modelu. Největší z nich je právě u veličiny vyska, která je zároveň o jeden řád vyšší než ostatní. To indikuje skutečnost, že proměnlivost výšky lze vysvětlit pomocí zbývajících proměnných v modelu. Podrobnější analýzou zjistíme, že 1) mezi spojitými veličinami je velikost korelace nejvyšší právě mezi věkem a výškou ($R = -0,2932$, $p < 0,001$) a potom mezi HDL a výškou ($R = -0.2378$, $p < 0,001$), 2) mezi muži a ženami existuje významný rozdíl ve výšce, přičemž průměrná výška mužů je 175,25 cm a průměrná výška žen 161,99 cm. Proto ji z redukovaného modelu vyloučíme:

```
> model.cad.step2 <- update(model.cad.step,~.-vyska)
```

Tím získáme výsledný model

```
> summary(model.cad.step2)

Call:
glm(formula = CAD ~ sex + beta.blok + statin + vek + EF + TKs +
    leu + HDL, family = binomial, data = data.cad)

Deviance Residuals:
```

```
    Min       1Q   Median       3Q      Max
-3.0915   0.1432   0.2960   0.4548   2.5121

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.733009   1.455969  -3.938 8.23e-05 ***
sexžena     -0.617212   0.245788  -2.511 0.012034 *
beta.blok1   1.904112   0.280362   6.792 1.11e-11 ***
statin1      3.039835   0.296005  10.270  < 2e-16 ***
vek          0.043773   0.011602   3.773 0.000161 ***
EF          -0.034273   0.010106  -3.391 0.000696 ***
TKs          0.017235   0.007065   2.439 0.014713 *
leu          0.240668   0.062552   3.847 0.000119 ***
HDL         -1.428044   0.346493  -4.121 3.77e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 960.31  on 940  degrees of freedom
Residual deviance: 558.98  on 932  degrees of freedom
AIC: 576.98

Number of Fisher Scoring iterations: 6
```

Ten porovnáme s předchozím redukovaným modelem

```
> lrtest(model.cad,model.cad.step2)
Likelihood ratio test

Model 1: CAD ~ sex + beta.blok + statin + vek + vyska + EF + TKs +
    leu + HDL
Model 2: CAD ~ sex + beta.blok + statin + vek + EF + TKs + leu + HDL
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  10 -277.07
2   9 -279.49 -1 4.8406     0.0278 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Srovnáním s původním modelem dostaneme

```
> lrtest(model.cad,model.cad.step2)
Likelihood ratio test

Model 1: CAD ~ sex + koureni + hypertenze + diabetes + renal.ins +
    ACEi + beta.blok + statin + MMP13 + vek + vyska + BMI + EF + TF +
    TKs + leu + trombo + fibrinogen + chol + HDL + glykemie +
```

```
    ln.CRP
Model 2: CAD ~ sex + beta.blok + statin + vek + EF + TKs + leu + HDL
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  24 -274.36
2   9 -279.49 -15 10.264     0.8028
```

Vidíme, že výsledný model bez veličiny vyska se významně odlišuje od modelu redukovaného, tzn. že došlo k signifikantní změně logaritmické věrohodnostní funkce u výsledného modelu, v porovnáním s modelem původním ovšem tato změna významná není.

6.3.1 Diagnostika výsledného modelu

Pro účely diagnostiky výsledného logistického regresního modelu byly vytvořeny funkce v R, které jsou uvedené v části Přílohy a které nyní využijeme. Pro srovnání uvedeme tytéž charakteristiky pro původní a redukovaný model.

Kendallovo τ_a, τ_b, Somersovo D_{asym}, Goodmanovo-Kruskalovo γ

```
> logreg.coeffs(model.cad)
    tau a     tau b    D asym     gamma
0.2593619 0.4522345 0.7885337 0.7885337
> logreg.coeffs(model.cad.step)
    tau a     tau b    D asym     gamma
0.2584032 0.4505629 0.7856190 0.7856190
> logreg.coeffs(model.cad.step2)
    tau a     tau b    D asym     gamma
0.2582811 0.4503500 0.7852478 0.7852478
```

Koeficienty determinace

```
> logreg.r2(model.cad)
    McFadden Cox-Snell Nagelkerke
R2 0.4286019 0.3542841  0.5539224
> logreg.r2(model.cad.step)
    McFadden Cox-Snell Nagelkerke
R2 0.4229542 0.3505516  0.5480868
> logreg.r2(model.cad.step2)
    McFadden Cox-Snell Nagelkerke
R2 0.4179135 0.3472022    0.54285
```

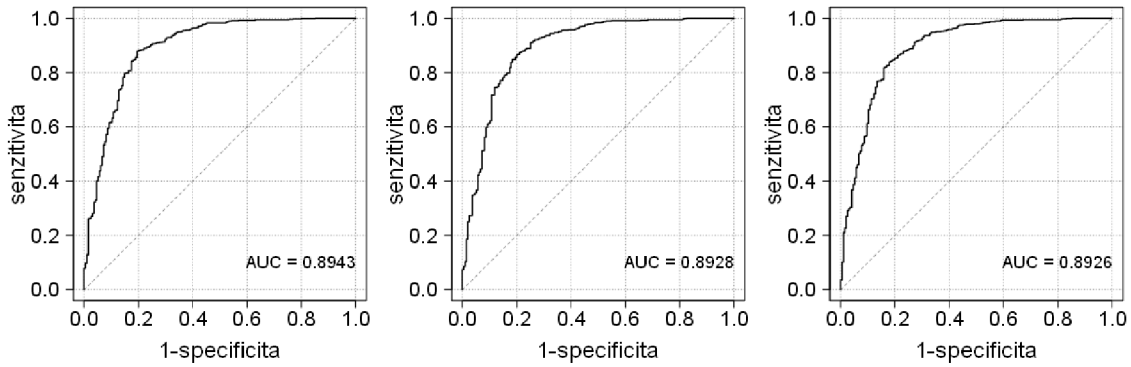
Testy dobré shody

```
> logreg.gof(model.cad,g.cg=8,g.hg=8)
         Chi-square  df          p
Cg         4.487746   6 0.61097443
Hg         1.780800   6 0.93871449
```

```
Pearson 1004.784182 917 0.02257321
D        548.717309 917 1.00000000
> logreg.gof(model.cad.step,g.cg=8,g.hg=8)
         Chi-square  df          p
Cg         5.852337   6 0.43993315
Hg         7.042732   6 0.31691349
Pearson 1012.527541 930 0.03039176
D        554.140913 930 1.00000000
> logreg.gof(model.cad.step2,g.cg=8,g.hg=8)
         Chi-square  df          p
Cg         5.436183   6 0.48920321
Hg         4.024934   6 0.67330199
Pearson 1010.457825 931 0.03539557
D        558.981476 931 1.00000000
```

ROC křivky

Byly vykresleny s použitím funkce `roc.plot`.



Obrázek 6.1: ROC křivky pro modely (směrem zleva doprava) a) původní, b) redukovaný a c) výsledný.

Výstupem této funkce jsou i hodnota *AUC*, prahového bodu a odhad dosažené specificity modelu při zvolené senzitivitě. Ukažme tyto hodnoty pro výsledný model při zvolené senzitivitě $0,95$:

```
> roc.plot(model.cad.step2,sens.cut=0.95)
          AUC   cut.off      sens      spec
fit 0.8926239 0.6058058 0.9497319 0.6461538
```

Hypotézu o „dobrém“ proložení dat modelem vzhledem k p-hodnotám pro oba Hosmerovy-Lemeshowovy koeficienty nezamítáme. Použití Pearsonova χ^2 testu a rozdílového testu deviancí vzhledem k velkému rozdílu mezi realizacemi jejich testových statistik χ^2 a D, kdy hodnota χ^2 je téměř dvakrát větší než hodnota D, se nejeví jako vyhovující.

Porovnáním Kendallova τ_a, Kendallova τ_b, Somersova D_{asym}, Goodmanova-Kruskalova γ koeficientu a též koeficietů McFaddenova, Coxova-Snellova a Nagelkerkova lze soudit, že kvalita proložení dat všemi třemi modely je téměř stejná. Stejný závěr lze učinit i

porovnáním ROC křivek a hodnot ploch pod křivkou AUC, které se liší řádově až na třetím desetinném místě.

Ačkoliv byl zjištěn významný rozdíl mezi redukovaným a výsledným modelem, z hlediska praktického se kvalita proložení dat oběma modely příliš neliší.

6.3.2 Výsledný model logistické regrese pro aterosklerózu

Výsledný model lze zapsat ve tvaru

$$\ln\left(\frac{P(CAD=1)}{1-P(CAD=1)}\right) = -5,7330 - 0,6172.I(sex) + 1,9041.I(beta.blok) + \\ + 3,0398.I(statin) + 0,0438.vek - 0,0343.EF + \\ + 0,0172.TKs + 0,2407.leu - 1,4280.HDL$$

přičemž $I(sex) = \begin{cases} 1 & , sex = \text{žena} \\ 0 & , sex = \text{muž} \end{cases}$, $I(beta.blok) = \begin{cases} 1 & , beta.blok = 1 \\ 0 & , beta.blok = 0 \end{cases}$ a $I(statin) = \\ = \begin{cases} 1 & , statin = 1 \\ 0 & , statin = 0 \end{cases}$.

Závěr

Byl vytvořen logistický regresní model pro predikci pravděpodobnosti diagnostikování aterosklerózy u pacientů, u nichž je podezření, že toto onemocnění mají. Jako významné faktory se ukazují být pohlaví, užívání či neužívání beta blokátorů a užívání či neužívání statinů. Mezi významné prediktory patří věk, ejekční frakce, systolický krevní tlak, koncentrace leukocytů a hodnota HDL. Ženy mají v porovnání s muži (při fixovaných stejných hodnotách ostatních veličin) nižší pravděpodobnost diagnostikování aterosklerózy. Dále pacienti užívající beta blokátory a statiny mají tuto pravděpodobnost vyšší. S rostoucím věkem, s vyšší hladinou krevního tlaku a vyšší koncentrací leukocytů podle vytvořeného modelu roste u pacientů pravděpodobnost diagnózy aterosklerózy, naopak vyšší hladina ejekční frakce a HDL tuto pravděpodobnost snižují. Tyto závěry jsou v souladu se známými skutečnostmi ohledně kvalitativního vztahu mezi hodnotami uvedených veličin u pacientů s CAD a bez CAD.

Pro účely klasifikace pacientů s podezřením na aterosklerózu byla pro zvolenou hladinu senzitivity $0,95$ z dat odhadnuta specificita na $0,6462$, přičemž hladina prahového bodu činí $0,6058$. To znamená, že pokud hodnota skóre přiřazená modelem určitému pacientovi bude větší než $0,6058$, tj. bude zařazen do skupiny pacientů s aterosklerózou, bude mít ve skutečnosti aterosklerózu s přibližně 95% pravděpodobností. Na druhou stranu, pokud hodnota skóre dosáhne hodnoty pod $0,6058$, tj. bude zařazen do skupiny pacientů, u nichž bylo podezření na aterosklerózu vyloučeno, tak s pravděpodobností přibližně $64,6\%$ se u něj toto onemocnění ve skutečnosti nevyskytuje.

Nebyla zjištěna významná asociace mezi přítomností či nepřítomností aterosklerózy a mezi polymorfismem na lokusu rs640198 genu kódujícího MMP13.

Přílohy

Zde jsou uvedeny vytvořené funkce v programu R použité k diagnostice logistického regresního modelu.

Kendallovo τ_a, τ_b, Somersovo D_{asym}, Goodmanovo-Kruskalovo γ

```
logreg.coeffs <- function(model,no=F) {
  y <- as.numeric(model$model[,1])-1 # realizace zavisle promenne
  s <- predict(model,type="response") # hodnota prirazenych skore
  if (length(y)!=length(s)) stop("y a s musi mit stejny rozsah.")
  n <- length(y)
  Sk <- Sd <- Sv1 <- Sv2 <- Sv12 <- 0
  for (i in 1:(n-1)) {
    for (j in (i+1):n) {
      y.ij <- sign(y[i]-y[j])
      s.ij <- sign(s[i]-s[j])
      if (y.ij*s.ij==1) {Sk <- Sk+1
        } else if (y.ij*s.ij==-1) {Sd <- Sd+1
          } else if (y.ij==0&s.ij!=0) {Sv1 <- Sv1+1
            } else if (y.ij!=0&s.ij==0) {Sv2 <- Sv2+1
              } else {Sv12 <- Sv12+1}
    }
  }
  sv <- c(Sk,Sd,Sv1,Sv2,Sv12,n*(n-1)/2)
  names(sv) <- c("concord","discord","ties y","ties s","ties y&s",
                 "no. of all pairs")
  tau.a <- (Sk-Sd)/(n*(n-1)/2) # Kendall tau a
  tau.b <- (Sk-Sd)/sqrt((Sk+Sd+Sv1)*(Sk+Sd+Sv2)) # Kendall tau b
  somers.D.asym <- (Sk-Sd)/(Sk+Sd+Sv2) # Somers' D asymetric
  gk.gamma <- (Sk-Sd)/(Sk+Sd) # Goodman-Kruskal gamma
  coeffs <- c(tau.a,tau.b,somers.D.asym,gk.gamma)
  names(coeffs) <- c("tau a","tau b","D asym","gamma")
  if (no) {return(list("number of pairs"=sv,"coefficients"=coefs))
    } else return(coeffs)
}
```

Koeficienty determinace

```
logreg.r2 <- function(model) {
  D1 <- deviance(model)
  D0 <- model$null.deviance
  n <- length(residuals(model))
  mcfadden <- 1-D1/D0
  cox.snell <- 1-exp((D1-D0)/n)
  nagel <- cox.snell/(1-exp(-D0/n))
  res <- matrix(c(mcfadden,cox.snell,nagel),1,3,dimnames=list(
                c("R2"),c("McFadden","Cox-Snell","Nagelkerke")))
  return(res)
}
```

Testy dobré shody

```
logreg.gof <- function(model,g.cg=10,g.hg=10,obs.exp=F) {
  y <- as.numeric(model$model[,1])-1
  yhat <- predict(model,type="response")
  cutyhat.cg <- cut(yhat,breaks=quantile(yhat,probs=seq(0,1,1/g.cg)),
                    include.lowest=T)
  cutyhat.hg <- cut(yhat,breaks=seq(0,1,1/g.hg),include.lowest=T)
  n <- length(y)
  m <- ncol(model$model)
  obs.cg <- xtabs(cbind(1-y,y)~cutyhat.cg) # Cg
  expect.cg <- xtabs(cbind(1-yhat,yhat)~cutyhat.cg)
  chisq.cg <- sum((obs.cg-expect.cg)^2/expect.cg)
  p.cg <- 1-pchisq(chisq.cg,g.cg-2)
  obs.hg <- xtabs(cbind(1-y,y)~cutyhat.hg) # Hg
  expect.hg <- xtabs(cbind(1-yhat,yhat)~cutyhat.hg)
  chisq.hg <- sum((obs.hg-expect.hg)^2/expect.hg)
  p.hg <- 1-pchisq(chisq.hg,g.hg-2)
  chisq.pear <- sum((y-yhat)^2/(yhat*(1-yhat))) # Pearson Chi^2
  p.pear <- 1-pchisq(chisq.pear,n-m-1)
  D <- -2*sum(yhat*log(yhat/(1-yhat))+log(1-yhat))  # D
  p.D <- 1-pchisq(D,n-m-1)
  mat <- matrix(c(chisq.cg,chisq.hg,chisq.pear,D,g.cg-2,g.hg-2,n-m-1,
                  n-m-1,p.cg,p.hg,p.pear,p.D),4,3,dimnames=list(c(
                  "Cg","Hg","Pearson","D"),c("Chi-square","df","p")))
  res <- list("Observed - Cg"=obs.cg,"Expected - Cg"=expect.cg,
              "Observed - Hg"=obs.hg,"Expected - Hg"=expect.hg,
              "Results"=mat)
  if(obs.exp) {return(res)
    } else return(mat)
}
```

ROC křivka

```
roc.plot <- function(model,case.level="1",figure=T,num.out=T,sens.
             cut=0.95,main="ROC křivka",mar=c(4,4,2.5,1),sub.AUC=T,
             text.AUC=F,cex.main=1,cex.axis=1,cex.lab=1,cex.text=0.7,
             pos.x=0.8,pos.y=0.1,mgp=c(3,1,0),grid.col="lightgray") {
  y <- factor(model$model[,1]) # zavisla promenna
  y.levels <- levels(y)
  if (y.levels[1]==case.level) y.levels <- rev(y.levels)
  data.fit <- data.frame(y=y,fit=model$fitted.values)
  data.fit <- data.fit[order(data.fit$fit),]
  Fy0 <- cumsum(data.fit$y==y.levels[1])/sum(data.fit$y==y.levels[1])
  Fy1 <- cumsum(data.fit$y==y.levels[2])/sum(data.fit$y==y.levels[2])
  data.fit$sens <- 1-Fy1
  data.fit$spec <- Fy0
  d.sens <- c(1,data.fit$sens,0)
  d.fpr <- c(1,1-data.fit$spec,0)
  yy <- rev(d.sens)
  xx <- rev(d.fpr)
  AUC <- sum(diff(xx)*yy[2:length(yy)])
  if (figure) {
    par(mar=mar,las=1,mgp=c(2.5,1,0),cex.main=cex.main,cex.axis=
        cex.axis,cex.lab=cex.lab,mgp=mgp)
    plot(1,xlim=c(0,1),ylim=c(0,1),type="n",xlab="",ylab="")
    grid(col=grid.col)
    lines(c(0,1),c(0,1),lty=2,col="grey50")
    par(new=T)
    plot(xx,yy,type="s",lwd=2,xlab="1-specificita",ylab="senzitivita"
         ,main=main)
    if (sub.AUC) {mtext(paste("AUC =",round(AUC,4)),cex=0.7)}
    if (text.AUC) text(pos.x,pos.y,paste("AUC =",round(AUC,4)),
                       cex=cex.text)
  }
  dd <- data.fit[data.fit$sens>=sens.cut,]
  dd[nrow(dd)+1,] <- data.fit[nrow(dd)+1,]
  if (dd[nrow(dd)-1,3]==sens.cut) {a <- dd[nrow(dd)-1,]
  } else {a <- c(1,apply(as.matrix(dd[c(nrow(dd)-1,nrow(dd)),2:4]),2,
                 mean))}
  num <- data.frame(AUC,cut.off=a[2],sens=a[3],spec=a[4])
  if (num.out) return(num)
}
```

Seznam použité literatury

[1] J. Anděl: *Základy matematické statistiky*. MATFYZPRESS, Praha 2007

[2] K. Zvára: *Regrese*. MATFYZPRESS, Praha 2008

[3] M. Meloun, J. Militký: *Statistická analýza experimentálních dat*. Academia, Praha 2004

[4] A. Agresti: *Categorical Data Analysis*. John Wiley & Sons, New Jersey 2002

[5] S. L. P. Ferrari, F. Cribari-Neto (2004). *Beta Regression for Modelling Rates and Proportions*. Journal of Applied Statistics, Vol. 31, No. 7, 799–815

[6] E. L. Lehmann, G. Casella: *Theory of Point Estimation*. Springer, New York 1998

[7] S. Pekár, M. Brabec: *Moderní analýza biologických dat*, Scientia, Praha 2009

[8] J. I. Myung, D. J. Navarro (2005). *Information Matrix*. Encyclopedia of Behaviaoral Statistics, Vol. 2, 923–924

[9] M. Rashid, N. Shaifa (2009). *Consistency of the Maximum Likelihood Estimator in Logistic Regression Model: A Different Approach*. Journal of Statistics, Vol. 16, 1–11

[10] S. A. Czepiel. *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation*, dostupné na http://czep.net/stat/mlelr/

[11] T. P. Minka: *Algorithms for maximum-likelihood logistic regression*, dostupné na https://research.microsoft.com/en-us/um/people/minka/papers/logreg/

[12] I. Horová, J. Zelinka: *Numerické metody*. Masarykova univerzita, Brno 2008

[13] A. Göktaş, O. İşçi (2011). *A Comparison of the Most Commonly Used Measures of Association for Doubly Ordered Square Contingency Tables via Simulation*. Metodološki zvezki, Vol. 8, No. 1, 17–37

[14] D. C. Hallett: *Goodness of Fits Test in Logistic Regression*, dizertační práce, 1999, dostupná na http://tspace.library.utoronto.ca

[15] A. Vašků *et al* (2012). *Matrix metalloproteinase 13 genotype in rs640198 polymorphism is associated with severe coronary artery disease*. Dis Markers, Vol. 33, No. 1, 43–49