Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

Kristýna Bílková

Logistická regrese s aplikacemi ve finančním sektoru

Katedra Pravděpodobnosti a Matematické Statistiky

Vedoucí bakalářské práce:	RNDr. Martin Branda, Ph.D.
Studijní program:	Matematika
Studijní obor:	Finanční matematika

Praha 2012

Na tomto místě bych chtěla poděkovat svému vedoucímu práce RNDr. Martinu Brandovi, Ph.D. za cenné rady, připomínky a ochotu při psaní této bakalářské práce. Dále bych ráda poděkovala panu RNDr. Ing. Jaroslavu Richterovi za poskytnutí a instalaci software. Velké díky patří také mé rodině za podporu při studiu.

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne ..... Kristýna Bílková

Název práce: Logistická regrese s aplikacemi ve finančním sektoru

Autor: Kristýna Bílková

Katedra: Katedra Pravděpodobnosti a Matematické Statistiky

Vedoucí bakalářské práce: RNDr. Martin Branda, Ph.D., Katedra Pravděpodobnosti a Matematické Statistiky

Abstrakt: V práci je popsán model binární logistické regrese. Jeho parametry jsou odhadnuty metodou maximální věrohodnosti. Pro numerické vyčíslení těchto odhadů je použit Newtonův-Raphsonův algoritmus. Pro měření statistické významnosti parametrů modelu jsou definovány některé statistiky. Dále je popsána konstrukce modelu iterační metodou. Pro posouzení kvality modelu jsou definovány testy dobré shody Perasonův Chí Kvadrát test a Hosmerův-Lemeshowův test. Diverzifikační schopnost modelu je ilustrována pomocí Lorenzovy křivky a kvantifikována Giniho koeficientem, Kolmogorovovou-Smirnovovou statistikou a zobecněným koeficientem determinace. Teoretické poznatky jsou aplikovány na data z oblasti pojišťovnictví.

Klíčová slova: binární logistická regrese, konstrukce modelu iterační metodou, testy dobré shody, diverzifikační schopnost modelu

Title: Logistic regression with applications in financial sector

Author: Kristýna Bílková

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Martin Branda, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this bachelor thesis binary logistic regression model is described. Its parameters are estimated by maximum likelihood method. Newton-Raphson's algorithm is used for enumeration of these estimates. There are defined some statistics for testing the significance of the coefficients. Then stepwise regression is described. For assessing the quality of the model Pearson's Chi Square Test and Hosmer-Lemeshow's Test of the goodness of fit are defined. Diversification ability of the model is illustrated by the Lorenz curve and is quantificated by Gini coefficient, Kolmogorov-Smirnov statistics and generalized coefficient of determination. The theoretical knowledge is applied to insurance area data.

Keywords: binary logistic regression, stepwise regression, goodness of fit tests, diversification ability of the model

Obsah

Úvod

Pomocí binární logistické regrese můžeme vytvořit model závislosti binární neboli dichotomické, nula-jedničkové veličiny na jiných nezávislých proměnných. Ty už mohou být jak binární, tak kategoriální či spojité. Na základě takového modelu lze predikovat pravděpodobnost výskytu jevu, který značí závislá veličina, pro nové hodnoty nezávislých proměnných. Tato metoda nalezne praktické využití ve všech oblastech, kde jsou k dispozici historická data týkající se výskytu jevu, jehož pravděpodobnost chceme modelovat, přičemž tato data jsou významná pro učinění nějakého rozhodnutí.

Velké uplatnění logistické regrese najdeme v medicíně a lékařských a farmaceutických studiích, kde modelujeme pravděpodobnost výskytu určitého onemocnění. Od toho se může odvíjet například na jakou skupinu lidí se máme více zaměřit v oblasti prevence, abchom tomuto onemocnění mohli pokud možno předejít.

V současnosti se logistická regrese hojně používá v oblasti řízení kreditního rizika. Modeluje se pravděpodobnost, že klient, jemuž společnost poskytla úvěr, tento úvěr nesplatí, na základě informací, které o klientovi máme k dispozici.

Další oblastí je pojišťovnictví, kde se modeluje například pravděpodobnost vzniku pojistné události nebo pravděpodobnost storna pojistné smlouvy. Díky tomu můžeme efektivněji zařadit klienta pojišťovny do bonusové třídy v případě havarijního pojištění nebo nabídnout klientovi, u nějž předpovíme vysokou pravděpodobnost storna, výhodnější podmínky pojištění.

Hlavním cílem této práce je srozumitelně popsat binární logistickou regresi, uvést ukazatele pro posouzení kvality modelu a teorii ilustrovat na numerické studii na datech z pojišťovnictví. Pro lepší názornost některých teoretických vztahů se budeme snažit představovat si je na příkladech ze života.

Organizace práce je následující. V první kapitole zavádíme model logistické regrese. Pro názornost se budeme v kapitole 1.1 nejprve zabývat jednorozměrným modelem a následně v kapitole 1.2 přejdeme k obecnějšímu vícerozměrnému. Poté, co si popíšeme základní vztahy, se zaměříme na odhad parametrů metodou maximální věrohodnosti včetně hledání numerického odhadu řešení pomocí Newtonova-Raphsonova algoritmu. Není pravidlem, že čím více parametrů do modelu zahrneme, tím lépe. Proto se v kapitole 1.3 podíváme na to, jak otestovat statistickou významnost parametrů. Jiný způsob, jak vhodně vybrat proměnné, které do modelu zahrneme, je krokový výběr (anglicky „stepwise regression“), který si ukážeme v kapitole 1.5.

Obsahem druhé kapitoly jsou metody vhodné k posouzení kvality již vybudovaného modelu. V kapitole 2.1 si ukážeme testy dobré shody, a to Pearsonův Chí Kvadrát test a Hosmerův-Lemeshowův test. Také nás bude zajímat diverzifikační schopnost modelu, které se věnuje kapitola 2.2. Tuto schopnost budeme ilustrovat pomocí Lorenzovy křivky a kvantifikovat pomocí Giniho koeficientu, Komogorovovy-Smirnovovy statistiky a zobecněného koeficientu determinace.

Náplní třetí kapitoly je aplikace na data z pojišťovnictví, která si popíšeme v kapitole 3.1. Provedeme numerickou studii, na níž ilustrujeme teorii vyloženou v předchozích kapitolách. V kapitole 3.2 tedy vybudujeme vhodný model logistické regrese a budeme se snažit posoudit jeho kvalitu. Výsledky v kapitole 3.3 shrneme.

K práci je přiloženo i CD, kde můžeme najít data použitá ve třetí kapitole,

projekt, v němž realizujeme numerickou studii v programu SAS, výstupy z tohoto programu a pdf verzi této bakalářské práce.

1. Model logistické regrese

V této kapitole si představíme model logistické regrese. Kromě odhadu jeho parametrů nás bude také zajímat výběr proměnných, které do modelu zahrneme. Vycházet budeme především z [3].

1.1 Jednorozměrný model

Nejdříve se budeme zabývat modelem s jedním regresorem. Můžeme si jej představit například jako modelování pravděpodobnosti rizika nesplacení dluhu v závislosti na věku klienta.

1.1.1 Základní popis

Nechť Y je dichotomická náhodná veličina, tj. nabývá hodnoty buď 0 (hovoříme o „neúspěchu“ nějakého jevu J) nebo 1 („úspěch“ jevu J). Hodnota této veličiny může být ovlivněna pozorovanou hodnotou x, a to způsobem, který popisuje právě logistická regrese:

$$\begin{aligned} E(Y \mid x) &= \pi(x) \\ \pi(x) &= \frac{e^{g(x)}}{1+e^{g(x)}} \\ g(x) &= \beta_0 + \beta_1 x. \end{aligned}$$

Y budeme nazývat odezva, x regresor a $g(x)$ logit.
$\pi(x)$ je zároveň pravděpodobnost „úspěchu“, $1-\pi(x)$ je tedy pravděpodobnost „neúspěchu“.

$$E(Y \mid x) = 0 \cdot P[Y=0 \mid x] + 1 \cdot P[Y=1 \mid x] = P[Y=1 \mid x] = \pi(x)$$

Veličina Y má alternativní rozdělení s parametrem $\pi(x)$.
Logit $g(x)$ je logaritmus šance, tedy

$$\begin{aligned} \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) &= \ln\left(\frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{1-\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}\right) \\ &= \ln\left(\frac{e^{\beta_0+\beta_1 x}\cdot\left(1+e^{\beta_0+\beta_1 x}\right)}{\left(1+e^{\beta_0+\beta_1 x}\right)\cdot\left(1+e^{\beta_0+\beta_1 x}-e^{\beta_0+\beta_1 x}\right)}\right) \\ &= \beta_0+\beta_1 x. \end{aligned}$$

1.1.2 Odhad parametrů

Mějme n nezávislých pozorování $(y_i, x_i), i = 1, \ldots, n$, kde y_i je hodnota náhodné veličiny Y v i-tém pozorování a x_i je hodnota regresoru pro i-tý pozorovaný subjekt (např. věk i-tého klienta). Pro odhad vektoru parametrů $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ logistické regrese použijeme metodu maximální věrohodnosti. Pro obecné poznatky

o této metodě můžeme nahlédnout do [1]. Tato metoda je založena na principu, že za odhad neznámého parametru vezmeme tu jeho hodnotu, pro kterou je dosažený výsledek nejpravděpodobnější. Tuto pravděpodobnost udává věrohodnostní funkce, budeme tedy hledat takový parametr, v němž nabývá svého maxima.

Všimněme si, že

$$P[Y_i = y_i \mid x_i] = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{(1-y_i)},$$

kde $y_i \in \{0, 1\} \forall i = 1, \ldots, n$, tedy pro $y_i = 1$ je $P[Y_i = y_i \mid x_i] = \pi(x_i)$ a pro $y_i = 0$ je $P[Y_i = y_i \mid x_i] = 1 - \pi(x_i)$. Protože pozorování (y_i, x_i), $i = 1, \ldots, n$ jsou nezávislá, můžeme věrohodnostní funkci $l(\boldsymbol{\beta})$ zapsat jako součin podmíněných pravděpodobností pro jednotlivá pozorování

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

Chceme najít takové $\boldsymbol{\beta}$, v němž $l(\boldsymbol{\beta})$ nabývá maxima. Protože logaritmus je rostoucí funkce, má $L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})]$ maximum ve stejném bodě jako $l(\boldsymbol{\beta})$.

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

Funkce $L(\boldsymbol{\beta})$ je funkce 2 proměnných na $\mathbb{R}^2$, její maximum najdeme pomocí parciálních derivací podle β_0 a β_1, které položíme rovné 0 (že se skutečně jedná o maximum rozebereme až v části týkající se vícerozměrného modelu). Tím získáme soustavu normálních rovnic, jejichž řešením je právě odhad vektoru $\boldsymbol{\beta}$, který budeme značit $\hat{\boldsymbol{\beta}}$.

Nejprve si uvědomme, že

$$\begin{aligned}
1 - \pi(x_i) &= \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} \\
\frac{\partial \pi(x_i)}{\partial \beta_0} &= \frac{e^{\beta_0 + \beta_1 x_i}\left(1 + e^{\beta_0 + \beta_1 x_i}\right) - e^{\beta_0 + \beta_1 x_i} e^{\beta_0 + 2\beta_1 x_i}}{\left(1 + e^{\beta_0 + \beta_1 x_i}\right)^2} \\
&= \frac{e^{\beta_0 + \beta_1 x_i}}{\left(1 + e^{\beta_0 + \beta_1 x_i}\right)^2} = \pi(x_i)\left(1 - \pi(x_i)\right) \\
\frac{\partial \pi(x_i)}{\partial \beta_1} &= \frac{x_i e^{\beta_0 + \beta_1 x_i}\left(1 + e^{\beta_0 + \beta_1 x_i}\right) - e^{\beta_0 + \beta_1 x_i} x_i e^{\beta_0 + \beta_1 x_i}}{\left(1 + e^{\beta_0 + \beta_1 x_i}\right)^2} \\
&= \frac{x_i e^{\beta_0 + \beta_1 x_i}}{\left(1 + e^{\beta_0 + \beta_1 x_i}\right)^2} = x_i \pi(x_i)\left(1 - \pi(x_i)\right).
\end{aligned}$$

První parciální derivace logaritmické věrohodnostní funkce mají tvar

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

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

Dostáváme soustavu normálních rovnic

$$\sum_{i=1}^{n}\left[y_i-\pi(x_i)\right] = 0 \tag{1.1}$$

$$\sum_{i=1}^{n}\left[x_i\left(y_i-\pi(x_i)\right)\right] = 0. \tag{1.2}$$

Vidíme, že tyto rovnice nejsou lineární v β_0 ani v β_1, což vyžaduje speciální metody pro jejich řešení, většina z nich je implementována v různých statistických software (např. SAS). Jednou z možností je Newtonův-Rapshonův algoritmus, který je popsán například v [4]. Jedná se o iterační alogritmus pro řešení nelineárních rovnic. Postupně počítáme $\boldsymbol{\beta}^{(t)} \in \mathbb{R}^{p+1}$: $\boldsymbol{\beta}^{(0)}$ určíme libovolně. Dále postupujeme podle vzorce

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(H^{(t)}\right)^{-1}\boldsymbol{q}^{(t)},$$

kde $\boldsymbol{q}^{(t)} = \nabla L(\boldsymbol{\beta}^{(t)})$ je gradient logaritmické věrohodnostní funkce s parametry $\boldsymbol{\beta}^{(t)}$ a $H^{(t)} = \left(\dfrac{\partial^2 L(\boldsymbol{\beta}^{(t)})}{\partial\beta_i\partial\beta_j}\right)_{i,j=0}^{p}$ matice druhých parciálních derivací této funkce. Algoritmus pokračuje až do doby, kdy se $\boldsymbol{\beta}^{(t+1)}$ liší od $\boldsymbol{\beta}^{(t)}$ maximálně o předem stanovenou mez.

Pro model s jedním regresorem mají $\boldsymbol{q}^{(t)}$ a $H^{(t)}$ tvar

$$\begin{aligned}\boldsymbol{q}^{(t)} &= \begin{pmatrix}\sum_{i=1}^{n}\left[y_i-\pi^{(t)}(x_i)\right]\\ \sum_{i=1}^{n}\left[x_i\left(y_i-\pi^{(t)}(x_i)\right)\right]\end{pmatrix}\\ H^{(t)} &= -\begin{pmatrix}\sum_{i=1}^{n}\pi^{(t)}(x_i)\left(1-\pi^{(t)}(x_i)\right) & \sum_{i=1}^{n}\left[x_i\pi^{(t)}(x_i)\left(1-\pi^{(t)}(x_i)\right)\right]\\ \sum_{i=1}^{n}\left[x_i\pi^{(t)}(x_i)\left(1-\pi^{(t)}(x_i)\right)\right] & \sum_{i=1}^{n}\left[x_i^2\pi^{(t)}(x_i)\left(1-\pi^{(t)}(x_i)\right)\right]\end{pmatrix},\end{aligned}$$

kde $\pi^{(t)}(x_i) = \dfrac{e^{\beta_0^{(t)}+\beta_1^{(t)}x_i}}{1+e^{\beta_0^{(t)}+\beta_1^{(t)}x_i}}$ a $\beta_0^{(t)}$ a $\beta_1^{(t)}$ jsou složky vektoru parametrů $\boldsymbol{\beta}^{(t)}$ odhadnutého v kroku t.Všimněme si, že matice $H^{(t)}$ nezávisí na pozorování y_i, $i = 1,\ldots,n$. Z rovnice (1.1) plyne, že součet pozorovaných hodnot y_i se rovná součtu přepovězených hodnot

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\pi(x_i).$$

1.2 Vícerozměrný model

Od teď budeme uvažovat, že hodnota náhodné veličiny Y může být ovlivněna více regresory. Například riziko, že klient nesplatí úvěr, bude souviset nejen s věkem klienta, ale i s jeho příjmem, zaměstnáním, počtem dětí a podobně.

1.2.1 Základní popis

Rozšíříme výše popsaný model o větší počet regresorů:

$$\begin{aligned} \mathbf{x}^\top &= (x_1, \ldots, x_p) \\ E(Y \mid \mathbf{x}) &= \pi(\mathbf{x}) \\ \pi(\mathbf{x}) &= \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} \\ g(\mathbf{x}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \end{aligned}$$

parametry $\beta_j,\ j = 0, \ldots, p$ jsou složky vektoru parametrů $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$.

1.2.2 Odhad parametrů

Mějme n nezávislých pozorování tvaru $(y_i, \mathbf{x_i})$, $i = 1, \ldots, n$. Tvar modelu logistické regrese je

$$\pi(\mathbf{x_i}) = \frac{e^{g(\mathbf{x_i})}}{1 + e^{g(\mathbf{x_i})}}\,, \tag{1.3}$$

kde

$$g(\mathbf{x_i}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\,, p \leq n$$

Stejně jako v případě jednorozměrného modelu použijeme k odhadu parametrů $\boldsymbol{\beta}$ metodu maximální věrohodnosti. Opět platí, že

$$P[Y_i = yi \mid \mathbf{x_i}] = \pi(\mathbf{x_i})^{y_i}[1 - \pi(\mathbf{x_i})]^{(1-y_i)},$$

kde $y_i \in \{0, 1\}\ \forall i = 1, \ldots, n$.

Vzhledem k tomu, že pozorování $(y_i, \mathbf{x_i})$, $i = 1, \ldots, n$ jsou nezávislá, můžeme věrohodnostní funkci $l(\boldsymbol{\beta})$ zapsat jako:

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

Pro snazší hledání maxima této věrohodnostní funkce se budeme zabývat hledáním maxima logaritmu této funkce, stejně jako v jednorozměrném případě.

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

První parciální derivace $\dfrac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}, j = 0, \ldots, p$ položíme rovné nule.

$$\begin{aligned} \frac{\partial \pi(\mathbf{x_i})}{\partial \beta_0} &= \frac{e^{g(\mathbf{x_i})}(1 + e^{g(\mathbf{x_i})}) - e^{2g(\mathbf{x_i})}}{(1 + e^{g(\mathbf{x_i})})^2} \\ &= \frac{e^{g(\mathbf{x_i})}}{1 + e^{g(\mathbf{x_i})}} = \pi(\mathbf{x_i})\,(1 - \pi(\mathbf{x_i})) \end{aligned}$$

$$\begin{aligned}\frac{\partial \pi(\mathbf{x_i})}{\partial \beta_j} &= \frac{e^{g(\mathbf{x_i})}x_{ij}(1+e^{g(\mathbf{x_i})}) - e^{2g(\mathbf{x_i})}x_{i,j}}{\left(1+e^{g(\mathbf{x_i})}\right)^2} \\ &= x_{i,j}\left(1-\pi(\mathbf{x_i})\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} &= \sum_{i=1}^{n}[y_i\frac{1}{\pi(\mathbf{x_i})}\frac{\partial \pi(\mathbf{x_i})}{\partial \beta_0} - (1-y_i)\frac{1}{1-\pi(\mathbf{x_i})}\frac{\partial \pi(\mathbf{x_i})}{\partial \beta_0}] \\ &= \sum_{i=1}^{n}[y_i(1-\pi(\mathbf{x_i})) - (1-y_i)\pi(\mathbf{x_i})] \\ &= \sum_{i=1}^{n}[y_i - \pi(\mathbf{x_i})]\end{aligned}$$

$$\begin{aligned}\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n}[y_i\frac{1}{\pi(\mathbf{x_i})}\frac{\partial \pi(\mathbf{x_i})}{\partial \beta_j} - (1-y_i)\frac{1}{1-\pi(\mathbf{x_i})}\frac{\partial \pi(\mathbf{x_i})}{\partial \beta_j}] \\ &= \sum[y_i x_{ij}(1-\pi(\mathbf{x_i})) - (1-y_i)x_{ij}\pi(\mathbf{x_i})] \\ &= \sum[xij(y_i - \pi(\mathbf{x_i}))], j = 1,\ldots,p\end{aligned}$$

Získali jsme p+1 věrohodnostních rovnic:

$$\begin{aligned}\sum_{i=1}^{n}[y_i - \pi(\mathbf{x_i})] &= 0 \\ \sum x_{ij}[y_i - \pi(\mathbf{x_i})] &= 0,\end{aligned}$$

$j = 1,\ldots,n$. Jejich řešení $\hat{\boldsymbol{\beta}}$ získáme například pomocí Newton-Raphsonova algoritmu, jak již bylo popsáno výše.

1.2.3 Vlastnosti odhadu parametrů metodou maximální věrohodnosti

Odhad $\hat{\boldsymbol{\beta}}$ pomocí metody maximální věrohodnosti je nestranný a konzistentní (viz [1]). Nejprve se zaměřme na rozptyly a kovariance odhadnutých parametrů, a to způsobem, jaký je popsán v [3]. Pro metodu odhadu rozptylů a kovariancí budeme potřebovat druhé parciální derivace logaritmické věrohodnostní funkce

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j^2} &= \sum_{i=1}^{n} -x_{i,j}\frac{\partial \pi(\mathbf{x_i})}{\partial j} \\ &= -\sum_{i=1}^{n} x_{i,j}^2\pi(\mathbf{x_i})(1-\pi(\mathbf{x_i})), j = 0,\ldots,p\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_l} &= \sum_{i=1}^{n} -x_{i,j}\frac{\partial \pi(\mathbf{x_i})}{\partial l} \\ &= -\sum_{i=1}^{n} x_{i,j}x_{i,l}\pi(\mathbf{x_i})(1-\pi(\mathbf{x_i})), j = 0,\ldots,p,\end{aligned}$$

kde $x_{i,0} = 1 \, \forall i = 1, \ldots, n$.

Hodnoty těchto derivací s opačným znaménkem jsou prvky tzv. informační matice $\mathbf{I}(\boldsymbol{\beta})$ o rozměrech $(p+1) \times (p+1)$

$$i(\boldsymbol{\beta})_{jj} = \sum_{i=1}^{n} x_{i,j}^2 \pi(\mathbf{x_i}) \left(1 - \pi(\mathbf{x_i})\right)$$

$$i(\boldsymbol{\beta})_{jl} = \sum_{i=1}^{n} x_{i,j} x_{i,l} \pi(\mathbf{x_i}) \left(1 - \pi(\mathbf{x_i})\right).$$

Informační matici můžeme maticově zapsat jako $\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{V} \mathbf{X}$, kde $\mathbf{X}$ je matice typu $n \times p + 1$ obsahující data pro každý subjekt a $\mathbf{V}$ diagonální matice typu $n \times n$ s prvky $\pi(x_i)\left(1 - \pi(x_i)\right))$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}$$

$$\mathbf{V} = \begin{pmatrix} \pi(x_1)\left(1 - \pi(x_1)\right) & 0 & \cdots & 0 \\ 0 & \pi(x_2)\left(1 - \pi(x_2)\right) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \pi(x_n)\left(1 - \pi(x_n)\right) \end{pmatrix}$$

Vzhledem k tomu, že diagonální prvky matice $\mathbf{V}$ jsou kladné, je tato matice pozitivně definitní, matice $\mathbf{I}(\boldsymbol{\beta})$ tedy bude pozitivně-semidefinitní nebo dokonce pozitivně definitní v případě plné řádkové hodnosti matice $\mathbf{X}$ (viz [6]). Matice druhých parciálních derivací $-\mathbf{I}(\boldsymbol{\beta})$ je tedy negativně-semidefinitní, případně negativně-definitní, což znamená, že logaritmická věrohodnostní funkce $L(\beta)$ je konkávní. Pro konkávní funkci platí, že každé její lokální maximum je zároveň maximem globálním (viz [2]). Bod, který jsme našli pomocí prvních parciálních derivací je tedy určitě maximem funkce $L(\boldsymbol{\beta})$, tudíž i $l(\boldsymbol{\beta})$.

Varianční matici získáme jako inverzi informační matice $Var(\boldsymbol{\beta}) = \mathbf{I}^{-1}(\boldsymbol{\beta})$. Rozptyl j-té složky vektoru $\boldsymbol{\beta}$, $var(\beta_j)$, je j-tý diagonální prvek matice $Var(\boldsymbol{\beta})$, kovariance složek $cov(\beta_j, \beta_l)$ je prvek této matice o souřadnicích $j, l = 0, \ldots, p$. Odhad varianční matice $\widehat{Var}(\hat{\boldsymbol{\beta}})$ získáme dosazením $\hat{\boldsymbol{\beta}}$. Odhad směrodatné odchylky $\hat{\sigma}(\hat{\beta}_j) = \left[\widehat{var}(\hat{\beta}_j)\right]^{1/2}$.

Zdůvodnění, že výše popsaná teorie funguje, najdeme například v [1]: matice $-\mathbf{I}(\hat{\beta})$ se nazývá výběrová Fisherova informační matice a konverguje v pravděpodobnosti k Fisherově informační matici. Pro odhad parametru $\hat{\boldsymbol{\beta}}$ získaný metodou maximální věrohodnosti platí, že $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, [-I(\boldsymbol{\beta})]^{(-1)}\right)$.

1.3 Testování statistické významnosti parametrů

Jednou ze základních otázek, kterou je třeba si položit, je, zda některé proměnné statisticky významně přispívají k vypovídající schopnosti modelu, či nikoliv. Abychom ji zodpověděli, porovnáme skutečné hodnoty pozorování odezvy s hodnotami, které nám vyjdou v modelu bez těchto proměnný a v modelu s nimi.

Toto porovnání je založeno na věrohodnostní funkci, která udává pravděpodobnost modelu. Zavedeme symbol D (z anglického Deviance):

$$D = -2\ln\left[\frac{(\textit{věrohodnost použitého modelu})}{(\textit{věrohodnost saturovaného modelu})}\right].$$

Saturovaný model obsahuje tolik parametrů, že s pravděpodobností 1 vystihuje data, v našem případě má tedy n parametrů. Věrohodnost takového modelu je tedy rovna jedné, takže logaritmická věrohodnost se rovná nule.
Potom

$$D = -2\ln\left[(\textit{věrohodnost použitého modelu})\right]$$

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

Čili

$$D = -2\ln[l(\boldsymbol{\beta})] = -2\sum_{i=1}^{n}[y_i\ln\pi(\mathbf{x_i}) + (1-y_i)\ln(1-\pi(\mathbf{x_i}))].$$

Nyní chceme zjistit, jestli nějakých konkrétních l proměnných je v modelu významných. Budeme testovat hypotézu H_0, jestli se koeficienty u těchto proměnných rovnají nule, proti alternativě H_1, že alespoň některý z nich je nenulový. Zavedeme si proto statistiku G:

$$\begin{aligned} G &= D(\textit{modelu bez } l \textit{ proměnných}) - D(\textit{modelu s proměnnými}) \\ &= -2\ln(\frac{l_b(\boldsymbol{\beta}_b)}{l_s(\boldsymbol{\beta}_s)}). \end{aligned}$$

Výraz $\dfrac{l_b(\boldsymbol{\beta}_b)}{l_s(\boldsymbol{\beta}_s)}$ se nazývá věrohodnostní poměr, $l_b(\boldsymbol{\beta}_b)$ značí věrohodnostní funkci modelu bez l proměnných, $l_s(\boldsymbol{\beta}_s)$ věrohodnostní funkci modelu s těmito proměnnými. Za platnosti hypotézy H_0, že koeficienty u daných l proměnných se rovnají 0, má G asymptotické rozdělení Chí-kvadrát o l stupních volnosti. Pro test na hladině α platí, že H_0 zamítáme ve prospěch H_1, jestliže je G větší než $(1-\alpha)$-kvantil Chí-kvadrát rozdělení o l stupních volnosti:

$$H_0 \textit{ zamítáme} \Leftrightarrow G > \chi^2_l(1-\alpha).$$

Jiný způsob, jak testovat statistickou významnost nějakého parametru, je Waldův test. Využívá asymptotické normality odhadnutých parametrů, tedy že $\dfrac{\hat{\beta}_i - \beta_i}{\widehat{\sigma}(\hat{\beta}_i)} \overset{as}{\sim} N(0,1)$. Hypotéza H_0 je, že $\beta_i = 0$, alternativa H_1 je, že $\beta_i \neq 0$.
Definujeme Waldovu statistiku

$$W = \frac{\hat{\beta}_i}{\hat{\sigma}(\hat{\beta}_i)} \overset{as}{\sim} N(0,1) \textit{ za platnosti } H_0.$$

Waldův test na hladině α: hypotézu H_0 zamítáme ve prospěch H_1, jestliže $|W| > z_{(1-\alpha/2)}$, kde $z_{(1-\alpha/2)}$ je $(1-\alpha/2)$-kvantil normálního rozdělení. Pomocí Waldovy statistiky můžeme také sestrojit $100(1-\alpha)\%$ interval spolehlivosti:

$$P\left[-z_{(1-\alpha/2)} \leq \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}(\hat{\beta}_i)} \leq z_{(1-\alpha/2)}\right] \approx 1-\alpha$$
$$\beta_i \in (\hat{\beta}_i - z_{1-\alpha/2}\hat{\sigma}(\hat{\beta}_i), \hat{\beta}_i + z_{1-\alpha/2}\hat{\sigma}(\hat{\beta}_i)).$$

1.4 Interpretace parametrů

Předpokládejme, že jsme už z dat napočítali hodnoty koeficientů a vybudovali model logistické regrese pro tato data tak, že všechny koeficienty jsou významné statisticky nebo z povahy problematiky, které se data týkají. Nyní se zamyslíme nad tím, co nám hodnoty odhadnutých koeficientů o této problematice říkají.

Pro jednoduchost uvažujme pouze jednorozměrný model. Budeme se ptát jednak na funkci závislosti mezi odezvou a regresorem a jednak na to, jak se změní odezva, změníme-li hodnotu regresoru o jednotku. Jestliže $\beta_1 = 0$, pak $\pi(x_i) = P[Y_i = 1 \mid x_i] = \frac{e^{\beta_0}}{1+e^{\beta_0}}, \forall i = 1, \dots n$. To znamená, že pravděpodobnost vůbec nezávisí na pozorováních x_i a je pro všechna stejná. Šance úspěchu ku neúspěchu je $odds = \frac{\pi(x_i)}{1-\pi(x_i)} = \frac{\frac{e^{\beta_0}}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0}}} = e^{\beta_0}$, logaritmus této šance je vlastně $logit(x_i)$, $\ln(odds) = \beta_0$. Parametr β_0 si můžeme představit jako úrovňový koeficient pro logit. Nyní se zaměříme na význam β_1. Změníme-li hodnotu x_i na x_j, podíl šancí (anglicky odds ratio) bude $OR = \frac{odds(x_i)}{odds(x_j)} = \frac{e^{(\beta_0+\beta_1 x_i)}}{e^{(\beta_0+\beta_1 x_j)}}$ a logaritmus podílu šancí $\ln\left(\frac{odds(x_i)}{odds(x_j)}\right) = \beta_1(x_i - x_j) = logit(x_i) - logit(x_j)$. Parametr β_1 tedy udává, jak moc je pravděpodobnější „úspěch“ v případě subjektu i oproti subjektu j. Nyní se zaměříme na konkrétní aplikace pro různé typy vstupních proměnných.

1.4.1 Binární proměnná

Předpokládejme, že regresor x nabývá pouze hodnoty 0 nebo 1, kde 1 znamená přítomnost nějakého jevu a 0 jeho nepřítomnost. Šance pro $x = 0$ je

$$odds(x=0) = \frac{\pi(0)}{1-\pi(0)} = \frac{\frac{e^{\beta_0}}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0}}} = e^{\beta_0}.$$

Parametr β_0 můžeme vyjádřit jako logaritmus šance, jestliže se $x = 0$. Šance

pro $x = 1$ je

$$odds(x = 1) = \frac{\pi(0)}{1 - \pi(0)} = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}}{\frac{1}{1+e^{\beta_0+\beta_1}}} = e^{\beta_0+\beta_1}.$$

Podíl šancí pro $x = 1$ a $x = 0$ je

$$OR = \frac{odds(x = 1)}{odds(x = 0)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}.$$

Odsud můžeme vyjádřit parametr β_1 jako logaritmus podílu šancí pro $x = 1$ a $x = 0$.

Abychom odhadli parametry β_0 a β_1 a jejich vlastnosti, nemusíme počítat věrohodnostní funkce, ale můžeme využít empirické ohdhady pravděpodobností, tj. relativní četnosti jednotlivých jevů. Označme $N_{i,j}$ četnost výskytu jevu, kdy $y = j$ a současně $x = i$, n_i počet pozorování, kdy se $x = i$, čili $n_0 = N_{0,0} + N_{0,1}$ a $n_1 = N_{1,0} + N_{1,1}$. Odhady šancí pak mají tvar

$$\widehat{odds}(x = 0) = \frac{\frac{N_{0,1}}{n_0}}{\frac{N_{0,0}}{n_0}} = \frac{N_{0,1}}{N_{0,0}}$$

$$\widehat{odds}(x = 1) = \frac{\frac{N_{1,1}}{n_1}}{\frac{N_{1,0}}{n_1}} = \frac{N_{1,1}}{N_{1,0}},$$

odhad podílu šancí:

$$\widehat{OR} = \frac{\frac{N_{1,1}}{N_{1,0}}}{\frac{N_{0,1}}{N_{0,0}}} = \frac{N_{1,1}N_{0,0}}{N_{1,0}N_{0,1}}.$$

Nyní už snadno vyjádříme odhady parametrů β_0 a β_1

$$\hat{\beta}_0 = \ln\left(\widehat{odds}(x = 0)\right) = \ln\left(\frac{N_{0,1}}{N_{0,0}}\right)$$

$$\hat{\beta}_1 = \ln\left(\widehat{OR}\right) = \ln\left(\frac{N_{1,1}N_{0,0}}{N_{1,0}N_{0,1}}\right).$$

Podíl šancí nám vlastně říká, jak je pravděpodobnější, že odezva se bude rovnat jedné v případě, že $x = 1$ oproti $x = 0$. Například Y značí, že klient nesplatí úvěr, x značí, jestli má klient práci nebo ne. Odhad $\widehat{OR} = 2$ znamená, že nesplacení úvěru je v testované populaci dvakrát pravděpodobnější u lidí bez práce než u pracujících.

Pro odhad varianční matice využijeme, že odhad střední hodnoty odezvy za podmínky x je $\hat{\pi}(x) = \frac{N_{x,1}}{n_x}$ a tudíž odhad rozptylu odezvy za podmínky x je

$\hat{\pi}(x)\,(1-\hat{\pi}(x)) = \frac{N_{x,1}}{n_x}\frac{N_{x,0}}{n_x} = \frac{N_{x,1}N_{x,0}}{n_x^2}$. Odhad informační matice je $\hat{I}(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{X}}^\top\hat{\mathbf{V}}\hat{\mathbf{X}}$. Diagonální matice $\hat{\mathbf{V}}$ obsahuje n_1 prvků $\frac{N_{1,0}N_{1,1}}{n_1^2}$ a n_0 prvků $\frac{N_{0,0}N_{0,1}}{n_0^2}$, matice $\hat{\mathbf{X}}$ obsahuje v prvním sloupci n jedniček a v druhém sloupci n_1 jedniček a n_0 nul.

$$\hat{I}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} n_0\frac{N_{0,1}N_{0,0}}{n_0^2} + n_1\frac{N_{1,1}N_{1,0}}{n_1^2} & n_1\frac{N_{1,1}N_{1,0}}{n_1^2} \\ n_1\frac{N_{1,1}N_{1,0}}{n_1^2} & n_1\frac{N_{1,1}N_{1,0}}{n_1^2} \end{pmatrix}$$

Inverzní matici spočteme například pomocí metody využívající adjungovanou matici a determinant. Determinant této matice je $\frac{N_{0,1}N_{0,0}N_{1,0}N_{1,1}}{n_0 n_1}$. Inverzní matice má tvar

$$\begin{aligned} \widehat{Var}(\hat{\beta}) &= \frac{n_0 n_1}{N_{0,1}N_{0,0}N_{1,0}N_{1,1}} \begin{pmatrix} \frac{N_{1,1}N_{1,0}}{n_1} & -\frac{N_{1,1}N_{1,0}}{n_1} \\ -\frac{N_{1,1}N_{1,0}}{n_1} & \frac{N_{0,1}N_{0,0}}{n_0} + \frac{N_{1,1}N_{1,0}}{n_1} \end{pmatrix} \\ &= \begin{pmatrix} \frac{n_0}{N_{0,1}N_{0,0}} & -\frac{n_0}{N_{0,1}N_{0,0}} \\ -\frac{n_0}{N_{0,1}N_{0,0}} & \frac{n_1}{N_{1,0}N_{1,1} + \frac{n_0}{N_{0,1}N_{0,0}}} \end{pmatrix}. \end{aligned}$$

Odhad rozptylu $\hat{\beta}_1$ můžeme vyjádřit jako $\hat{\beta}_1 = \frac{1}{N_{1,0}} + \frac{1}{N_{1,1}} + \frac{1}{N_{0,1}} + \frac{1}{N_{0,0}}$. Nyní známe odhad parametrů β_0, β_1 i jejich rozptyly, můžeme tedy testovat hypotézy o jejich nulovosti například pomocí Waldova testu tak, jak byl popsán výše.

1.4.2 Kategoriální proměnná

Nyní může proměnná nabývat $k > 2$ celočíselných hodnot. Takovéto proměnné můžeme dále dělit na ordinální, jejichž hodnoty lze uspořádat, a nominální, jejichž hodnoty nelze nijak logicky uspořádat. Příkladem ordinální může být nejvyšší dosažené vzdělání klienta (např. v pořadí: základní, středoškolské bez maturity, středoškolské s maturitou, vysokoškolské,...), nominální je třeba proměnná představující kraj, ve kterém klient žije - hodnoty přiřazené krajům nemají žádný význam, fungují jen jako identifikace jednotlivých krajů. Kategoriální proměnné používáme v modelu tak, že je zakódujeme do tzv. „dummy“ proměnných. Máme-li k kategorií, potřebujeme k-1 „dummy“ proměnných. Jednu kategorii zvolíme jako referenční a k ní budeme vztahovat ostatní kategorie. Tuto metodu si ilustrujeme na příkladě uvedeném v [3].

Studie, pro niž sestavujeme model, se týká přítomnosti chronické nemoci v závislosti na rase pacienta, která může být: běloch, černoch, hispánec a ostatní. Tabulka 1.1 ukazuje jeden ze způsobů zakódování proměnné x značící rasu do „dummy“ proměnných, jako referenční kategorie se zde bere běloch.

Logit tohot modelu bude $g(x) = \beta_0 + \beta_1 D1 + \beta_2 D2 + \beta_3 D3$. Ukažme si nyní

		Dummy		
Rasa	**(kód)**	**D1**	**D2**	**D3**
běloch	(1)	0	0	0
černoch	(2)	1	0	0
hispánec	(3)	0	1	0
ostatní	(4)	0	0	1

Tabulka 1.1: Kódování proměnné rasa

logaritmus podílu šancí černocha a bělocha:

$$\begin{aligned}
\ln\left[\widehat{OR}(\check{c}ernoch, b\check{e}loch)\right] &= \hat{g}(\check{c}ernoch) - \hat{g}(b\check{e}loch) \\
&= \left[\hat{\beta}_0 + \hat{\beta}_1(D1=1) + \hat{\beta}_2(D2=0) + \hat{\beta}_3(D3=0)\right] \\
&- \left[\hat{\beta}_0 + \hat{\beta}_1(D1=0) + \hat{\beta}_2(D2=0) + \hat{\beta}_3(D3=0)\right] \\
&= \hat{\beta}_1.
\end{aligned}$$

Podíly šancí a varianční matici pro koeficienty můžeme odhadnout využitím empirických odhadů pravděpodobností analogicky jako v případě binární proměnné, stejně tak i Waldův test pro testování hypotéz o nulovosti koeficientů a stanovení intervalů spolehlivosti pro jednotlivé koeficienty.

1.4.3 Spojitá proměnná

Příkladem spojité proměnné může být například měsíční příjem klienta. Šance pro klienta se vstupní hodnotou x je $odds(x) = \dfrac{\dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{\dfrac{1}{1+e^{\beta_0+\beta_1 x}}}$. Podíl šancí pro dvě hodnoty vstupní proměnné lišící se o jednotku je $OR(x+1, x) = \dfrac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$. Parametr β_1 udává, jak se změní logaritmus šancí při jednotkovém přírůstku regresoru x. Nulovost tohoto parametru by znamenala, že šance by nezávisela na hodnotě vstupní proměnné a tudíž by byla stejná pro všechny klienty. Mnohdy se stává, že jednotkový přírůstek vstupní proměnné pro model nemusí být vůbec významný, protože 1 je v kontextu příliš „drobná“, například změna platu o 1Kč neznamená prakticky nic, ale změna o 10000Kč už významná je, nebo příliš „hrubá“, třeba když x nabývá hodnoty mezi nulou a jedničkou. Chceme vědět, co se stane, změní-li se hodnota x o nějaké námi libovolně zvolené c. Logaritmus podílu šancí bude rozdíl logitů $\ln(OR) = \ln\left(\dfrac{odds(x+c)}{odds(x)}\right) = c\beta_1$, podíl šancí je tedy $OR = e^{c\beta_1}$. Odhad tohoto podílu šancí $\widehat{OR}$ dostaneme dosazením maximálně věrohodného odhadu $\hat{\beta}_1$ do OR. Odhad směrodatné odchylky $c\hat{\beta}_1$ je $c\hat{\sigma}(\hat{\beta}_1)$. Již víme, že $\dfrac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}(\hat{\beta}_1)} = \dfrac{c\hat{\beta}_1 - c\beta_1}{c\hat{\sigma}(\hat{\beta}_1)} \overset{as}{\sim} N(0,1)$. Nyní můžeme sestrojit $100(1-\alpha)\%$-ní

asymptotický interval spolehlivosti pro tento podíl šancí:

$$\begin{aligned} P\left[-z_{(1-\alpha/2)} \leq \frac{c\hat{\beta}_1 - c\beta_1}{c\hat{\sigma}(\hat{\beta}_1)} \leq z_{(1-\alpha/2)}\right] &= P\left[e^{c\hat{\beta}_1 - z_{1-\alpha/2}c\hat{\sigma}(\hat{\beta}_1)} \leq e^{c\beta_1} \leq e^{c\hat{\beta}_1 + z_{1-\alpha/2}c\hat{\sigma}(\hat{\beta}_1)}\right] \\ &\approx 1-\alpha \\ OR = e^{c\beta_1} &\in (e^{c\hat{\beta}_1 - z_{1-\alpha/2}c\hat{\sigma}(\hat{\beta}_1)}, e^{c\hat{\beta}_1 + z_{1-\alpha/2}c\hat{\sigma}(\hat{\beta}_1)}). \end{aligned}$$

Někdy se může stát, že přímé použití spojité proměnné v modelu není zcela vhodné. Vezměme si třeba příjem. Riziko nesplacení úvěru bude na příjmech závislé jinak u lidí s velmi vysokými příjmy než u běžných klientů. Tento problém se dá řešit třeba vhodnou tranformací, například že do modelu nedosadíme přímo x, ale nějakou vhodnou funkci $f(x)$ (například $\sqrt{x}$). Jiným způsobem řešení tohoto problému je vytvořit ze spojité proměnné kategoriální proměnnou, například lidé s měsíčním platem do 20000Kč, 20000Kč-40000Kč, 40000Kč-70000Kč, více než 70000Kč. Pro tuto kategoriální proměnnou se vytvoří tzv. „dummy" proměnné, viz část 1.4.2.

1.5 Konstrukce modelu iterační metodou

Nyní si popíšeme rozšiřující se iterační algoritmus pro vhodný výběr regresorů, které zahrneme do modelu. Nazývá se Stepwise regression anglicky, krokový výběr česky. Je založen na kontrolování statistické významnosti jednotlivých regresorů. V každém kroku algoritmu budeme za nejvýznamnější regresor považovat takový, pro který bude co největší rozdíl logaritmických věrohodností modelu s ním a bez něj, což můžeme stanovit třeba právě statistikou G popsanou v části 1.3. V části 1.4.2 jsme si ukázali, že kategoriální proměnnou s k možnými hodnotami modelujeme pomocí $k-1$ „dummy" proměnných. Statistika G má asymptotické rodělení Chí-kvadrát, počet stupňů volnosti však závisí na počtu proměnných, jejichž absenci chceme testovat. Z tohoto důvodu při vzájemném porovnávání modelů s různými proměnnými a bez nich musíme počítat s různými stupni volnosti. Proto bude výhodnější používat jako kritérium významnosti p-hodnotu G.

Krok(0):
Předpokládejme, že máme k dispozici celkem p různých regresorů. Nejprve sestavíme model, který bude obsahovat pouze konstantu (anglicky intercept only model), a spočítáme jeho logaritmickou věrohodnost, kterou označíme jako L_0. Potom spočítáme $L_j^{(0)}\, \forall j = 1, \ldots, p$, což značí logaritmickou věrohodnost modelu obsahujícího pouze konstantu a regresor x_j. Takovéto značení budeme používat v průběhu celého algoritmu. Horní index (0) znamená nultý krok, dolní index j znamená, že jsme do modelu přidali proměnnou x_j. Označme $G_j^{(0)} = -2(L_0 - L_j^{(0)})$ statistiku použitou v testu, kterým budeme porovnávat model obsahující regresor x_j s modelem bez něj, $p_j^{(0)} = P\left[\chi_l^2(1-\alpha) > G_j^{(0)}\right]$ p-hodnotu tohoto testu, kde $p = 1$ pro x_j spojité a $p = k-1$ pro x_j kategoriální s k kategoriemi. Nejdůležitější proměnnou je ta s nejmenší p-hodnotou. Označme e_1 index regresoru, který je „kandidátem" pro vstup do kroku (1), $p_{e_1}^{(0)} = min\left(p_j^{(0)}\right)$. Je důležité uvědomit si, že x_{e_1} ještě nemusí být statisticky významný. Vyjde-li například $p_{e_1}^{(0)} = 0,83$,

v analýze asi nebudeme chtít pokračovat, protože odezva nejspíš nezávisí ani na nejdůležitější proměnné. Proto pro p-hodnotu zvolíme prahovou hodnotu p_E (E jako Entry). Dle [3] je vhodné volit p_E mezi 0,15 a 0,20. Proměnnou potom budeme chtít zahrnout do modelu, jestliže p-hodnota modelu s touto proměnnou bude menší než p_E. Jestliže $p_{e_1}^{(0)} < p_E$, přejdeme ke kroku (2), jinak následuje krok (S).

Krok(1):
Označme $L_{e_1}^{(1)}$ logaritmickou věrohodnost modelu obsahujícího konstantu a regresor x_{e_1}. Chceme zjistit, jestli mezi zbývajícími $p-1$ regresory je ještě nějaký významný. Spočítáme tedy věrohodnosti $L_{e_1,j}^{(1)}$ modelů obsahujících konstantu, x_{e_1} a x_j pro každé $j \in \{1,\ldots,p\} \setminus e_1$. Statistiku $G_j^{(1)}$ pro porovnání takovéhoto modelu s modelem bez x_j definujeme jako $G_j^{(1)} = -2(L_{e_1}^{(1)} - L_{e_1,j}^{(1)})$. Nechť x_{e_2} je takový regresor, pro nějž vyšla nejmenší p-hodnota věrohodnostního testu $p_{e_2}^{(1)} = min\left(p_j^{(1)}\right)$. Je-li $p_{e_2}^{(1)} < p_E$, pokračujeme krokem (2), jinak přejdeme ke kroku (S).

Krok(2):
Může se stát, že přidáním proměnné x_{e_2} do modelu ztrácí význam mít v modelu proměnnou x_{e_1}. V tomto kroku tedy budeme kontrolovat a případně vypouštět „zbytečné“ proměnné, jedná se o tzv. zpětnou regresi (anglicky backward regression). $L_{-e_j}^{(2)}$ je logaritmická věrohodnost modelu, ze kterého jsme odstranili x_{e_j}. Test v tomto kroku bude založen na statistice $G_{-e_j}^{(2)} = -2(L_{-e_j}^{(2)} - L_{e_1,e_2}^{(2)})$, $p_{e_j}^{(2)}$ bude p-hodnota tohoto testu. Kritérium pro rozhodnutí, zda vyřadit proměnnou x_{r_2}, kde $p_{r_2}^{(2)} = max\left(p_{-e_j}^{(2)}, p_{-e_2}^{(2)}\right)$, stanovíme pomocí námi předem zvolené kritické hodnoty pro vyřazování proměnných p_R (R z anglického slova Remove). Hodnota p_R musí být ostře větší než p_E, abychom předešli zacyklení algoritmu v důsledku toho, že budeme pak znovu zavádět do modelu proměnné, které jsme již jednou vyloučili. Regresor x_{r_2} vyloučíme z modelu, jestliže $p_{r_2}^{(2)} \geq p_R$, jinak v modelu zůstává. Pak následuje výběr další proměnné. Porovnáme model obsahující konstantu, x_{e_1}, x_{e_2} s modelem obsahujcím ještě navíc x_j pro každé $j = 1,2,\ldots,p$, $j \neq e_1, e_2$ na základě G statistiky, spočítáme příslušné p-hodnoty. Označme x_{e_3} proměnnou a takovou p-hodnotou $p_{e_3}^{(2)} = min\left(p_j^{(2)}\right)$. Platí-li $p_{e_3}^{(2)} < p_E$, zahrneme x_{e_3} do modelu a pokračujeme krokem (3), jinak přejdeme ke kroku (S).

Krok(3):
Tento krok je vlastně identický s krokem (2). Nejprve se provede zpětná regrese pro proměnnou vybranou v předchozím kroku a pak se vybere nový kandidát pro zařazení do modelu a napočítají se pro něj p-hodnoty, vše tak, jako v kroku (2). Toto se opakuje do té doby, než se algoritmus dostane do posledního kroku (S).

Krok(S):
K tomuto kroku dojde buď když model obsahuje všech p regresorů, nebo když všechny proměnné obsažené v modelu mají p-hodnotu pro vyřazování menší než p_R a současně všechny proměnné nezahrnuté v modelu mají p-hodnotu pro přidání

do modelu větší než p_E. V tomto stádiu tedy model obsahuje všechny proměnné, které jsou významné vzhledem ke konstantám p_E a p_R.

2. Posouzení kvality modelu

Nyní už máme pro naše data model a zajímá nás, jak dobře je vystihuje. V této kapitole popíšeme několik způsobů, jak kvalitu modelu měřit. Bude nás zajímat především shoda modelu s reálnými daty a diverzifikační síla modelu. Při testech budeme nyní předpokládat, že model dobře vystihuje závislost odezvy na regresorech, tedy že jsme pomocí výše uvedených metod a postupů zvolili vhodné proměnné a jejich koeficienty.

2.1 Testy dobré shody

Vyjdeme především z [3]. Zavedeme následující značení. Nechť náš model obsahuje p nezávislých proměnných, tedy $\mathbf{x} = (x_1, \ldots, x_p)$ a J udává počet různých pozorovaných hodnot $\mathbf{x}$. Jestliže u nějakých klientů byly napozorovány stejné hodnoty, bude $J < n$, n značí počet pozorovaných klientů. Označme m_j počet subjektů u nichž $\mathbf{x} = \mathbf{x_j}$, $j = 1, \ldots, J$. Je zřejmé, že $\sum_{i=1}^{J} m_j = n$. Dále označme z_j počet odezev rovnajících se jedné u subjektů jejichž $\mathbf{x} = \mathbf{x_j}$. Potom $\sum_{i=1}^{J} z_j = n_1$, kde n_1 je celkový počet odezev rovnajících se jedné. Pokud model obsahuje alespoň jednu spojitou proměnnou, předpokládáme, že $J \approx n$. Pro názornost si tyto informace shrneme do tabulky 2.1.

Skupina	**1**	**2**	$\cdots$	**J**
$\mathbf{x}$	$\mathbf{x_1}$	$\mathbf{x_2}$	$\cdots$	$\mathbf{x_J}$
m_j	m_1	m_2	$\cdots$	m_J
z_j	z_1	z_2	$\cdots$	z_J
$\hat{\pi}(\mathbf{x})$	$\hat{\pi}(\mathbf{x_1})$	$\hat{\pi}(\mathbf{x_2})$	$\cdots$	$\hat{\pi}(\mathbf{x_J})$
$\hat{z}_j$	$m_1\hat{\pi}(\mathbf{x_1})$	$m_2\hat{\pi}(\mathbf{x_2})$	$\cdots$	$m_J\hat{\pi}(\mathbf{x_J})$

Tabulka 2.1: Shrnutí informací pro testy dobré shody

Do skupiny j spadá m_j subjektů, z_j je pozorovaná četnost takových subjektů ze skupiny j, jejichž odezva se rovnala jedné. z_j nazveme pozorované četnosti. Můžeme je odhadnout tak, že $\hat{z}_j = m_j\hat{\pi}(\mathbf{x_j})$. $\hat{z}_j$ pak budeme říkat odhadnuté nebo teoretické četnosti.

2.1.1 Pearsonův Chí-kvadrát test

Chceme posoudit, jak moc se od sebe liší pozorované a teoretické četnosti z_j a $\hat{z}_j$. $\hat{z}_j = m_j\hat{\pi}(\mathbf{x_j}) = m_j\dfrac{e^{\hat{g}(x_j)}}{1 + e^{\hat{g}(x_j)}}$, kde $\hat{g}(x_j)$ je odhad logitu. Pearsonovo reziduum je definováno jako

$$r(z_j, \hat{\pi}(\mathbf{x_j})) = \frac{(z_j - m_j\hat{\pi}(\mathbf{x_j}))}{\sqrt{m_j\hat{\pi}(\mathbf{x_j})(1 - \hat{\pi}(\mathbf{x_j}))}}.$$

Pro testování zavedeme statistiku X^2

$$X^2 = \sum_{j=1}^{J} r\left(z_j, \hat{\pi}(\mathbf{x_j})\right)^2.$$

Ta má asymptoticky rozdělení Chí-kvadrát o $J - (p+1)$ stupních volnosti, viz [3].

Tento test má smysl pouze pokud $J \not\approx n$, tedy pokud se v modelu vyskytují jen kategoriální proměnné. Díky tomu totiž existuje omezený počet možností pro hodnoty $\mathbf{x}$, můžeme tedy zafixovat J a s rostoucím n rostou i m_j.

2.1.2 Hosmerův-Lemeshowův test

Pro každý subjekt máme pozorování $\mathbf{x}$ a odhad $\hat{\pi}(\mathbf{x})$. Subjekty rozdělíme do g skupin (obvykle se volí $g = 10$ a skupinám se potom říká „decily rizika“) tak, že první skupina bude obsahovat indexy klientů, jejichž hodnota $\hat{\pi}(\mathbf{x})$ je mezi $\frac{1}{g}100\%$ nejmenších hodnot, druhá skupina bude mít $\hat{\pi}(\mathbf{x})$ mezi $\frac{1}{g}100\%$ a $\frac{2}{g}100\%$ nejmenšími hodnotami a tak dále, až v g-té skupině budou indexy klientů, jejichž $\hat{\pi}(\mathbf{x})$ patří mezi $\frac{1}{g}\%$ největších. Označmě c_l l-tou skupinu indexů, n'_l počet indexů ve skupině c_l. Zavedeme $o_l = \sum_{i \in c_l} y_i$, což je vlastně počet subjektů, jejichž indexy jsou ve skupině c_l a odezva se rovná jedné. Můžeme si je představit například jako počet klientů ve skupině c_l, kteří nesplatili dluh. Dále $\bar{\pi}_l = \dfrac{\sum_{i \in c_l} \hat{\pi}(\mathbf{x_i})}{n'_l}$ je průměr odhadů $\hat{\pi}(\mathbf{x_i})$ ve skupině c_l. Hosmer-Lemeshowova statistika má tvar

$$\hat{C} = \sum_{l=1}^{g} \frac{(o_l - n'_l \bar{\pi}_l)^2}{n'_l \bar{\pi}_l (1 - \bar{\pi}_l)}.$$

Hosmer a Lemeshow na základě simulací ukázali, že pokud se $J = n$, pak rozdělení statistiky $\hat{C}$ lze dobře aproximovat rozdělením Chí-kvadrát o $g - 2$ stupních volnosti. Tuto aproximaci lze použít i když $J \approx n$, viz [3].

2.2 Diverzifikační schopnost modelu

Diverzifikační schopnost modelu je jednou z jeho stěžejních vlastností. Každému subjektu i přidělíme na základě příslušných pozorovaných dat skóre $s_i \in \mathbb{R}$ a budeme chtít podle tohoto skóre odhadnout, jestli se hodnota y_i rovná jedné nebo nule. Jako toto skóre můžeme použít například odhad $\hat{\pi}(x_i)$. Ideálně bychom chtěli stanovit hranici s_0 tak, aby odezva všech subjektů, jejichž skóre je menší než s_0, byla nula a odezva subjektů, jejichž skóre je větší než s_0, byla jedna. Divezifikační schopnost tedy udává, jak dobře dokáže model rozlišit subjekty s $y_i = 0$ od subjektů s $y_i = 1$ na základě hodnoty skóre s_i. Pro názornost si budeme představovat, že modelujeme riziko, že subjekty nesplatí svůj dluh. Klienty, kteří nesplatili, a tedy jejich $y_i = 1$ označíme jako špatné, ty, kteří splatili a tudíž $y_i = 0$ označíme jako dobré.

2.2.1 Lorenzova křivka

Lorenzova křivka, někdy též nazývána ROC křivka (z anglického Receiver Operating Characteristics), vhodným způsobem graficky znázorňuje diverzifikační sílu modelu.
Setřídíme klienty sestupně podle skóre a podíváme se na jejich hodnoty y_i. V ideálním případě by měly být samé jedničky a až po nich následovat nuly, což ovšem většinou v praxi nelze dosáhnout, protože se může stát, že klient se špatnými předpoklady dluh bez problému splatí, zatímco klient s dobrými předpoklady svůj dluh neuhradí. Myšlenka Lorenzovy křivky je založena na porovnání, jak moc se reálná posloupnost jedniček a nul liší od ideálního seřazení. Potřebujeme distribuční funkci skóre pro dobré klienty a pro špatné klienty. Tyto distribuční funkce však většinou neznáme, takže je odhadneme empirickými distribučními funkcemi. Pro dobré klienty („Good“) dostáváme

$$F_G(s) = \frac{1}{n_G} \sum_{j=1}^{n} \mathbb{I}_{(-\infty,s]}(s_j) y_j, \ s \in \mathbb{R},$$

pro špatné

$$F_B(s) = \frac{1}{n_B} \sum_{j=1}^{n} \mathbb{I}_{(-\infty,s]}(s_j)(1 - y_j), \ s \in \mathbb{R},$$

kde n_G je počet dobrých klientů, n_B špatných a výraz $\mathbb{I}_{(-\infty,s]}(s_i)$ dá jedničku, jestliže $s_i \in (-\infty, s]$, jinak nulu.

Lorenzova křivka vznikne spojením bodů $[F_G(s_i), F_B(s_i)]$, $i = 1, \ldots, n$, leží tedy uvnitř jednotkového čtverce. Vychází z bodu $[0, 0]$ a končí v bodě $[1, 1]$.

Máme-li stanovenu mezní hodnotu skóre s_0 pro zařazní klientů mezi dobré nebo špatné na základě výpočtu jejich skóre, můžeme Lorenzovu křivku interpretovat také pomocí klasifikační tabulky 2.2.

skutečnost	0	1
predikce		
1	TP	FP
0	FN	TN

Tabulka 2.2: Klasifikační tabulka

TP („True Positive“) značí počet subjektů, jejichž skóre je větší než s_0 a odezva se rovná jedné, FP („False Postive“) počet subjektů se skóre větším než s_0, ale odezvou rovnou nule, FN („False Negative“) počet subjektů se skóre menším, než s_0, ale odezvou rovnou jedné a TN („True Negative“) počet subjektů se skóre menším, než s_0, a odezvou rovnou nule. Představme si na chvíli, že výpočet skóre je test, zda-li bude odezva rovna jedné nebo nule. Senzitivita testu je $P(\hat{Y} = 1 \mid Y = 1)$, její odhad se rovná $\frac{TP}{TP + FN}$. Specificita testu je $P(\hat{Y} = 1 \mid Y = 0)$, odhad se rovná $\frac{FP}{FP + TN}$. Lorenzova křivka je potom grafem ukazujícím závislost (1-specificity) na senzitivitě.

V ideálnímm případě křivka kopíruje nejdříve levou a pak horní hranu čtverce. Naopak prakticky žádnou diversifikační schopnost nemá model, u nějž Lorenzova křivka leží blízko diagonály. Model, jehož Lorenzova křivka je pod diagonálou, neodpovídá dobře datům, může být například „převrácený“. Jednou z možností, jak měřit schopnosti diverzifikace modelu, je spočítat plochu pod křivkou - AUC (z anglického Area Under Curve). Na tomto je založen například Giniho koeficient.

2.2.2 Giniho koeficient

Označme A jako plochu nad Lorenzovou křivkou, B jako plochu mezi Lorenzovou křivkou a diagonálou. $A + AUC$ tedy dává celou plochu čtverce ($A + AUC = 1$), $A + B$ dává polovinu plochy čtverce ($A + B = \frac{1}{2}$). Ještě si všimněme, že $A = 1 - AUC$. Giniho koeficient definujeme jako poměr plochy B k ploše nad diagonálou:

$$Gini = \frac{B}{A+B} = 2B = 1 - 2A = 1 - 2(1 - AUC) = 2AUC - 1.$$

Je zřejmé, že Giniho koeficient nabývá hodnot z intervalu $[-1, 1]$, kde hodnota 1 znamená nejlepší diverzifikační schopnost, nula žádnou diverzifikaci a záporné hodnoty ukazují, že se jedná o naopak postavený model.

2.2.3 Kolmogorova-Smirnovova statistika

Další možností, jak měřit diverzifikační schopnost modelu, je Kolmogorova-Smirnovova statistika. Tu definujeme pomocí distribučních funkcí špatných klientů a dobrých klientů, respektive výběrových distribučních funkcí dobrých a špatných klientů zadefinovaných v části 2.2.1, jako

$$KS = \sup_{s \in \mathbb{R}} |F_B(s) - F_R(s)|.$$

KS může nabývat hodnot z intervalu $[0, 1]$. $KS = 0$, pokud model diverzifikační schopnost nemá vůbec, naopak $KS = 1$, jestliže je diverzifikační schopnost modelu dokonalá. Kolmogorovu-Smirnovovu statistiku můžeme interpretovat i pomocí Lorenzovy křivky a to tak, že KS se přibližně rovná maximální vzdálenosti bodů Lorenzovy křivky od diagonály vynásobené $\sqrt{2}$. Odvození tohoto vztahu najdeme v [4].

2.2.4 Zobeněný koeficient determinace

Jedním z měřítek vypovídajícím o kvalitě modelu je také zobecněný koeficient determinace, který využívá věrohodnostní funkce modelu. Cox a Snell jej definovali jako

$$R^2 = 1 - \left(\frac{l(\mathbf{0})}{l(\boldsymbol{\beta})}\right)^{\frac{2}{n}},$$

kde $l(\mathbf{0})$ je věrohodnost modelu obsahujícího pouze konstantu a $l(\hat{\boldsymbol{\beta}})$ věrohodnost použitého modelu s vektorem parametrů $\boldsymbol{\beta}$ (viz [5], strana 4115). Tento koeficient

nabývá maxima $R^2_{max} = 1 - \left(l(\hat{\mathbf{0}})\right)^{\frac{2}{n}}$, což je menší než jedna. Rádi bychom však měli pevnou maximální hodnotu, které může ukazatel nabývat pro ideální model. Proto se používá modifikace tohoto ukazatele, a to Nagelkerkeho koeficient determinace

$$R^2_N = \frac{R^2}{R^2_{max}},$$

což je vlastně jen úprava R^2 tak, aby pro ideální model byl výsledek jedna.

3. Aplikace na data

Doposud jsme se logistické regresi věnovali teoreticky. Ukázali jsme si, jak vybudovat model a jak otestovat jeho vypovídající schopnosti. Tuto teorii si ukážeme na datech ze souboru datalogreg1.sas7bdat (najdeme ho na přiloženém CD ve složce data) pomocí statistického software SAS. Budeme modelovat pravděpodobnost škody v závislosti na tarifní skupině, věku, regionu a pohlaví. Přiložené CD dále obsahuje ve složce projekt soubor projekt.epg, což je vlastně celé zpracování úlohy v programu SAS, a výstupní soubory ve složce report.

3.1 Data

Data v souboru datalogreg1.sas7bdat jsou simulovaná dle mírně upravených reálných charakteristik. Obsahují informace o šedesáti tisících uzavřených smlouvách o pojištění odpovědnosti z provozu motorového vozidla, a to identifikační číslo smlouvy, číslo tarifní skupiny, region, věk, pohlaví a informaci, zda nastala škoda. Identifikační číslo smlouvy (v datech označeno smlo_id) je pouze pomocná informace, v modelu jej nijak nepoužijeme. Tarifních skupin (v datech TS) je celkem pět podle objemu motoru vozidla: do 1000 cm^3, 1000 cm^3-1350 cm^3, 1350 cm^3-1850 cm^3, 1850 cm^3 - 2500 cm^3, nad 2500 cm^3. Tarifní skupinu zahrneme do modelu jako kategoriální proměnnou, kterou zakódujeme pomocí „dummy“ proměnných způsobem uvedeným v tabulce 3.1. Tarifní skupina číslo pět představuje

	Dummy			
Tarifní skupina	**T1**	**T2**	**T3**	**T4**
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1
5	0	0	0	0

Tabulka 3.1: Zakódování proměnné TS

referenční kategorii.

Další proměnnou je *region*, má čtyři kategorie podle počtu obyvatel v místě bydliště, a to nad 500000, 50000-500000, 5000-50000, do 5000. Zakódování je v tabulce 3.2. Region číslo čtyři představuje referenční kategorii.

V datech najdeme také informaci o věku klientů a to v proměnné *veks*, kde je věk vyjádřed v letech, a v proměnné *vek*, která rozděluje klienty do tří kategorií: 18-30 let, 30-65 let, 65 a více let. Kódování kategoriální proměnné *vek* vidíme v tabulce 3.3. Třetí věková skupina je referenční kategorií. Do modelu samozřejmě nebudeme zahrnovat proměnnou *vek* i *veks*, vzhledem k tomu, že by nám to nepřineslo žádnou další informaci oproti modelu s pouze jednou z těchto proměnných. Pro některé statistiky bude vhodnější použít proměnnou *vek*, pro jiné zase *veks*, takže si vybudujeme dva modely.

Data obsahují také informace o pohlaví, proměnná *pohlavi*, kde 1 představuje

	Dummy		
region	**R1**	**R2**	**R3**
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

Tabulka 3.2: Zakódování proměnné *region*

	Dummy	
věková skupina	**V1**	**V2**
1	1	0
2	0	1
3	0	0

Tabulka 3.3: Zakódování proměnné *vek*

ženu, 2 muže. V modelu tuto proměnnou opět zakódujeme, a to tak, že $poh = 1$ pro ženu, 0 pro muže, tedy muž představuje referenční kategorii.

Proměnná *skoda* se rovná jedné, pokud u klient skutečně způsobil škodu během sledovaného roku, nule, pokud ne. Toto je odezva, kterou budeme modelovat.

Jak vhodně stanovit jednotlivé kategorie pro kategoriální proměnné není předmětem této práce, a tudíž se tím ani nebudeme zabývat.

3.2 Vybudování modelů

Vybudujeme dva modely, jeden s proměnnou *veks* a druhý s proměnnou *vek*. Druhý model bude obsahovat pouze kategoriální proměnné. Metodou maximální věrohodnosti odhadneme koeficienty u jednotlivých regresorů pro oba modely a dostáváme logity

$$\begin{aligned} g_1(\mathbf{x}) &= -1,4817 - 0,3226T1 - 0,3070T2 - 0,2264T3 - 0,1885T4 \\ &+ 0,4653R1 + 0,2386R2 + 0,1342R3 + 0,5142poh + 0,00435veks \end{aligned}$$

a

$$\begin{aligned} g_2(\mathbf{x}) &= -1,2927 - 0,3227T1 - 0,3077T2 - 0,2261T3 - 0,1880T4 \\ &+ 0,4633R1 + 0,2377R2 + 0,1350R3 + 0,5129poh + 0,265V1 \\ &+ 0,00438V2. \end{aligned}$$

Výsledky Waldova testu pro každou proměnnou shrnují tabulky 3.4 a 3.5
V prvním modelu vycházejí všechny p-hodnoty velmi malé, tudíž bychom podle tohoto kritéria do modelu měli zahrnou všechny uvedené proměnné. Ve druhém modelu jsou ovšem p-hodnoty u proměnných $V1$ a $V2$ větší než $0,05$ a tedy na 95% hladině tyto proměnné nejsou pro model významné a tudíž bychom je podle tohoto kritéria do modelu zahrnovat neměli.

proměnná	**odhad koeficientu**	**směrodatná chyba**	**Waldova statistika**	**p -hodnota**
konstanta	-1,4817	0,0447	1100,8789	<0,001
$T1$	-0,3226	0,0370	75,9729	<0,001
$T2$	-0,3070	0,0317	93,6898	<0,001
$T3$	-0,2264	0,0315	51,6387	<0,001
$T4$	-0,1885	0,0314	36,0577	<0,001
$R1$	0,4633	0,0324	206,765	<0,001
$R2$	0,2386	0,0289	68,0915	<0,001
$R3$	0,1342	0,0292	21,1688	<0,001
poh	0,5142	0,0188	744,9331	<0,001
veks	0,00435	0,000599	52,7675	<0,001

Tabulka 3.4: Výsledky Waldova testu v prvním modelu

proměnná	**odhad koeficientu**	**směrodatná chyba**	**Waldova statistika**	**p -hodnota**
konstanta	-1,2927	0,0388	1110,2777	<0,0001
$T1$	-0,3227	0,0370	76,0935	<0,0001
$T2$	-0,3077	0,0317	94,1791	<0,0001
$T3$	-0,2261	0,0315	51,5188	<0,0001
$T4$	-0,1880	0,0314	35,9009	<0,0001
$R1$	0,4633	0,0323	205,1756	<0,0001
$R2$	0,2377	0,0289	67,6152	<0,0001
$R3$	0,1350	0,0292	21,4289	<0,0001
poh	0,5129	0,0188	742,15801	<0,0001
$V1$	0,265	0,0263	1,0151	0,3137
$V2$	0,00438	0,0229	0,0364	0,8486

Tabulka 3.5: Výsledky Waldova testu ve druhém modelu

Dalším způsobem, jak rozhodnout, které regresory do modelu zahrnout, je krokový výběr (anglicky Stepwise regression). Pro oba modely nastavíme prahové hodnoty $p_E = 0,20$ a $p_R = 0,25$. Průběh algoritmu pro první model je následující:

Krok(0): Pracujeme s modelem obsahujícím pouze konstantu, jejíž odhad vyšel $-1,\overline{0246}$, standartní chyba $0,00926$, Waldova statistika $12242,7220$ a p-hodnota Waldova testu menší než $0,0001$. Analýzu kandidátů pro vstup do modelu shrnuje tabulka 3.6.

kandidát	$\mathbf{G}^{(0)}_{\mathbf{kandidát}}$	$\mathbf{p}^{(0)}_{\mathbf{kandidát}}$
TS	112,8175	<0,0001
region	234,8101	<0,0001
pohlavi	745,7303	<0,0001
veks	48,3509	<0,0001

Tabulka 3.6: Stepwise regression pro první model, Krok(0), kandidáti pro vstup

Největší hodnotu G statistiky a tudíž i nejmenší p-hodnotu pozorujeme u regresoru pohlavi, zahrneme jej tedy do modelu a pokračujeme dál.

Krok(1): Model obsahuje konstantu a proměnnou *pohlavi*. Maximálně věrohodné odhady vychází $-1,2947$ pro konstantu a $0,5102$ pro koeficient regresoru pohlavi. Provedeme ještě zpětnou regresi pro proměnnou pohlavi. Statistika $G^{(1)}_{-pohlavi}$ vyjde $738,8054$, p-hodnota $p^{(1)}_{r_1} < 0,0001$, takže *pohlavi* v modelu zůstává. Dále v tabulce 3.7 analyzujeme zbylé kandidáty pro vstup. Zařadíme pro-

kandidát	$\mathbf{G^{(1)}_{kandidát}}$	$\mathbf{p^{(1)}_{kandidát}}$
TS	113,4994	<0,0001
region	237,4739	<0,0001
veks	51,0782	<0,0001

Tabulka 3.7: Stepwise regression pro první model, Krok(1), kandidáti pro vstup

měnnou region.

Krok(2): V modelu máme konstantu, *pohlavi* a *region*. Maximálně věrohodné odhady uvádíme v tabulce 3.8.

regresor	**odhad koeficientu**
konstanta	-1,5012
$R1$	0,4614
$R2$	0,2369
$R3$	0,1336
poh	0,5121

Tabulka 3.8: Stepwise regression pro první model, Krok(2),maximálně věrohodné odhady koeficientů

Výsledky zpětné regrese jsou v tabulce 3.9.

kandidát	$\mathbf{G^{(2)}_{-kandidát}}$	$\mathbf{p^{(2)}_{kandidát}}$
region	236,1981	<0,0001
pohlavi	741,3293	<0,0001

Tabulka 3.9: Stepwise regresion pro první model, Krok(2), výsledky zpětné regrese

P-hodnoty jsou malé, proto nevyřadíme žádnou proměnnou. V tabulce 3.10 analyzujeme kandidáty pro vstup do modelu. Do modelu vstoupí proměnná TS.

kandidát	$\mathbf{G^{(2)}_{kandidát}}$	$\mathbf{p^{(2)}_{kandidát}}$
TS	114,5965	<0,0001
veks	53,5101	<0,0001

Tabulka 3.10: Stepwise regression pro první model, Krok(2), kandidáti pro vstup

Krok(3): Maximálně věrohodné odhady koeficientů regresorů pro současný model jsou v tabulce 3.11.

regresor	**odhad koeficientu**
konstanta	-1,2837
$T1$	-0,3228
$T2$	-0,3079
$T3$	-0,2261
$T4$	-0,1880
$R1$	0,4634
$R2$	0,2375
$R3$	0,1350
poh	0,5129

Tabulka 3.11: Stepwise regression pro první model, Krok(3), maximálně věrohodné odhady koeficientů

Výsledky zpětné regrese obsahuje tabulka 3.12.

kandidát	$\mathbf{G^{(3)}_{-kandidát}}$	$\mathbf{p^{(3)}_{kandidát}}$
TS	114,2265	<0,0001
region	237,2975	<0,0001
pohlavi	742,0614	<0,0001

Tabulka 3.12: Stepwise regresion pro první model, Krok(3), výsledky zpětné regrese

Vidíme, že všechny p-hodnoty jsou výrazně menší než p_R, žádnou proměnnou tedy nevyřadíme. Ještě se podívejme na posledního kandidáta pro vstup do modelu, proměnnou *veks*. Hodnota $G_j^{(3)} = 52,8360$ a $p_j^{(3)} < 0,0001$, *veks* tedy do modelu zahrneme.

Krok(4): Maximálně věrohodné odhady koeficientů u proměnných nyní najdeme v tabulce 3.13.
Výsledky zpětné regrese uvádíme v tabulce 3.14.
Opět nebudeme vyřazovat žádnou proměnnou.

Krok (S): Model obsahuje všechny proměnné, které máme k dispozici, což odpovídá i výsledkům získaným pomocí Waldova testu.

Tento algoritmus provedeme analogicky pro druhý model. Vzhledem k tomu, že model obsahuje stejné proměnné kromě regresoru *veks*, který je nahrazen regresorem *vek*, jsou výsledky algoritmu stejné až do kroku (3), s vyjímkou u informací týkajících se proměnné *vek*. Ukažme si je jen stručně v tabulkách 3.16,3.19,3.22,3.25,3.15,3.17,3.20,3.23,3.18,3.21 a 3.24.

Poznamenejme, že už v tabulce 3.16 si můžeme povšimnout poměrně vysoké p-hodnoty u proměnné vek.

regresor	odhad koeficientu
konstanta	-1,4817
$T1$	-0,3226
$T2$	-0,3070
$T3$	-0,2264
$T4$	-0,1885
$R1$	0,4653
$R2$	0,2386
$R3$	0,1342
poh	0,5142
$veks$	0,00435

Tabulka 3.13: Stepwise regression pro první model, Krok(4),maximálně věrohodné odhady koeficientů

kandidát	$\mathbf{G^{(4)}_{-kandidát}}$	$\mathbf{p^{(4)}_{kandidát}}$
TS	113,5529	<0,0001
$region$	239,7202	<0,0001
$pohlavi$	744,9331	<0,0001
$veks$	52,7675	<0,0001

Tabulka 3.14: Stepwise regresion pro první model, Krok(4), výsledky zpětné regrese

regresor	odhad koeficientu
konstanta	-1,0246

Tabulka 3.15: Stepwise regression pro druhý model, Krok(0),maximálně věrohodné odhady koeficientů

kandidát	$\mathbf{G^{(0)}_{kandidát}}$	$\mathbf{p^{(0)}_{kandidát}}$
TS	112,8175	<0,0001
$region$	234,8101	<0,0001
$pohlavi$	745,7303	<0,0001
vek	1,4395	0,4869

Tabulka 3.16: Stepwise regression pro druhý model, Krok(0), kandidáti pro vstup

regresor	odhad koeficientu
konstanta	-1,2947
$pohlavi$	0,5102

Tabulka 3.17: Stepwise regression pro druhý model, Krok(1),maximálně věrohodné odhady koeficientů

kandidát	$G^{(1)}_{-\text{kandidát}}$	$p^{(1)}_{\text{kandidát}}$
pohlavi	738,8054	<0,0001

Tabulka 3.18: Stepwise regresion pro druhý model, Krok(1), výsledky zpětné regrese

kandidát	$G^{(1)}_{\text{kandidát}}$	$p^{(1)}_{\text{kandidát}}$
TS	113,4994	<0,0001
region	237,4739	<0,0001
vek	1,5234	0,4669

Tabulka 3.19: Stepwise regression pro druhý model, Krok(1), kandidáti pro vstup

regresor	odhad koeficientu
konstanta	-1,5012
*R*1	0,4614
*R*2	0,2369
*R*3	0,1336
poh	0,5121

Tabulka 3.20: Stepwise regression pro druhý model, Krok(2),maximálně věrohodné odhady koeficientů

kandidát	$G^{(2)}_{-\text{kandidát}}$	$p^{(2)}_{\text{kandidát}}$
region	236,1981	<0,0001
pohlavi	741,3293	<0,0001

Tabulka 3.21: Stepwise regresion pro druhý model, Krok(2), výsledky zpětné regrese

kandidát	$G^{(2)}_{\text{kandidát}}$	$p^{(2)}_{\text{kandidát}}$
TS	114,5965	<0,0001
veks	53,5101	<0,0001

Tabulka 3.22: Stepwise regression pro druhý model, Krok(2), kandidáti pro vstup

regresor	odhad koeficientu
konstanta	-1,2837
$T1$	-0,3228
$T2$	-0,3079
$T3$	-0,2261
$T4$	-0,1880
$R1$	0,4634
$R2$	0,2375
$R3$	0,1350
poh	0,5129

Tabulka 3.23: Stepwise regression pro druhý model, Krok(3),maximálně věrohodné odhady koeficientů

kandidát	$\mathbf{G}^{(3)}_{-\mathbf{kandidát}}$	$\mathbf{p}^{(3)}_{\mathbf{kandidát}}$
TS	114,2265	<0,0001
region	237,2975	<0,0001
pohlavi	742,0614	<0,0001

Tabulka 3.24: Stepwise regresion pro druhý model, Krok(3), výsledky zpětné regrese

kandidát	$\mathbf{G}^{(3)}_{\mathbf{kandidát}}$	$\mathbf{p}^{(3)}_{\mathbf{kandidát}}$
vek	1,2485	0,5357

Tabulka 3.25: Stepwise regression pro druhý model, Krok(3), kandidáti pro vstup

Vidíme, že p-hodnota pro proměnnou *vek* je mnohem větší, než p_E, nebudeme ji tedy do modelu zahrnovat. Tím jsme vyčerpali všechny možnosti proměnných, čímž se dostáváme do kroku (S). Výsledný model tedy obsahuje všechny regresory, kromě *vek*. Tento výsledek se shoduje i s výsledkem získaným pomocí Waldova testu.

3.2.1 Výsledné modely

Na základě výše uvedené analýzy do prvního modelu zahrneme proměnné TS, *region*, *pohlavi*, *veks* a do druhého TS, *region*, *pohlavi*. Metodou maximální věrohodnosti odhadneme koeficienty u těchto proměnných a výsledné odhadnuté

logity budou tvaru:

$$\begin{aligned} \textit{logit prvního modelu} &= -1,4817 - 0,3226T1 - 0,3070T2 - 0,2264T3 \\ &- 0,1885T4 + 0,4653R1 + 0,2386R2 + 0,1342R3 \\ &+ 0,5142poh + 0,00435veks \\ \textit{logit druhého modelu} &= -1,2837 - 0,3228T1 - 0,3079T2 - 0,2261T3 \\ &- 0,1880T4 + 0,4634R1 + 0,2375R2 + 0,1350R3 \\ &+ 0,5129poh. \end{aligned}$$

3.2.2 Posouzení kvality modelů

Začneme testy dobré shody. První model obsahuje spojitou proměnnou *veks*, použijeme tedy Hosmerův-Lemeshovův test. Rozdělení klientů do skupin udává tabulka 3.26. $\mathbf{o_{skupina}}$ reprezentuje vlastně pozorovanou četnost výskytu $skoda = 1$ ve skupině a $\mathbf{n'_{skupina}} \cdot \bar{\pi}_{\mathbf{skupina}}$ si můžeme představit jako očekávanou četnost výskytu jevu $skoda = 1$ v dané skupině.

Skupina	**Počet klientů ve skupině**	$\mathbf{o_{skupina}}$	$\mathbf{n'_{skupina}} \cdot \bar{\pi}_{\mathbf{skupina}}$
1	6009	1064	1057,74
2	6010	1221	1175,68
3	5995	1249	1260,35
4	5999	1370	1369,78
5	5997	1468	1511,52
6	6013	1656	1644,03
7	6025	1750	1754,96
8	6000	1823	1853,97
9	5992	1962	1986,19
10	5960	2285	2235,35

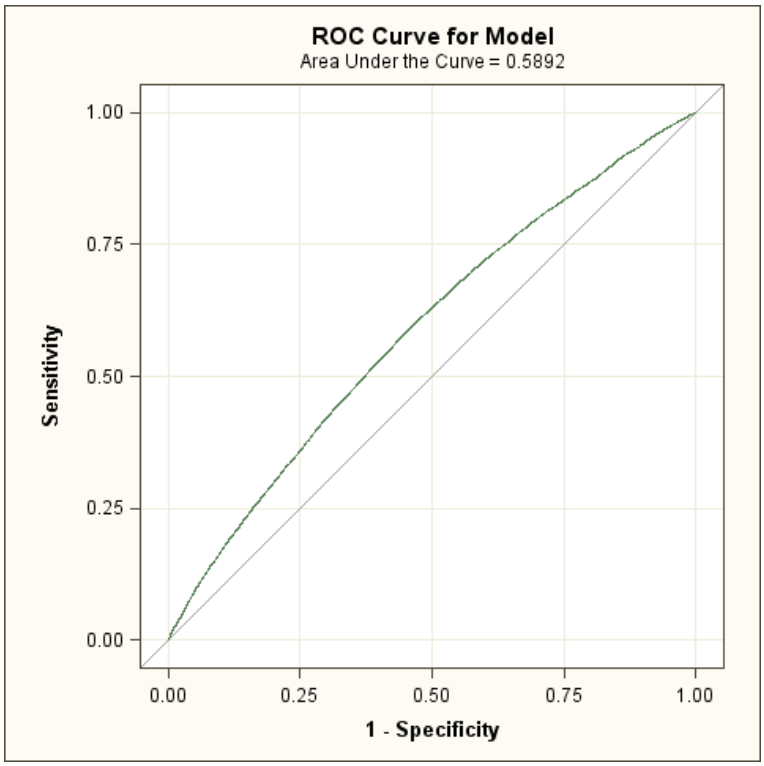
Tabulka 3.26: Hosmer-Lemeshowův test

Hodnota Hosmer-Lemeshowovy statistiky je 7,1151, počet stupňů volnosti 8 a p-hodnota testu 0,5243. Vidíme, že p-hodnota je poměrně vysoká, takže na hladině 95% zamítáme hypotézu, že model se shoduje s reálnými daty.

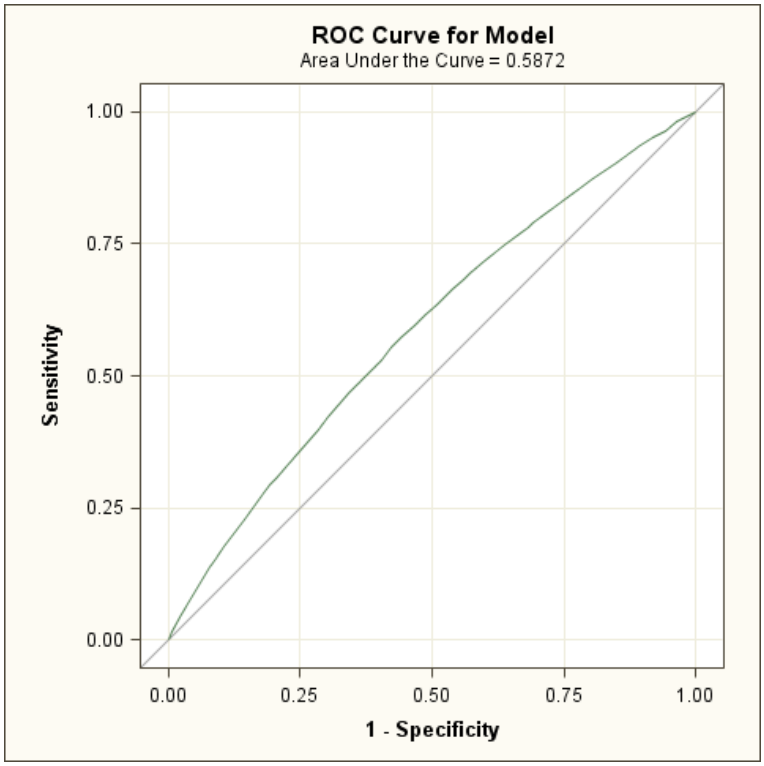
Druhý model obsahuje pouze kategoriální proměnné, a tak na něj provedem Pearsonův Chí-kvadrát test. Počet různých hodnot pozorování regresorů $J = 2200$. Hodnota statistiky X^2 vyjde 3333,8465, počet stupňů volnosti 2199 a p-hodnota tohoto testu menší než 0,0001. Nezamítáme tedy hypotézu, že se model dobře shoduje s daty.

Zaměřme se ještě na Diverzifikační schopnost modelů. Obrázek 3.2.2 ukazuje Lorenzovu křivku pro první model. Velikost oblasti pod křivkou je $AUC = 0,589$. Tomu také odpovídá Giniho koeficient, který se rovná $0,178$.

Lorenzovu křivku pro druhý model je znázorněna na obrázku 3.2.2. Hodnota AUC se rovná $0,587$ a Giniho koeficient $0,174$.

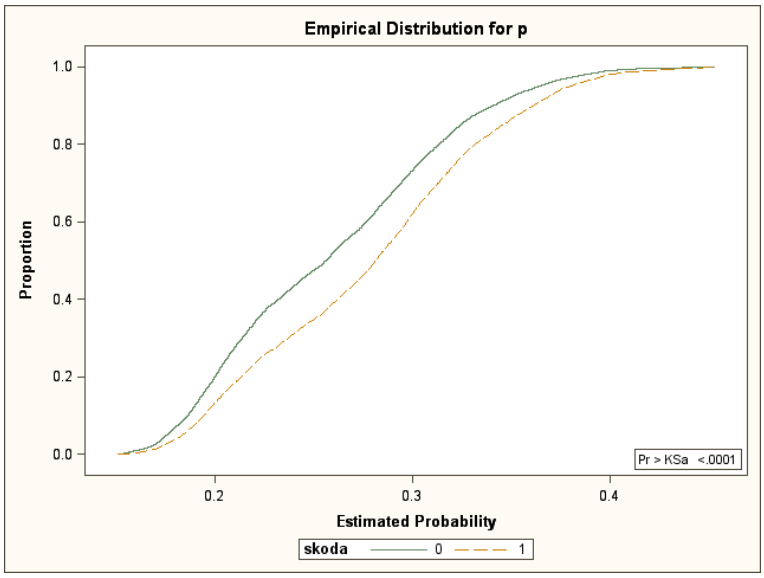


Obrázek 3.1: Lorenzova křivka pro první model

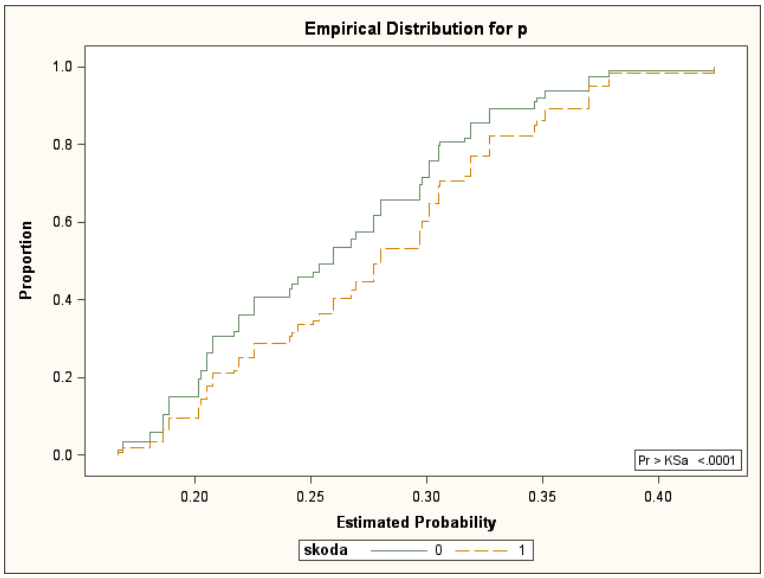


Obrázek 3.2: Lorenzova křivka pro druhý model

Zobrazíme si ještě empirické distribuční funkce „dobrých“ a „špatných“ klientů a spočítáme Kolmogorov-Smirnovovu statistiku. Tyto distribuční funkce pro první model můžeme vidět na obrázku 3.2.2, pro druhý model na obrázku. Hodnota Kolmogorovy-Smirnovovy statistiky pro první model je 0,059165, pro druhý 0,057329.



Obrázek 3.3: Empirické distribuční funkce dobrých a špatných klientů pro první model



Obrázek 3.4: Empirické distribuční funkce dobrých a špatných klientů pro druhý model

Jak Lorenzova křivka, tak i Giniho koeficient a Kolmogorova-Smirnovova statistika ukazují na lepší diverzifikační schopnost prvního modelu. To, že první model je kvalitnější, ukazuje i zobecněný koeficient determinace, který vychází $R^2 = 0,0190$ pro první model a $R^2 = 0,0181$. Pro lepší představu si uvedeme ještě Nagelkerkeho koeficient determinace $R^2_N = 0,0277$ pro první model, $R^2_N = 0,0264$ druhý.

3.3 Shrnutí praktické části

Můžeme si všimnout, že transformace spojité proměnné *veks* na kategoriální proměnnou *vek* byla tak zásadní, že se proměnné *vek* stala statisticky nevýznamnou pro model jak podle Waldova testu, tak podle výsledku algorimu Stepwise regression. To může být i konkrétním způsobem, jak se proměnná „zkategorizovala“. Dalo by se diskutovat o tom, jak zvolit kategorie vhodněji. Z praktického hlediska by ale každá kategorie měla být ve tvaru intervalu a ne libovolné množiny, už jen kvůli jednoduché interpretovatelnosti. Model zohledňující věk má poměrně vysokou p-hodnotu pro Hosmerův-Lemeshowův test dobré shody, naproti tomu durhý model má p-hodnotu Pearsonova Chí-kvadrát testu menší než 0.0001. Tyto dvě hodnoty však nemůžeme mezi sebou přímo porovnávat, protože se jedná o dva odlišné testy. Přesto nám však výsledky napovídají, že datům bude lépe odpovídat druhý model. Co se ale diverzifikační schopnosti týče, tam vynechání informace o věku způsobilo její zhoršení.

Závěr

Binární logistická regrese v mnoha oborech zajisté patří k hojně používaným statistickým metodám. Po teoretické stránce jsme se jí věnovali v prvních dvou kapitolách. Popsali jsme, jak vypadá základní model a metodou maximální věrohodnosti jsme odhadli jeho parametry. Ukázali jsme si statistiky pro testování významnosti parametrů. V jednorozměrném modelu jsme se zabývali vlastnostmi regresorů. Odhadu koeficientu binárního regresoru jsme zkonstruovali pomocí četností, díky čemuž v takovém případě není nutno používat metodu maximální věrohodnosti. Pro kategoriální proměnnou jsme na příkladě převzatém z [3] ukázali, jak takovouto proměnnou v modelu použít. U spojité proměnné jsme se zaměřili na význam koeficientu regresoru v jednorozměrném modelu. Pro vhodný výběr proměnných jsme popsali konstrukci modelu iterační metodu. Abychom změřili kvalitu modelu, ukázali jsme si dva testy dobré shody, a to Pearsonův Chí-kvadrát test a Hosmerův-Lemeshowův test, a znázornili a kvantifikovali diverzifikační sílu pomocí Lorenzovy křivky, Giniho koeficientu a Kolmogorovovy-Smirnovovy statistiky. Jako poslední měřítko kvality modelu jsme uvedli Zobeněný koeficient determinace. Ve třetí kapitole jsme provedli aplikaci teoretických poznatků na data z pojišťovnictví, konkrétně se jednalo o data obsahující informace o pojistných smlouvách o pojištění odpovědnosti z provozu motorového vozidla (známé jako povinné ručení). Zajímavým závěrem bylo, že tuto pravděpodobnost z regresorů, které jsme měli k dispozici, nejméně výrazně ovlivňuje věk klientů.

Seznam použité literatury

[1] Jiří Anděl. *Základy matematické statistiky*. MatfyzPress, 2005. ISBN 80-86732-40-1.

[2] Jitka Dupačová, Petr Lachout. *Úvod do optimalizace*. MatfyzPress, 2011. ISBN 978-80-7378-041-8.

[3] David W. Hosmer, Stanley Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, Inc., 2000. ISBN 0-471-35632-8.

[4] Markéta Ondrušková. *Odhadování a kritéria těsnosti modelu logistické regrese*. Bakalářská práce, Praha 2011. Univerzita Karlova v Praze, Matematicko-fyzikální Fakulta

[5] *SAS/STAT® 9.3 User's Guide*. SAS Institute Inc., 2011.

[6] Karel Zvára. *Regrese*. MatfyzPress, 2008. ISBN 978-80-7378-041-8.

Seznam obrázků

Seznam tabulek

Přílohy

Příloha č. 1

K práci je přiloženo CD obsahující data pro numerickou studii popsanou ve třetí kapitole, projekt vytvořený v programu SAS Enterprise Guide zpracovávající tato data a výstupy z tohoto projektu. Na CD dále nalezneme tuto práci ve formátu pdf.