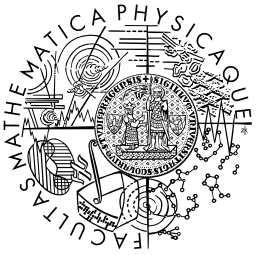


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Martin Šíma

Statistické učení a určování rizikovosti klientů

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Pešta Michal, Ph.D.

Studijní program: Matematika

Studijní obor: Finanční matematika

Praha 2021

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

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 má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Chtěl bych poděkovat vedoucímu práce doc. RNDr. Michalu Peštovi, Ph.D. za rady a připomínky při psaní této práce.

Název práce: Statistické učení a určování rizikovosti klientů

Autor: Martin Šíma

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Pešta Michal, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Práce popisuje výběr modelu pro klasifikační problém, představuje binární logistickou regresi. Přibližuje odhad parametrů maximalizací věrohodnosti a nutnost numerického vyčíslení, iterativně převažovaná metoda nejmenších čtverců je používána v práci. V práci jsou definovány Waldův a Wilksův test pro test statistické významnosti parametrů. Byly popsány rozdíly mezi kvalitativními a kvantitativními proměnnými a jejich interpretacemi. Popsány byly iterativní strategie konstrukce modelu, grafická analýza residuí, metody měření kvality modelu a odhadu testové chyby. Aplikační část přenáší nabyté poznatky na datový soubor bankovních klientů.

Klíčová slova: statistické učení určování rizikovosti klientů skóringové modely logistická regrese míry diskriminace

Title: Statistical learining and client risk assessment

Author: Martin Šíma

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Pešta Michal, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This bachelor thesis tackles model selection for classification problems and presents binary logistic regression. A method of parameter estimation is discussed and the necessity of numerical approach is explained, iteratively reweighted least squares method is used. The Wald and Wilks tests are defined to measure statistical significance of the parameters. The difference between qualitative and quantitative data and their interpretation is discussed. Model selection method of the stepwise selection is defined, residual plots and other methods of the model fit measurement are introduced. Acquired knowledge is applied to a data set of bank clients.

Keywords: statistical learning client risk assessment scoring models logistic regression discrimination measures

Obsah

Úvod

V dnešní době se často setkáváme s daty a jejich interpretacemi. Takové úsudky mohou mít dalekosáhle negativní i pozitivní dopady na fungování společnosti nebo firem.

Cílem statistického učení je tvorba prediktivních modelů, které nám umožňují lépe porozumět různým problematikám ať už jde o medicínu a farmacii, kde se modeluje například pravděpodobnost výskytu určitého onemocnění, o algoritmy, které přiřazují daným lidem personalizované reklamy s cílem maximalizovat prodeje, nebo o v této práci zkoumanou predikci pravděpodobnosti, zda je klient schopný splácet úvěr, nebo nikoliv na základě klientem poskytnutých informací, jeho dosavadního chování a jiných zjištěných skutečnostech.

Učení takovýchto modelů může probíhat jako *učení s učitelem* (Supervised learning), kde pro *nezávisle proměnné*, v textu je zaměňováno s pojmy regresory, prediktory, kovariáty nebo vysvětlující veličiny, jako je např. věk a průměrný příjem žadatele o úvěr, známe správnou *odezvu* v textu je zaměňováno s pojmem závisle proměnná, tedy zda klient úspěšně splácí nebo nesplácí úvěr. Klasickými příklady tohoto učení je lineární nebo logistická regrese, kterou si v tomto textu popíšeme detailněji. Moderní metody tohoto typu učení jsou kupříkladu zobecněný aditivní model (GAM), boosting, metoda podpůrných vektorů (Support vector machines).

Jiným typem učení, které musí využívat modely, které sledují měření nezávisle proměnných X_i, ale neznají již závisle proměnnou s nimi spjatou. Takové učení se nazývá *učení bez učitele* (Unsupervised learning). Taková situace může nastat třeba při hledání vzorců chování jednotlivců nebo skupin spotřebitelů na základě dat bez informace o jejich nákupní historii nebo tendencích. Tomu slouží metody shlukové analýzy.

Tvorba modelů závislosti odezvy na jiných nezávislých proměnných bude hlavním předmětem našeho zkoumání. Konkrétně se budeme věnovat rozdělení nula-jedničkové veličiny – Default ANO $\equiv 1$, default NE $\equiv 0$. Popíšeme, jak zhodnotit vliv jednotlivých regresorů (nezávisle proměnných) na odezvu a jak zhodnotit celkovou kvalitu modelu pomocí statistických testů. Projdeme také, jak vhodně nezávisle proměnné vybrat, aby nedošlo k tzv. overfittingu nebo zahrnutí proměnných, které na predikci buď vliv nemají nebo její kvalitu dokonce zhoršují.

1. Co je statistické učení

Obecně budeme hledat vztah závisle proměnné Y na prediktorech $\mathbf{X} = (X_1, X_2, \ldots, X_p)^\top$. Tento vztah zapíšeme jako

$$Y = f(\mathbf{X}) + \varepsilon, \tag{1.1}$$

kde f je předem daná, ale neznáma funkce $f : \mathbf{X} \longrightarrow \mathbb{R}$ a ε je náhodná veličina, jejíž $\mathbb{E}\varepsilon = 0$, nazýváme ji *chybový člen* (Error term). Náhodná veličina $\mathbf{X}$ je nezávislá na ε. Předpokládáme, že vektor $\mathbf{X}$ může ovlivňovat střední hodnotu Y, ale ne $\mathbb{V}\mathrm{ar}(Y)$. Zajímá nás, jestli a jak jednotlivé prvky vektoru $\mathbf{X}$ ovlivňují $\mathbb{E}Y$.

Rovnici 1.1 lze ekvivalentně zapsat jako

$$\begin{aligned} \mathbb{E}[Y|\mathbf{X}] &= f(\mathbf{X}), \\ \mathbb{V}\mathrm{ar}(Y|\mathbf{X}) &= \varepsilon^2. \end{aligned} \tag{1.2}$$

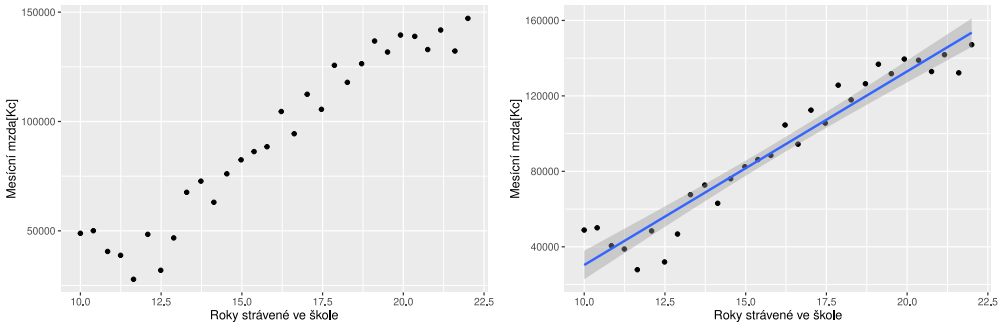
Nalezení f může být motivováno různě. První možností je, že chceme provést predikci pro Y. To může nastat, pokud bychom chtěli modelovat např. EAD (exposure at deafault) pro určitý úvěr, skupinu úvěrů nebo skupinu dlužníků. V takovém případě věřitele zajímá co nejpřesnější odhad očekávané ztráty pro tvorbu rezerv. Druhou motivací pak může být zájem o hlubší porozumění určité problematice. V situaci, kdy máme sice k dispozici velké množství prediktorů, ale jejich získávání je drahé, nás může zajímat, jaké prediktory mají na výstup nejpodstatnější vliv. Cílem pak může být eliminace sběru nepotřebných dat.

Odhad zapíšeme takto

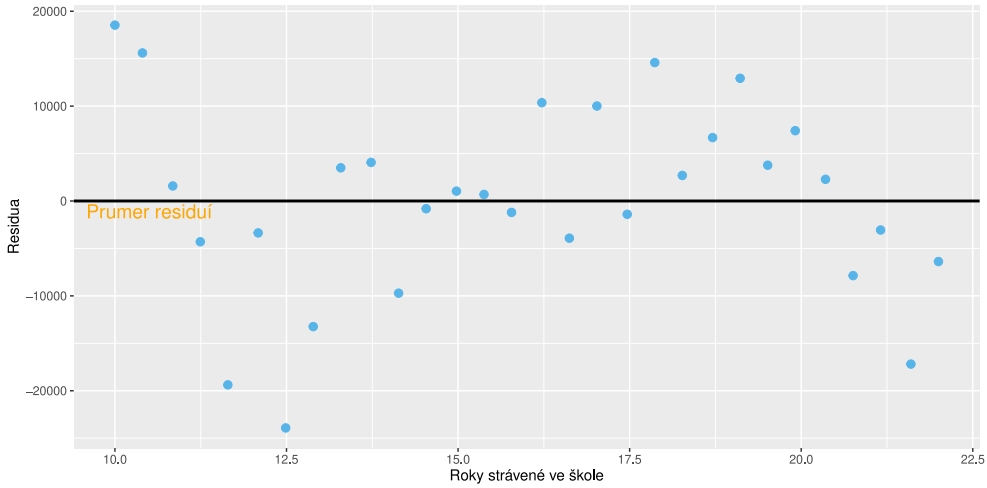
$$\hat{Y} = \hat{f}(\mathbf{X}), \tag{1.3}$$

kde $\hat{f}$ je nějakým odhadem funkce f a $\hat{Y}$ je predikce odezvy založená na tomto odhadu.

Jednoduchý příklad odhadu $\hat{f}$ si ukážeme na data setu přiloženému k James a kol. (2013). Data obsahují napozorované roční příjmy v USD proti letem vzdělání u 30 osob. V levém obrázku 1.1 vidíme, že by mohl existovat vztah mezi dosaženým vzděláním v podobě let strávených ve školském systému a mzdou. Modrá křivka na pravém obrázku ukazuje možný lineární vztah těchto veličin. Šedě zvýrazněná oblast je 95% interval spolehlivosti tohoto vztahu při užití modelu lineární regrese. Triviální vhled, zda tento model funguje, nám poskytne obrázek 1.2, kde vidíme graf *residuí* lineárního modelu, body tohoto grafu by neměly vykazovat žádný trend, to by nasvědčovalo nelineárnímu vlivu dat.



Obrázek 1.1: Income data set, přepočteno na Kč. Na dvojici obrázků vidíme napozorované mzdy a roky strávené ve škole.



Obrázek 1.2: Residua modelu z obrázku 1.1.

1.1 Odhadování

Způsobů, jak odhadovat $\hat{f}$, je mnoho. Abychom mohli určit, která metoda je pro daný problém vhodná je potřeba popsat potřebný aparát.

Přesnost predikce Y pomocí odhadu $\hat{Y}$ závisí na dvou faktorech – *redukovatelná* (reducible) a *neredukovatelná* (irreducible) chyba.

Věta 1. *Nechť $\hat{Y} = \hat{f}(\boldsymbol{X})$ je předem daný odhad $Y = f(\boldsymbol{X}) + \varepsilon$ a vektor prediktorů $\boldsymbol{X}$, $\boldsymbol{x}$ je pevná realizace tohoto náhodného vektoru. Podle značení v* (1.1) *a* (1.3) *lze chyby odhadu vyjádřit jako*

$$\mathbb{E}[(Y - \hat{Y})^2 | \boldsymbol{X} = \boldsymbol{x}] = \Big(f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})\Big)^2 + \mathbb{V}ar(\varepsilon),$$

kde první sčítanec je redukovatelná chyba a druhý neredukovatelná chyba.

Důkaz. Rozepíšeme:

$$\begin{aligned}\mathbb{E}[(Y - \hat{Y})^2 | \mathbf{X} = \mathbf{x}] &= \mathbb{E}[\Big(f(\mathbf{X}) + \varepsilon - \hat{f}(\mathbf{X})\Big)^2 | \mathbf{X} = \mathbf{x}] \\ &= \mathbb{E}[\Big(f(\mathbf{X}) - \hat{f}(\mathbf{X})\Big)^2 | \mathbf{X} = \mathbf{x}] \\ &+ \mathbb{E}[2\Big(f(\mathbf{X}) - \hat{f}(\mathbf{X})\Big)\varepsilon | \mathbf{X} = \mathbf{x}] + \mathbb{E}[\varepsilon^2 | \mathbf{X} = \mathbf{x}].\end{aligned}$$

Podle (1.1) předpokládáme, že $\mathbb{E}\varepsilon = 0$. Protože $\mathbf{X}$ a ε jsou nezávislé náhodné veličiny, platí

$$\begin{aligned}\mathbb{E}[(Y - \hat{Y})^2|\mathbf{X} = \mathbf{x}] &= \Big(f(\mathbf{x}) - \hat{f}(\mathbf{x})\Big)^2 \\ &\quad + 2\mathbb{E}[f(\mathbf{X}) - \hat{f}(\mathbf{X})|\mathbf{X} = \mathbf{x}]\mathbb{E}[\varepsilon|\mathbf{X} = \mathbf{x}] + \mathbb{E}[\varepsilon^2|\mathbf{X} = \mathbf{x}] \\ &= \Big(f(\mathbf{x}) - \hat{f}(\mathbf{x})\Big)^2 + \mathbb{E}[\varepsilon^2] \\ &= \Big(f(\mathbf{x}) - \hat{f}(\mathbf{x})\Big)^2 + \mathbb{V}\mathrm{ar}(\varepsilon),\end{aligned}$$

kde poslední rovnost plyne z $\mathbb{E}\varepsilon = 0$ a $\mathbb{V}\mathrm{ar}(\varepsilon) = \mathbb{E}[(\varepsilon - \mathbb{E}\varepsilon)^2]$.

□

Kvalita odhadu $\hat{f}$ může být obecně různá. Věta 1 nám ukazuje, že pokud chceme odhad vylepšovat, je to možné pouze snížením redukovatelné chyby např. vhodnější volbou metody odhadu $\hat{f}$, jakkoliv blízko se dobrým odhadem dostaneme, chyby, kterou přináší chybový člen ε, se zbavit nelze. Chybový člen může obsahovat kupříkladu námi nenaměřené prediktory. My se budeme soustředit na snižování redukovatelné chyby.

1.1.1 Měření kvality odhadů

Popišme si, co dělá model „dobrým“. V dalším textu bude chybový člen ε mít vlastnosti popsané v (1.1). Uvažujme nyní, že $\mathbf{X} = \mathbf{x}$ a $\hat{f}(\mathbf{x})$ je náhodnou veličinou, protože vzniká na základě náhodného výběru.

Definice 1 (Podmíněné vychýlení)**.** *Nechť $\hat{f}(\boldsymbol{X})$ je odhadem $Y = f(\boldsymbol{X}) + \varepsilon$ s konečnou střední hodnotou, $\boldsymbol{x}$ je pevnou realizací náhodného vektoru $\boldsymbol{X}$. Pak* podmíněné vychýlení odhadu $\hat{f}$ *je*

$$Bias(\hat{f}|\boldsymbol{x}) \equiv \mathbb{E}[\hat{f(\boldsymbol{x})}] - f(\boldsymbol{x}). \tag{1.4}$$

Definice 2 (Střední čtvercová odchylka (Mean squared error))**.** *Nechť $\hat{f}(\boldsymbol{X})$ je odhadem $Y = f(\boldsymbol{X}) + \varepsilon$ s konečným rozptylem. Pak* střední čtvercová odchylka $\hat{f}$ *je*

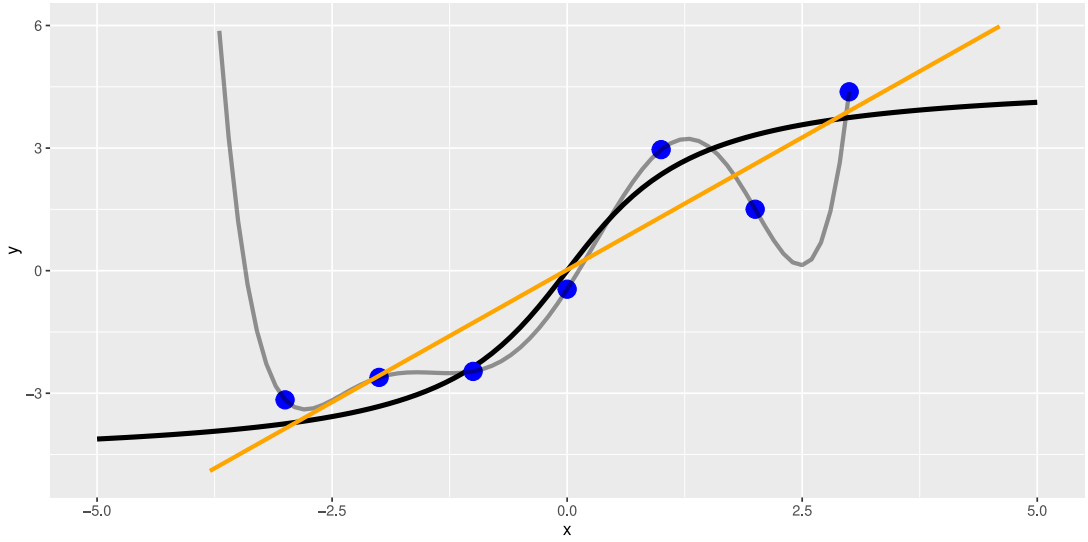
$$MSE(\hat{f}) \equiv \mathbb{E}[\hat{f}(\boldsymbol{X}) - f(\boldsymbol{X})]^2. \tag{1.5}$$

Definice 2 nám dává základní nástroj k měření přesnosti odhadu $\hat{f}$ od napozorovaných dat. Protože, jak ukazuje věta 1, chceme snižovat redukovatelnou chybu, je přirozené, že snižování $\mathrm{MSE}(\hat{f})$ bude k tomuto našemu cíli přispívat. Pro blízké předpovězené hodnoty napozorovaným datům dostaneme nízkou hodnotu MSE, naopak pro vysoké odchylky odhadu od naměřených dat je hodnota MSE vyšší.

Definice 3. *Nechť $\boldsymbol{D}_{training}$ a $\boldsymbol{D}_{test}$ jsou disjunktní množiny.*
$\boldsymbol{D}_{training} = \{(Y_i,\boldsymbol{X}_i) : i = 1,2,\ldots,k\}$, $\boldsymbol{D}_{test} = \{(Y_j,\boldsymbol{X}_j) : j = 1,2,\ldots,l\}$ a $k + l = n$, pak definujeme

$$\boldsymbol{D} \equiv \boldsymbol{D}_{training} \cup \boldsymbol{D}_{test},$$

kde $\boldsymbol{D}$ jsou data, $\boldsymbol{D}_{training}$ jsou tréninková data, $\boldsymbol{D}_{test}$ jsou testovací data a n je rozsah dat.



Obrázek 1.3: Srovnání modelu lineární regrese s interpolací Lagrangeovým polynomem. Černě $f(x)$, modře napozorovaná tréninková data, šedě polynom šestého stupně, oranžově model lineární regrese.

Pro model, který učíme na množině $\mathbf{D}_{training}$, může být $\text{MSE}(\hat{f}) = 0$, to nám ale neříká nic o prediktivních schopnostech a vlastnostech modelu $\hat{f}$ na množině zatím nenapozorovaných dat $\mathbf{D}_{test}$. Obecně nás totiž zajímá hlavně jak spolehlivě dokáže daný model přinášet předpovědi na datech, která nebyla využita pro jeho učení. Pokud bychom například vytvářeli algoritmus, který predikuje vývoj akciových trhů na základě historických dat, nezajímá nás, jak přesně model dokáže predikovat ceny pro minulý týden, důležité je jestli model správně určí vývoj pro následující den.

Pokud počítáme MSE z dat použitých pro učení, nazveme jej *tréninková střední čtvercová odchylka*, nebo tréninková MSE, a značíme $\text{MSE}_{training}$. Naopak MSE pro data, která nebyla využita pro učení modelu, dostáváme *testovací střední čtvercovou odchylku*, nebo testovací MSE, a značíme MSE_{test}.

Příklad: Mějme realizaci dat $\mathbf{D}_{training} = \{(y_1,x_1)^\top, (y_2,x_2)^\top, \ldots, (y_7,x_7)^\top\}$ a $\mathbf{D}_{test} = \{(y_8,x_8)^\top, \ldots, (y_{13},x_{13})^\top\}$. Funkce $f : x \to 3\arctan(x)$, $\varepsilon_i \sim \mathsf{N}(0,1)$ a $y_i = f(x_i) + \varepsilon_i$. Mějme dva modely, proložme tréninková data přímkou (lineární regrese) a proveďme interpolaci tréninkových dat pomocí Lagrangeova polynomu.

	Interpolace	Lin. regrese
$\text{MSE}_{training}$	0	6,3
MSE_{test}	$3{,}1 \times 10^6$	28,4
Var	$3{,}2 \times 10^5$	1,5

Tabulka 1.1: Srovnání modelu lineární regrese s interpolací Lagrangeovým polynomem. Hodnoty byly zaokrouhleny.

Modely jsou znázorněny na obrázku 1.3, ze kterého vidíme, že ačkoliv interpolace byla na tréninkových datech „bezchybná“, když došlo na predikce, byla

jednoznačně překonána modelem lineární regrese, to nám potvrzuje tabulka 1.1. S rostoucí flexibilitou metod učení totiž klesá $\text{MSE}_{training}$, což ovšem neznamená, že se sníží i MSE_{test}. Pokud zvolená metoda dosahuje nízkých hodnot $\text{MSE}_{training}$, ale vysokých hodnot MSE_{test}, nazveme tento jev *overfitting*. Problém je způsoben tím, že daná metoda nalézá v napozorovaných tréninkových datech vzory, které mohou vznikat pouze na základě náhodné veličiny ε, namísto hledané $f(\mathbf{X})$. Všimněme si také hodnot výběrového rozptylu obou modelů na náhodném výběru tvořeným testovacími daty.

Pokud ve větě 1 oslabíme předpoklady, dostaneme následující.

Věta 2 (Bias variance tradeoff)**.** *Nechť $\hat{f}(\boldsymbol{X})$ je náhodnou veličinou a odhadem $Y = f(\boldsymbol{X}) + \varepsilon$, $\boldsymbol{x}$ je pevnou realizací náhodného vektoru $\boldsymbol{X}$ a $\mathbb{V}ar\Big(f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})\Big)$ je konečný. Pak*

$$\mathbb{E}[(Y - \hat{Y})^2 | \boldsymbol{X} = \boldsymbol{x}] = \Big(Bias(\hat{f}|\boldsymbol{x})\Big)^2 + \mathbb{V}ar\Big(\hat{f}|\boldsymbol{x}\Big) + \mathbb{V}ar(\varepsilon),$$

kde $\mathbb{V}ar\Big(\hat{f}|\boldsymbol{x}\Big) = \mathbb{E}\Big[\Big(\hat{f}(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]\Big)^2\Big]$.

Důkaz. Rozepíšeme:

$$\begin{aligned}
\mathbb{E}[(Y - \hat{Y})^2 | \mathbf{X} = \mathbf{x}] &= \mathbb{E}[\Big(Y - \hat{f}(\mathbf{X})\Big)^2 | \mathbf{X} = \mathbf{x}] \\
&= \mathbb{E}[\Big(\varepsilon + f(\mathbf{x}) - \hat{f}(\mathbf{x})\Big)^2] \\
&= \mathbb{E}[\Big(f(\mathbf{x}) - \hat{f}(\mathbf{x})\Big)^2] + \mathbb{E}[\varepsilon^2].
\end{aligned}$$

Využijeme, že pro libovolnou náhodnou veličinu Z s konečným rozptylem platí

$$\mathbb{V}\text{ar}(Z) = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2.$$

Dostáváme tedy, že

$$\begin{aligned}
\mathbb{E}[(Y - \hat{Y})^2 | \mathbf{X} = \mathbf{x}] &= \Big(\mathbb{E}[f(\mathbf{x}) - \hat{f}(\mathbf{x})]\Big)^2 + \mathbb{V}\text{ar}\Big(f(\mathbf{x}) - \hat{f}(\mathbf{x})\Big) + \mathbb{V}\text{ar}(\varepsilon) \\
&= \Big(\text{Bias}(\hat{f}|\mathbf{x})\Big)^2 + \mathbb{V}\text{ar}\Big(\hat{f}|\mathbf{x}\Big) + \mathbb{V}\text{ar}(\varepsilon),
\end{aligned}$$

kde poslední rovnost plyne z definice 1 a toho že,

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x})] &= \mathbb{E}[f(\mathbf{x}) + \varepsilon - \varepsilon] = \mathbb{E}[Y | \mathbf{X} = \mathbf{x}] = f(\mathbf{x}), \\
\mathbb{V}\text{ar}\Big(f(\mathbf{x}) - \hat{f}(\mathbf{x})\Big) &= \mathbb{E}\Big[\Big(f(\mathbf{x}) - \hat{f}(\mathbf{x}) - \mathbb{E}[f(\mathbf{x}) - \hat{f}(\mathbf{x})]\Big)^2\Big] \\
&= \mathbb{E}\Big[\Big(f(\mathbf{x}) - \hat{f}(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})] + \mathbb{E}[\hat{f}(\mathbf{x})]\Big)^2\Big] \\
&= \mathbb{E}\Big[\Big(\hat{f}(\mathbf{x}) - \mathbb{E}[\hat{f}(\mathbf{x})]\Big)^2\Big] = \mathbb{V}\text{ar}\Big(\hat{f}|\mathbf{x}\Big),
\end{aligned}$$

což platí dle předpokladu (1.2).

□

Tato rovnost se dá nazvat kompromisem při hledání správné metody, která najde rovnováhu mezi dobrým proložením dat a nízkým rozptylem. Jak nám ukázal předchozí příklad, nebylo nijak obtížné najít funkci, která přesně kopírovala data ovšem za cenu vysokého rozptylu. Naopak lineární funkce, která sice dosahovala nízkého rozptylu, ale nemusí být vždy vhodnou volbou modelu, jak si ukážeme později.

1.2 Lineární regrese

Připomeňme si nejprve základní model, kterým je lineární regrese. Ta zkoumá vztah mezi spojitou veličinou Y a p-rozměrným vektorem $\mathbf{X}$ obsahujícím konečně mnoho spojitých nebo diskrétních veličin. Definice viz Kulich (2014, Kapitola 10), označení parametrů upraveno.

Definice 4. *Nechť chybové členy ε_i jsou nezávislé náhodné veličiny, pro něž platí $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{V}ar(\varepsilon_i) = \sigma^2$ (homoskedasticita), vektor $\boldsymbol{\beta} = (\beta_0,\beta_1,\cdots,\beta_p)^\top$ je vektor neznámých parametrů, kde $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$. Pro data $(Y_i, \boldsymbol{X}_i), i = 1, 2, \cdots, n$ a $\boldsymbol{X} \in \mathbb{R}^p$ splňující* lineární regresní model *platí*

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i. \tag{1.6}$$

Dle Kulich (2014, str. 88) můžeme definici 1.6 zapsat jako

$$\mathbb{E}[Y_i|\mathbf{X}_i] = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}, \tag{1.7}$$
$$\mathbb{V}\mathrm{ar}(Y_i|\mathbf{X}_i) = \sigma^2. \tag{1.8}$$

1.2.1 Vlastnosti

Vhodný odhad parametrů $\hat{\boldsymbol{\beta}}$, který minimalizuje euklidovskou vzdálenost $\hat{Y}_i$ od Y_i dostaneme *metodou nejmenších čtverců*. Rovnice (1.3) má pak tvar

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$$

pro data $(Y_i, \mathbf{X}_i), i = 1, 2, \ldots, n$. Metodu nejmenších čtverců můžeme použít, protože lineární regrese je lineární v hledaných parametrech.

Interpretace parametrů je v tomto případě přímočará a plyne z (1.7), kde pro $j = 1, 2, \cdots, p$ a $\mathbb{0} \in \mathbb{R}^p$ dostáváme

$$\begin{aligned}\beta_j &= \mathbb{E}[Y_i|X_{i1} = x_1, \cdots, X_{ij} = x_j + 1, \cdots, X_{ip} = x_p] - \\ &\quad - \mathbb{E}[Y_i|X_{i1} = x_1, \cdots, X_{ij} = x_j, \cdots, X_{ip} = x_p], \\ \beta_0 &= \mathbb{E}[Y_i|\mathbf{X}_i = \mathbb{0}],\end{aligned}$$

tedy parametr β_j udává změnu $\mathbb{E}[Y_i]$ při změně závisle proměnné X_{ij} o jedna, při ponechání zbylých regresorů konstantních. Parametr β_0 nazýváme absolutní člen, anglicky intercept.

Definice 5. *Nechť $\hat{Y}_i$ je nějaký odhad $Y_i, i = 1,2,\cdots,n$ a $\bar{Y} \equiv \frac{1}{n}\sum_{i=1}^n Y_i$, pak*

$$TSS \equiv \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad \ldots \textit{ nazveme celkový součet čtverců,}$$

$$RSS \equiv \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad \ldots \textit{ nazveme residuální součet čtverců,}$$

$$RSE \equiv \sqrt{\frac{RSS}{n-p-1}} \qquad \ldots \textit{ nazveme residuální standardní chyba (residual standard error).}$$

Pro $\varepsilon \sim \mathsf{N}_n(0,\sigma^2 I_n)$ je RSE nestranným a konsistentním odhadem σ viz věta 10.4 Kulich (2014, str. 93) a věta o spojité transformaci. Pro testování hypotéz o parametrech a jejich lineárních kombinací využíváme větu 10.5 Kulich (2014, str. 94).

Věta 3. *Nechť platí lineární model* (1.6)*, navíc pro chybový člen platí* $\varepsilon \sim N_n(0,\sigma^2 I_n)$ *a* $\boldsymbol{c} \in \mathbb{R}^{p+1}$*, pak*

$$\frac{\boldsymbol{c}^\top \hat{\boldsymbol{\beta}} - \boldsymbol{c}^\top \boldsymbol{\beta}}{\sqrt{\frac{RSS}{n-p-1}\boldsymbol{c}^\top (X^\top X)^{-1}\boldsymbol{c}}} \sim t_{n-p-1}.$$

Pro zjištění, zda má na model vliv jen určitá podmnožina prediktorů $\boldsymbol{\beta}$ využíváme *F-test* Kulich (2014, Věta 10.6). Bez újmy na obecnosti volíme testované regresory jako posledních $d \in \mathbb{N}_0 \wedge d \leq p+1$ prvků vektoru $\boldsymbol{\beta}$.

Věta 4. *Nechť platí lineární model* (1.6)*, kde* $\boldsymbol{X} = (\mathbb{1}|\boldsymbol{X}_A|\boldsymbol{X}_B)$*,* $\boldsymbol{\beta} = (\boldsymbol{\beta}_A^\top, \boldsymbol{\beta}_B^\top)^\top$*,* $\boldsymbol{\beta}_A \in \mathbb{R}^{p+1-d}, \boldsymbol{\beta}_B \in \mathbb{R}^d$*. Pak model lze zapsat jako*

$$Y = (\mathbb{1}|\boldsymbol{X}_A)\boldsymbol{\beta}_A + \boldsymbol{X}_B\boldsymbol{\beta}_B + \varepsilon.$$

Nechť navíc platí hypotéza $H_0 : \boldsymbol{\beta_B} = \mathbb{0}$*, pak*

$$F \equiv \frac{n-p-1}{d}\frac{RSS_0 - RSS}{RSS} \sim F_{d,n-p-1},$$

kde $RSS_0 = \sum_{i=1}^n (Y_i - (\mathbb{1}|\boldsymbol{X}_A)(\boldsymbol{\beta}_A^\top, \mathbb{0}^\top)^\top)^2$*, tedy RSS na „zúženém“ modelu. Symbol* $\mathbb{0}$ *zde označuje d-dimenzionální nulový vektor a* $\mathbb{1} = (1,\cdots,1)^\top \in \mathbb{R}^n$*.*

Definice 6 (Koeficient determinace)**.** *Veličinu*

$$R^2 \equiv 1 - \frac{RSS}{TSS}$$

nazveme koeficient determinace *nebo* R^2 statistika*.*

V modelu lineární regrese, kde koeficienty $\boldsymbol{\beta}$ byly odhadnuty metodou nejmenších čtverců, navíc platí

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\frac{1}{n-1}\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

Kde poslední výraz odpovídá podílu výběrového rozptylu odhadnuté odezvy v čitateli a výběrového rozptylu napozorovaných dat ve jmenovateli. Důkaz v William (2009, str. 507).

Všimněme si, že hodnota R^2 je vždy v intervalu [0,1]. Hodnoty blízké 1 naznačují dobré vysvětlení odezvy Y, hodnoty blízké nule napovídají, že lineární regrese dostatečně nevysvětluje variabilitu dat. To může být způsobeno tím, že je model lineární regrese pro daný problém nevhodný, že je rozptyl chybového členu σ^2 moc vysoký nebo obojí.

1.3 Klasifikace

Model lineární regrese je sice přímočarý a jednoduše interpretovatelný pro určitý typ problémů, jako je například hledání vztahu mezi marketingovým rozpočtem, jeho součástmi, regresory, a tržbami z prodeje, odezvou. Problémem, který se snažíme řešit v tomto textu, je ale posuzování „kvality“ žadatelů o úvěry, což nemusí nutně odpovídat lineárnímu modelu (1.6), obzvláště pokud se je budeme snažit rozřadit do skupin, např. každý klient je buď dobrý (akceptovatelný), nebo špatný (neúvěrovatelný), skupiny a jejich počet volíme podle aplikace.

Nechť $(Y_i, \mathbf{X}_i), i = 1, 2, \cdots, n$ a $\mathbf{X} \in \mathbb{R}^p$ jsou data a $Y_i \in \{1,2,\cdots,k\}, k \in \mathbb{N}$. Mezi hodnotami Y_i nemusí existovat smysluplné uspořádání, taková data nazýváme *kategoriální*.

Definice 7 (Chybovost)**.** *Nechť $\hat{Y}_i$ je odhadem Y_i, pro $i = 1, 2, \cdots, n$, pak*

$$Err \equiv \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(Y_i \neq \hat{Y}_i)}, \tag{1.9}$$

kde symbol $\mathbb{1}$ zde značí charakteristickou funkci.

Cílem je opět minimalizace chybovosti.

1.3.1 Logistická regrese

Uvažujme nyní, že množina skupin je dvouprvková, tedy $Y_i \in \{0,1\}$, pro $i = 1, 2, \ldots, n$. Mějme data o bankovních klientech, kterým byl v určitý okamžik přistaven úvěr. Rozdělme tyto klienty do věkových skupin a podívejme se, jak velká část které skupiny úspěšně splácí své závazky, tyto hodnoty najdeme v tabulce 1.2. Z této tabulky si všimněme, že s rostoucím věkem klesá relativní četnost špatných úvěrů s výjimkou skupiny seniorů, kterým jsou úvěry nabízeny a poskytovány pouze zřídka, to je ostatně vidět z nízkého počtu poskytnutých úvěrů v této skupině. Hodnoty z tabulky 1.2 vidíme na obrázku 1.4, kde byla invertována osa x z důvodu jasnější interpretace. Vidíme totiž, že by křivka spojující relativní četnosti mohla vzdáleně připomínat zatím neznámou distribuční funkci nějaké náhodné veličiny, na našem data setu to není tak evidentní, jako kdybychom analyzovali např. výskyt onemocnění pro různé věkové skupiny viz David W. Hosmer a kol. (2013, Figure 1.2).

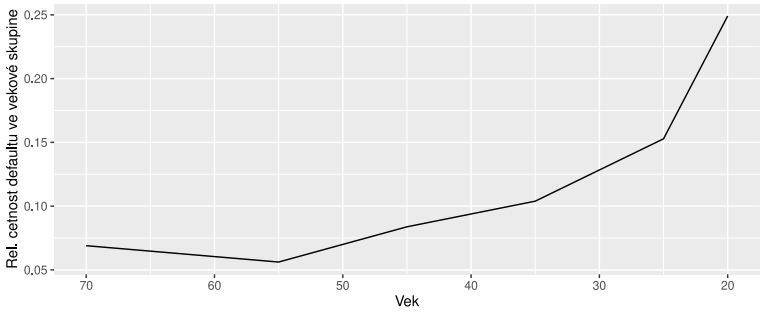
Uvažujme nyní, že relativní četnosti z tabulky 1.2 jsou odhadem pro $\mathbb{E}[Y|\mathbf{X}]$, kde Y je *dichotomická* závisle proměnná a regresor X je věk. Zde je evidentně model (1.7) nevhodný, protože pro libovolné X může nabývat libovolné hodnoty mezi $-\infty$ a $+\infty$, zároveň ale jistě $\forall \mathbf{x} \in \mathbb{R}^p : 0 \leq \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] \leq 1$. Vhodnou funkcí pro tuto aplikaci je logistická křivka, jejíž parametry jsou, jak uvidíme, jednoduše interpretovatelné.

Značení přebíráme z David W. Hosmer a kol. (2013, Kapitola 2.2). Pro zjednodušení budeme nyní označovat podmíněnou střední hodnotu odezvy jako:

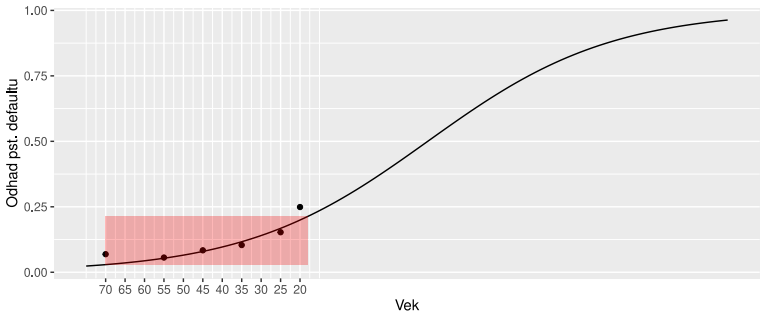
$$\pi(\mathbf{X}) \equiv \mathbb{E}[Y|\mathbf{X}].$$

Věk. skupina	n	OK	Default	Rel. četnost deafultů
18-23	823	618	205	0,332
23-30	2140	1813	327	0,180
30-40	2647	2372	275	0,116
40-50	1910	1750	160	0,091
50-65	1122	1059	63	0,059
65+	261	243	18	0,074
Celkem	8903	7855	1048	0,133

Tabulka 1.2: Default rate podle věkové skupiny. Intervaly jsou zleva uzavřené a zprava otevřené.



(a) Empirické relativní četnosti po věk. skupinách.



(b) Odhady defaultu pomocí log. regrese.

Obrázek 1.4: Odhad je založen na logistické regresi s parametry $\beta_0 = -0{,}544$ a $\beta_1 = 0{,}0423$, červeně je zvýrazněna oblast, která by nás v rámci analýzy zajímala.

Definice 8. *Nechť chybové členy ε_i jsou nezávislé náhodné veličiny, vektor $\boldsymbol{\beta} = (\beta_0,\beta_1,\ldots,\beta_p)^T$ je vektor neznámých parametrů, kde $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$. Pro data $(Y_i, \boldsymbol{X}_i)$, $i = 1, 2, \ldots, n$, $\boldsymbol{X}_i \in \mathbb{R}^p$, $Y_i \in \{0,1\}$ splňující* model logistické regrese *platí*

$$Y_i = \frac{e^{\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip}}}{1 + e^{\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip}}} + \varepsilon_i. \tag{1.10}$$

Ekvivalentně:

$$\pi(\boldsymbol{X}) = \frac{e^{\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip}}}{1 + e^{\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip}}}. \tag{1.11}$$

Za použití *logit transformace*

$$g(\mathbf{X}) = \ln\left(\frac{\pi(\mathbf{X})}{1-\pi(\mathbf{X})}\right), \tag{1.12}$$

dostáváme, že

$$g(\mathbf{X}) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}. \tag{1.13}$$

Odvoďme vlastnosti chybového členu ε_i v definici 8. Význam tohoto chybového členu je evidentně jiný než v chybového členu vystupujícího v (1.6), tato náhodná veličina nemůže nabývat jakékoliv hodnoty. Mohou nastat pouze dva případy, vyjádříme z $Y_i = \pi(\mathbf{X}_i) + \varepsilon_i$,

$$Y_i = 1 \qquad\qquad Y_i = 0.$$

Tedy pro chybový člen platí buď a nebo

$$\varepsilon_i = 1 - \pi(\mathbf{X}_i) \qquad\qquad \varepsilon_i = -\pi(\mathbf{X}_i).$$

Protože navíc:

$$\pi(\mathbf{X}_i) = \mathbb{E}[Y_i|X_i] = 1P[Y_i = 1|X_i] + 0P[Y_i = 1|X_i] = P[Y_i = 1|X_i],$$

z čehož vidíme, že

$$\begin{aligned} P[\varepsilon_i = 1 - \pi(\mathbf{X}_i)] &= \pi(\mathbf{X}_i), \\ P[\varepsilon_i = -\pi(\mathbf{X}_i)] &= 1 - \pi(\mathbf{X}_i). \end{aligned}$$

Nyní snadno dopočítáme momenty

$$\begin{aligned} \mathbb{E}[\varepsilon_i] &= [1-\pi(\mathbf{X}_i)]\pi(\mathbf{X}_i) + (-1)\pi(\mathbf{X}_i)[1-\pi(\mathbf{X}_i)] = 0, \\ \mathbb{V}\mathrm{ar}(\varepsilon_i) &= [1-\pi(\mathbf{X}_i)]^2\pi(\mathbf{X}_i) + \Big(-\pi(\mathbf{X}_i)\Big)^2[1-\pi(\mathbf{X}_i)] = \pi(\mathbf{X})[1-\pi(\mathbf{X})]. \end{aligned} \tag{1.14}$$

Odhad parametrů

Pro odhad parametrů v modelu logistické regrese se používá metody maximální věrohodnosti. Předpokládejme, že máme náhodný výběr o rozsahu n, hledáme $\arg\max_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} L_n(\boldsymbol{\beta})$, kde *věrohodnostní funkce* má tvar podle David W. Hosmer a kol. (2013, (1.3))

$$L_n(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\mathbf{X}_i)^{Y_i}[1-\pi(\mathbf{X}_i)]^{1-Y_i},$$

logaritmická věrohodnost je tedy rovna

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \ln(\pi(\mathbf{X}_i)) + \sum_{i=1}^{n}(1-Y_i)\ln(1-\pi(\mathbf{X}_i)). \tag{1.15}$$

Hledáme takový vektor $\hat{\boldsymbol{\beta}}$, že $\nabla l_n(\hat{\boldsymbol{\beta}}) = \mathbb{0}$. Budeme tedy parciálně derivovat podle jednotlivých složek vektoru $\hat{\boldsymbol{\beta}}$. Položíme $X_{i0} = 1$, pro i-tý sčítanec (1.15),

$j = 0,1,\ldots,p$ dostáváme

$$\pi'(\mathbf{X}_i) \equiv \frac{\partial \pi(\mathbf{X}_i)}{\partial \beta_j} = X_{ij} \frac{\pi(\mathbf{X}_i)}{1 + e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}, \tag{1.16}$$

$$\begin{aligned} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j} &= Y_i \frac{\pi'(\mathbf{X}_i)}{\pi(\mathbf{X}_i)} + (1 - Y_i) \frac{-\pi'(\mathbf{X}_i)}{1 - \pi(\mathbf{X}_i)} \\ &= \frac{Y_i \pi'(\mathbf{X}_i) - Y_i \pi'(\mathbf{X}_i)\pi(\mathbf{X}_i) - \pi'(\mathbf{X}_i)\pi(\mathbf{X}_i) + Y_i \pi'(\mathbf{X}_i)\pi(\mathbf{X}_i)}{\pi(\mathbf{X}_i)[1 - \pi(\mathbf{X}_i)]} \\ &= \frac{Y_i \pi'(\mathbf{X}_i) - \pi'(\mathbf{X}_i)\pi(\mathbf{X}_i)}{\pi(\mathbf{X}_i)[1 - \pi(\mathbf{X}_i)]} = \frac{\pi'(\mathbf{X}_i)[Y_i - \pi(\mathbf{X}_i)]}{\pi(\mathbf{X}_i)[1 - \pi(\mathbf{X}_i)]}. \end{aligned} \tag{1.17}$$

Po dosazení (1.16) do (1.17) dostáváme, že

$$\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j} = 0 \Leftrightarrow \sum_{i=1}^{n} X_{ij}[(Y_i) - \pi(\mathbf{X}_i)] = 0. \tag{1.18}$$

Z (1.18) pro $\beta_j = \beta_0$ dostáváme, že od modelu budeme požadovat, aby

$$\sum_{i=1}^{n} Y_i = \hat{\pi}(\mathbf{X}_i).$$

Protože je výraz (1.18) nelineární v parametrech $\beta_j, j = 0,1,\ldots,p$, není možné získat $\hat{\boldsymbol{\beta}}$ explicitní vyjádřením a je nutné využít numerických metod. Ve statistickém jazyce R je logistická regrese implementována pod

```
glm(formula, family = binomial, data, method = "glm.fit", ...)
```

Výchozí metoda této implementace, glm.fit, používá *iterativně převažovanou metodu nejmenších čtverců* (Iteratively reweighted least squares – IRLS), nebude dále rozebíráno, více k odhadům parametrů např. v textu Zvára (2019, kap. 12.2) nebo v Fahrmeir a kol. (2013, kap. 5.1.2).

Interpretace parametrů

Poznamenejme nejprve, že obecně, než se budeme pokoušet jakýkoliv model interpretovat, bychom měli již mít vhodně vyhodnocenou jeho významnost.

Definice 9 (Poměr šancí (Odds ratio))**.** *(David W. Hosmer a kol., 2013, (3.1)) Nechť* $\boldsymbol{X}_a = (X_1,\ldots,a,\ldots,X_p)$,$\boldsymbol{X}_b = (X_1,\ldots,b,\ldots,X_p)\boldsymbol{\beta} \in \mathbb{R}^{p+1}$,$a,b \in \mathbb{R}$ *jsou hodnoty kovariátu* X_j*.Pak výraz*

$$OR(a,b) \equiv \frac{\frac{\pi(\boldsymbol{X}_a)}{1-\pi(\boldsymbol{X}_a)}}{\frac{\pi(\boldsymbol{X}_b)}{1-\pi(\boldsymbol{X}_b)}} \tag{1.19}$$

nazýváme poměr šancí.

Poměr šancí (1.19) lze zjednodušit na

$$\begin{aligned} OR(a,b) &= \frac{\pi(\mathbf{X}_a)}{1-\pi(\mathbf{X}_a)}\frac{1-\pi(\mathbf{X}_b)}{\pi(\mathbf{X}_b)} \\ &= \frac{\frac{e^{\beta_0+\cdots+\beta_j a+\cdots+\beta_p X_p}}{1+e^{\beta_0+\cdots+\beta_j a+\cdots+\beta_p X_p}}}{\frac{1}{1+e^{\beta_0+\cdots+\beta_j a+\cdots+\beta_p X_p}}}\frac{\frac{1}{1+e^{\beta_0+\cdots+\beta_j b+\cdots+\beta_p X_p}}}{\frac{e^{\beta_0+\cdots+\beta_j b+\cdots+\beta_p X_p}}{1+e^{\beta_0+\cdots+\beta_j b+\cdots+\beta_p X_p}}} \\ &= \frac{e^{\beta_0+\cdots+\beta_j a+\cdots+\beta_p X_p}}{e^{\beta_0+\cdots+\beta_j b+\cdots+\beta_p X_p}} \\ &= e^{\beta_j(a-b)}. \end{aligned}$$

Pokud je kovariát X_j dichotomický, dostáváme, že $a = 1$ a $b = 0$, tedy že

$$OR(1,0) = e^{\beta_j}.$$

Interpretujeme, že subjekt s příznakem $X_j = 1$ má e^{β_j}-krát větší šanci na výskyt výstupu Y než subjekt bez příznaku ($X_j = 0$).

Pokud je kovariát X_j spojitý, budeme zkoumat jeho nárůst o předem určenou jednotku $c \in \mathbb{R}$. Máme tedy, že $a = x + c, b = x$, kde $x \in \mathbb{R}$, pak

$$OR(x + c,x) = e^{c\beta_j}. \tag{1.20}$$

Vhodné a snadno pochopitelné hodnoty c pro spojité změny kovariátů jsou kupříkladu násobky 2, 5 nebo 10 (výška, věk aj.).

Statistická významnost modelu a podmodelů

Jakmile máme odhad $\hat{\boldsymbol{\beta}}$, je důležité, abychom posoudili jeho kvalitu. Podmodely je možné testovat na základě dále popsaných testových statistik, které mají stejná asymptotická rozdělení.

Definice 10 (Deviance)**.** *(Zvára, 2019, (12.12)) Uvažujme nejbohatší možný model, který má právě tolik parametrů jako je různých hodnot vektoru $\boldsymbol{X}_i$. Takový model nazveme* saturovaný*. Označme l_{max} maximální hodnotu věrohodnostní funkce v saturovaném modelu, pak*

$$D(\boldsymbol{\beta}) \equiv 2(l_{max} - l_n(\boldsymbol{\beta}) \tag{1.21}$$

nazýváme deviance.

Následující požadavky a věta dle Zvára (2019, A.3). Nechť $\hat{\boldsymbol{\beta}} \in \Omega$ je maximálně věrohodným odhadem vektoru $\boldsymbol{\beta}$, $\tilde{\boldsymbol{\beta}} \in \omega \subset \Omega$, vektor $\tilde{\boldsymbol{\beta}}$ je tedy vektorem parametrů v podmodelu. Původně byl tento test navržen v článku Wilks (1938).

Věta 5 (Test věrohodnostním poměrem (Wilksův test))**.** *Předpokládejme, že jsou splněny podmínky regularity, pak*

$$G \equiv D(\tilde{\boldsymbol{\beta}}) - D(\hat{\boldsymbol{\beta}}). \tag{1.22}$$

Za platnosti nulové hypotézy H_0 : platí podmodel

$$G \xrightarrow{D} \chi^2_q,$$

kde q je rozdíl dimenze prostorů Ω a ω (rozdíl počtu parametrů).

Pokud tedy na hladině α zamítáme nulovou hypotézu, tak usuzujeme, že alespoň jeden z testovaných vynechaných koeficientů je nenulový.

Věta 6 (Waldův test)**.** *(David W. Hosmer a kol., 2013, str. 42)*

$$W \equiv \hat{\boldsymbol{\beta}}^{\top}[\widehat{\mathbb{V}ar}(\hat{\boldsymbol{\beta}})]^{-1}\hat{\boldsymbol{\beta}},$$

kde odhad rozptylu je na základě nám známé Fischerovy informační matice. *Více v Zvára (2019, A.3).*

Za platnosti nulové hypotézy H_0*:* $\boldsymbol{\beta} = \mathbb{0} \in \mathbb{R}^q$ *platí*

$$W \xrightarrow{D} \chi^2_q.$$

Speciálně pro jednu proměnnou

$$W = \frac{\hat{\beta}_i}{\hat{\sigma}(\hat{\beta}_i)} \xrightarrow{D} \chi^2_1 = \mathsf{N}(0,1).$$

2. Tvorba modelu

2.1 Typy proměnných

Než se pustíme do tvorby modelu, věnujme pár slov typům regresorů, na které můžeme narazit. Nejjednodušší dělení je na nominální (kvantitativní) a kategoriální (kvalitativní). V nominálních datech, pod kterými si představme například měsíční obrat na účtě, existuje uspořádání. Pro kategoriální proměnnou takové uspořádání existovat nemusí, mezi takové proměnné patří např. pohlaví, národnost a další. Uspořádání může být eticky nevhodné či na určité úrovni zakázané jako např. jednotné tarify životních pojištění žen a mužů v EU. Abychom se řazení vyhnuli, zavádíme *designové* nebo tzv. *dummy* proměnné.

Prediktor se dvěma úrovněmi

Uvažujme nyní nezávislou proměnnou X_i definovanou jako

$$X_i = \begin{cases} 0, & \text{je-li } i\text{-tý subjekt žena,} \\ 1, & \text{je-li } i\text{-tý subjekt muž.} \end{cases}$$

Dummy proměnná nabývá pouze dvou možných hodnot, pokud dále uvažujeme, že platí model

$$g(X_i) = \beta_0 + \beta_1 X_i,$$

speciálně pro lineární regresi

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

pak nastávají pouze tyto dva případy:

$$g(X_i) = \begin{cases} \beta_0, & \text{je-li } i\text{-tý subjekt žena,} \\ \beta_0 + \beta_1, & \text{je-li } i\text{-tý subjekt muž.} \end{cases}$$

Uvažujme, že jde o logistickou regresi, kde $g(X)$ je logit funkcí. Zde interpretujeme β_0 jako průměrný logaritmus šance defaultu pro ženy, $\beta_0 + \beta_1$ je průměrný logaritmus šance defaultu pro muže a β_1 je průměrný rozdíl logaritmů šance mezi ženami a muži.

Pokud bychom ale uvažovali kódování nezávisle proměnné jako

$$X_i = \begin{cases} -1, & \text{je-li } i\text{-tý subjekt žena,} \\ 1, & \text{je-li } i\text{-tý subjekt muž,} \end{cases}$$

dostali bychom výraz

$$g(X_i) = \begin{cases} \beta_0 - \beta_1, & \text{je-li } i\text{-tý subjekt žena,} \\ \beta_0 + \beta_1, & \text{je-li } i\text{-tý subjekt muž.} \end{cases}$$

Při tomto kódování interpretujeme β_0 jako průměrný logaritmus šance v celém náhodném výběru a β_1 je rozdíl logaritmů šance mužů a žen od průměru. Pro náš data set, kde byly proměnné pohlaví kódovány v souladu s právě popsaným, dostáváme koeficienty viz tabulka 2.1. Důležité je, že predikce jsou pro oba tyto případy totožné, rozdílná je pouze interpretace proměnných.

Kódování	$\hat{\beta}_0$	$\hat{\beta}_1$
0/1	−2,231	0,332
-1/1	−2,065	0,166

Tabulka 2.1: Porovnání odhadu koeficientů pro různé kódování kategoriální proměnné.

Národnost	baseline	X_{i1}	X_{i2}	X_{i3}	X_{i4}	X_{i5}
ČR	1	0	0	0	0	0
EU	1	1	0	0	0	0
mimo EU	1	0	1	0	0	0
PL	1	0	0	1	0	0
SR	1	0	0	0	1	0
UA	1	0	0	0	0	1

Tabulka 2.2: Designová matice pro národnosti.

Prediktor s více úrovněmi

Uvažujeme nyní, že proměnná státní příslušnost X_i nabývá šesti hodnot: ČR, SR, PL, Mimo EU (kromě UA), UA, EU (jiné než PL, SR, ČR) – dále bude označováno pouze jako EU. Jedna dummy proměnná v tomto případě nemůže reprezentovat všechny možné hodnoty, proto nadefinujeme celkem pět proměnných

$$X_{i1} = \begin{cases} 1, & \text{je-li státní příslušnost } i\text{-tého subjektu EU,} \\ 0, & \text{není-li státní příslušnost } i\text{-tého subjektu EU,} \end{cases}$$

$$X_{i2} = \begin{cases} 1, & \text{je-li státní příslušnost } i\text{-tého subjektu mimo EU,} \\ 0, & \text{není-li státní příslušnost } i\text{-tého subjektu mimo EU,} \end{cases}$$

$$X_{i3} = \begin{cases} 1, & \text{je-li státní příslušnost } i\text{-tého subjektu PL,} \\ 0, & \text{není-li státní příslušnost } i\text{-tého subjektu PL,} \end{cases}$$

$$X_{i4} = \begin{cases} 1, & \text{je-li státní příslušnost } i\text{-tého subjektu SR,} \\ 0, & \text{není-li státní příslušnost } i\text{-tého subjektu SR,} \end{cases}$$

$$X_{i5} = \begin{cases} 1, & \text{je-li státní příslušnost } i\text{-tého subjektu UA,} \\ 0, & \text{není-li státní příslušnost } i\text{-tého subjektu UA.} \end{cases}$$

Jiný způsob definice je tzv. *designová matice*, která by pro tento případ vypadala jako 2.2. β_0 se nazývá *baseline*. Takto definované proměnné pak můžeme použít

Národnost	n	Koeficient	Odhad koeficientu	p − hodnota
ČR	7883	β_0	−2,15	0,00
EU	64	β_1	1,13	0,00
mimo EU	101	β_2	−0,18	0,62
PL	47	β_3	1,29	0,00
SR	716	β_4	0,98	0,00
UA	92	β_5	0,59	0,03

Tabulka 2.3: Koeficienty logistické regrese na základě státní příslušnosti.

pro tvorbu regresního modelu

$$g(X_i) = \begin{cases} \beta_0 + \beta_1, & \text{je-li státní příslušnost EU,} \\ \beta_0 + \beta_2, & \text{je-li státní příslušnost mimo EU,} \\ \beta_0 + \beta_3, & \text{je-li státní příslušnost PL,} \\ \beta_0 + \beta_4, & \text{je-li státní příslušnost SR,} \\ \beta_0 + \beta_5, & \text{je-li státní příslušnost UA,} \\ \beta_0, & \text{je-li státní příslušnost ČR.} \end{cases}$$

Parametr β_0 zde interpretujeme jako průměrnou logaritmickou šanci defaultu pro Čecha, zbylé parametry $\beta_j, j = 1, \ldots, 5$ jako rozdíl logaritmů šance mezi ČR a j-tou kategorií.

V tabulce 2.3 vidíme výsledné odhady parametrů. Z tabulky na základě p-hodnot vidíme, že všechny koeficienty až na β_2 jsou statisticky významné, p-hodnota je zde počítána na základě Waldova testu pro jednu proměnnou, nemůžeme ale vyloučit, že kategorie „mimo EU“ má stejný logaritmus šance jako kategorie „ČR“.

Otestujme ještě hypotézu $H_0 : \beta_2 = \beta_5 = 0$ pomocí Waldova testu. Dostáváme hodnotu testové statistiky $W = 19{,}999$ a p-hodnotu $4{,}542 \times 10^{-5}$. Hypotézu H_0 tedy zamítáme na hladině $\alpha = 0{,}05$ ve prospěch alternativy H_1 : alespoň jeden z koeficientů β_2,β_5 je nenulový.

2.2 Strategie tvorby modelu

Statistické testy, které byly popsány v dřívějších kapitolách, nám umožňují určit, které ze závisle proměnných jsou významné. Jak tedy určit a vyloučit ty regresory, které danému modelu nijak neprospívají, a jak vybrat mezi rozdílnými obdobně schopnými modely? Pro porovnání modelů budeme využívat některou z následujících statistik, od nich požadujeme, aby braly v potaz nejen kvalitu fitu, ale nějakým způsobem adresovaly problém overfittingu.

Definice 11 (AIC)**.** *(Zvára, 2019, kap 10.1.7) Nechť $p + 1$ je počet složek maximálně věrohodného odhadu vektoru $\hat{\boldsymbol{\beta}}$ a $l_n(\hat{\boldsymbol{\beta}})$ je logaritmická věrohodnostní funkce, pak*

$$AIC \equiv -2l_n(\hat{\boldsymbol{\beta}}) + 2(p + 1) \tag{2.1}$$

se nazývá Akaikeho informační kritérium.

Vybavíme-li si definici 6, vidíme, že klasická definice R^2 statistiky nebude pro srovnávaní modelů vhodná, protože nebere v potaz počet použitých proměnných a vždy upřednostní bližší odhad. Zároveň formulace není vhodná pro nelineární model.

Definice 12 (Zobecněný koeficient determinace)**.** *Cox a Snell (1989) Nechť n je rozsah náhodného výběru, $l_n(\mathbb{0})$ je věrohodnost nulového modelu (obsahujícího pouze intercept) a $l_n(\boldsymbol{\beta})$ je věrohodnost modelu s vektorem parametrů $\boldsymbol{\beta}$. Veličinu*

$$R^2_{gen} \equiv 1 - \left(\frac{l_n(\mathbb{0})}{l_n(\boldsymbol{\beta})}\right)^{\frac{2}{n}}, \tag{2.2}$$

nazveme zobecněný koeficient determinace.

Existují i jiné definice např. od Nagelkerke (1991).

Validační množiny a křížová validace

V oddílu 1.1.1 jsme se zabývali rozdílem mezi testovou chybou a tréninkovou chybou. Připomeňme, že chceme co nejnižší testovou chybu. Alternativním způsobem k porovnání modelů je odhadování testové chyby, to ale vzhledem k tomu, že ve většině případů nebudeme mít testová data k dispozici, vyžaduje specifický přístup. Myšlenka spočívá v náhodném rozdělení dostupných dat na tréninkovou a validační množinu.

Bohužel náhodné dělení dat s sebou přináší možnou variabilitu výsledků. Učení na datech o menším rozsahu může mít za výsledek horší kvalitu modelu, tedy může dojít k nadhodnocování testové chyby. Tento problém napravuje *LOOCV validace* (Leave-One-Out-CV)(James a kol., 2013, kapitola 5.1). Ta spočívá ve vynechání právě jednoho pozorování (Y_i,$\mathbf{X}_i$) a učení modelu na zbylé množině. Průměrnou testovou chybu spočteme postupným vynecháním všech pozorování. Pro klasifikační problém vypadá odhad testové chyby jako

$$CV_{(n)} \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(Y_i \neq \hat{Y}_i)}. \tag{2.3}$$

Pro rozsáhlé náhodné výběry je tento způsob výpočetně náročný, protože jsou parametry modelu odhadovány n-krát, proto jako kompromis mezi LOOCV a dělením na dvě náhodné množiny zavádíme k-násobnou LOOCV, kde data $\mathbf{D}$ náhodně rozdělíme na k disjunktních množin $\mathbf{D}_{test,j}$,$j = 1,\ldots,k$ s alespoň přibližně stejným počtem prvků. LOOCV je pak speciálním případem, kdy $k = n$. Nechť Err_j značí chybovost, viz definice 7, na j-té testové podmnožině $\mathbf{D}_{test,j}$, pak odhad testové chyby pomocí k-násobné LOOCV získáme jako

$$CV_{(k)} \equiv \frac{1}{k}\sum_{j=1}^{k} \mathrm{Err}_j. \tag{2.4}$$

Pro lineární regresi bychom v definicích použili MSE namísto Err_j pro danou podmnožinu dat.

Residua v binární logistické regresi

Z lineární regrese víme, že pro měření kvality modelu bývají používány residua, ty ale v případě nelineárních modelů nemusí jít jednoduše využít k jejich analýze. Podle David W. Hosmer a kol. (2013, 5.2.1) zavádíme následující značení. Nechť $\mathbf{X}_j$,$j = 1,2,\ldots,J$ značí jednotlivé různé konfigurace kovariátu $\mathbf{X}$, pokud se pro některá pozorování hodnoty shodují, pak $J < n$. Označme m_j, $j = 1,2,\ldots,J$ počet subjektů s konfigurací $\mathbf{X}_j$. Nechť dále y_j je počet m_j, pro něž $y = 1$. Podotkněme, že je časté, že $J \approx n$.

Pro j-tou konfiguraci dostáváme

$$\hat{y}_j = m_j \hat{\pi}(\mathbf{X}_j), \tag{2.5}$$

kde $\hat{\pi}$ je odhadem v souladu s (1.11). Pro konkrétní konfiguraci kovariátu $\mathbf{X}$ definujeme *Pearsonovo* residuum jako

$$r(y_j,\hat{\pi}(\mathbf{X}_j)) \equiv \frac{y_j - m_j\hat{\pi}(\mathbf{X}_j)}{\sqrt{m_j\hat{\pi}(\mathbf{X}_j)(1-\hat{\pi}(\mathbf{X}_j))}}. \tag{2.6}$$

Dále za vyžití zápisu (2.5) definujeme *devianční* residuum jako

$$d(y_j,\hat{\pi}(\mathbf{X}_j)) \equiv sgn(y_j - \hat{y}_j)\left\{2\left[y_j \ln\left(\frac{y_j}{\hat{y}_j}\right) + (m_j - y_j)\ln\left(\frac{m_j - y_j}{m_j - \hat{y}_j}\right)\right]\right\}^{\frac{1}{2}}. \tag{2.7}$$

Na základě těchto residuí je založen Pearsonův Chí-kvadrát test a D test, které jsou zadefinovány v David W. Hosmer a kol. (2013, (5.2),(5.4)).

Residua jsou často vyobrazována v podobě grafů, existuje několik možných přístupů s cílem odhalit nenáhodné chování. Residua můžeme vyobrazit proti jednotlivým regresorům nebo například vůči času.

2.2.1 Výběr proměnných

Mějme nyní p prediktorů, z nichž budeme vybírat nejlepší kombinace, ve smyslu vylepšování statistik (2.1) nebo (2.2). Počet takových kombinací je 2^p, tedy potenciálně velmi vysoký. Proto jsou pro výběr vhodné podmnožiny regresorů z množiny všech jejich kombinací využívány zjednodušující postupy. Výběr může probíhat například podle dále popsaných dvou postupů a častěji jejich kombinací (James a kol., 2013, kap. 6.1.2). Takto vzniklé modely ovšem nemusí být těmi nejlepšími ze všech možných.

Vzestupný výběr

1. Nechť M_0 je *nulový model* neobsahující žádný z prediktorů.
2. Pro $k = 0, \ldots, p-1$:
   - najdeme všech $p-k$ modelů, které do modelu M_k přidávají jeden další prediktor,
   - z těchto modelů vybereme nejlepší a označíme ho M_{k+1}. Nejlepší model je ten s nejnižší hodnotou AIC nebo R^2_{gen}.
3. Z množiny modelů $M_0, \ldots, M_p$ vybereme celkově nejlepší podle odhadu testové chyby křížovou validací, AIC, R^2_{gen} nebo porovnáním AUC.

Namísto 2^p modelů zde uvažujeme pouze $1+\sum_{k=0}^{p-1}(p-k) = 1+p(p+1)/2$ modelů, což je značné ulehčení.

Sestupný výběr

1. Nechť M_p je *úplný model* obsahující všechny prediktory.
2. Pro $k = p, p-1, \ldots, 1$:
   - najdeme všech k modelů, které z modelu M_k odebírají jeden prediktor,
   - z těchto modelů vybereme nejlepší a označíme ho M_{k-1}. Nejlepší model je ten s nejnižší hodnotou AIC nebo R^2_{gen}.
3. Z množiny modelů $M_0, \ldots, M_p$ vybereme celkově nejlepší podle odhadu testové chyby křížovou validací, AIC, R^2_{gen} nebo porovnáním AUC.

Opět uvažujeme pouze $1 + p(p+1)/2$ modelů.

2.3 Diskriminační schopnost modelu

Binární klasifikační model může udělat pouze dva typy chyb. Pokud se budeme již dále bavit o modelech defaultu daného klienta, tak buď klienta, který zdefaultuje, přiřadí do kategorie řádně splácejících, nebo klienta, který by splácel, přiřadí do kategorie defaultujících. Pro nás bude důležité mezi těmito chybami rozlišovat, proto zavedeme tzv. *confusion matrix*. V souvislosti s tímto zavádíme pojmy *senzitivita* a *specificita* a další, které si vysvětlíme na tabulce 2.4 a v následující definici.

Definice 13. *Pozitivní chápejme jako hodnotu* 1*, má vlastnost, negativní chápejme jako hodnotu* 0*, nemá vlastnost, pak definujeme tyto pojmy:*

- *P – Condition positive – pozitivní z celého náhodného výběru,*
- *N – Condition negative – negativní z celého náhodného výběru,*
- *PP – Predicted condition positive – predikovaní pozitivní,*

		Reálný status		
		0	1	celkem
Predikovaný status	0	TN	FN	PN
	1	FP	TP	PP
	celkem	N	P	n

Tabulka 2.4: Confusion matrix pro binární klasifikátor obecně.

- *PN – Predicted condition negative – predikovaní negativní,*
- *TP – True positive – pozitivní z predikovaných pozitivních,*
- *TN – True negative – negativní z predikovaných negativních,*
- *FP – False positive – negativní z predikovaných pozitivních,*
- *FN – False negative – pozitivní z predikovaných negativních.*

Na základě těchto pojmů pak

$$TPR \equiv \frac{TP}{P} \textit{nazveme} \text{ senzitivita},$$
$$TNR \equiv \frac{TN}{N} \textit{nazveme} \text{ specificita}.$$

Prahem *rozhodnutí nazveme* $treshold \in (0,1)$*, jestliže klasifikační model rozhoduje podle*

$$\hat{Y}_i = \begin{cases} 1, & \textit{je-li } P[Y_i = 1|\boldsymbol{X}_i] > treshold, \\ 0, & \textit{jinak.} \end{cases}$$

V dalším textu budou využívány převážně zkratky těchto pojmů. Pro model, jehož výstup vidíme v tabulce 2.5, spočteme tyto hodnoty jednoduše. Senzitivita vychází 216/1048, což je pouze 20,1%. Specificita je 7152/7855, tedy 91%.

Tabulky jako 2.5 jsou založeny pouze na jedné hodnotě *treshold*. Lepší by bylo uvažovat těchto hodnot více. Graficky se různé hodnoty typicky vyobrazují pomocí tzv. *ROC křivky*, jejíž název pochází z teorie detekce signálu (Receiver Operating Characteristic), a AUC – plochou pod ROC křivkou (David W. Hosmer a kol., 2013, kap 5.2.4). Graf vyobrazuje pravděpodobnost detekce TP, tedy TPR, proti pravděpodobnost detekce FP, tedy $1 - \text{TNR} = 1 - spcificita$ pro různé hodnoty prahu. Plocha pod ROC křivkou je z intervalu [0,5 , 1] a měří schopnost

		Reálný status defaultu		
		OK	Default	celkem
Predikovaný status defaultu	OK	7152	832	7984
	Default	703	216	919
	celkem	7855	1048	8903

Tabulka 2.5: Confusion matrix pro model 2.3 a práh 0,11.

modelu diskriminovat mezi subjekty, kteří vykazují zkoumanou vlastnost a těmi, kteří ne. Podle téže kapitoly David W. Hosmer a kol. (2013) vyhodnocujeme AUC takto:

$$\text{Pokud} = \begin{cases} \text{AUC} = 0{,}5, & \text{diskriminace je na úrovni hodu mincí,} \\ 0{,}5 < \text{AUC} < 0{,}7, & \text{diskriminační schopnost je spíše nízká,} \\ 0{,}7 \leq \text{AUC} < 0{,}8, & \text{diskriminační schopnost je akceptovatelná,} \\ 0{,}8 \leq \text{AUC} < 0{,}9, & \text{diskriminační schopnost je dobrá,} \\ \text{AUC} \geq 0{,}9, & \text{diskriminační schopnost je vynikající.} \end{cases} \tag{2.8}$$

V programovacím jazyce R v aplikační části budeme pro znázornění ROC křivky používat knihovnu Robin a kol. (2011).

3. Aplikace

S vybudovaným aparátem se nyní můžeme pustit do již kompletní analýzy data setu a nikoliv pouze samostatných částí, jak bylo v rámci demonstrativních příkladů ukázáno v předchozích kapitolách. Budeme vyšetřovat skutečnou sedmiletou historii úvěrů poskytnutých jednou z českých bank a sledovat závislosti mezi informacemi, které jsou bankou o klientech uchovávány, dvěma indikátory zveřejňovanými Českým statistickým úřadem, konkrétně kvartální nezaměstnanost a kvartální průměrná mzda, velikostmi okresů v ČR podle podle počtu obyvatel ze sčítání lidu v roce 2011, a tím, zda byl klient schopen řádně a bez prodlení splácet závazky plynoucí z poskytnutých úvěrů, což uvažujeme, že je dichotomická závisle proměnná.

3.1 Popis datového souboru

K analýze máme k data o rozsahu $n = 8903$, nejstarší ze záznamů je z 1.1.2012 a nejnovější z 30.3.2019, povahou jsou data velmi citlivá, nebudou tedy až na některé výběrové charakteristiky sdílena. Referenční kvartál je Q4 2011, k němuž jsou vztaženy data ze ČSÚ, vyobrazeno na 3.1. Každému ze záznamů je přiřazena hodnota nezaměstnanosti a průměrné mzdy z předchozího kvartálu než v tom, v němž byl úvěr poskytnut. Cílem našeho modelu je odhad selhání v okamžiku podání nové žádosti o úvěr a hodnoty jsou statistickým úřadem zveřejňovány se zpožděním, proto záznamům přiřazujeme starší období.

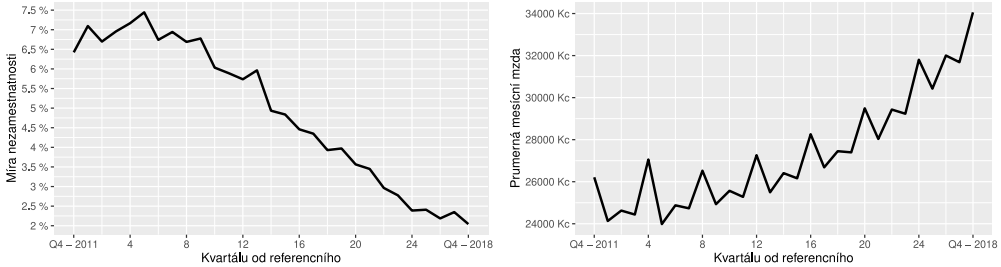
3.1.1 Dvouúrovňové regresory

O každém klientovi víme, zda pro něj platí následující:

1. je muž,

	Ženy − 0	Muži − 1
Rel. četnost	0,3775	0,6225

2. je vysokoškolsky vzdělaný – uvedl v bance svůj titul,



(a) Míra nezaměstnanosti v celé ČR. (b) Průměrná měsíční mzda v celé ČR.

Obrázek 3.1: Data z ČSÚ, kvartálně pro celou ČR.

Vlastnosti	Ne − 0	Ano − 1
VŠ	0,8969	0,1031
OSVČ	0,8554	0,1446
OSVČ - účet	0,8682	0,1318
Jiný úvěr	0,9589	0,0411
Spoř. účet	0,7921	0,2079
Not. úschova	0,9997	0,0003
Term. vklad	0,9817	0,0183
BÚ v cizí meně	0,8690	0,1310

Tabulka 3.1: Výběrové charakteristiky dvouúrovňových kategoriálních proměnných.

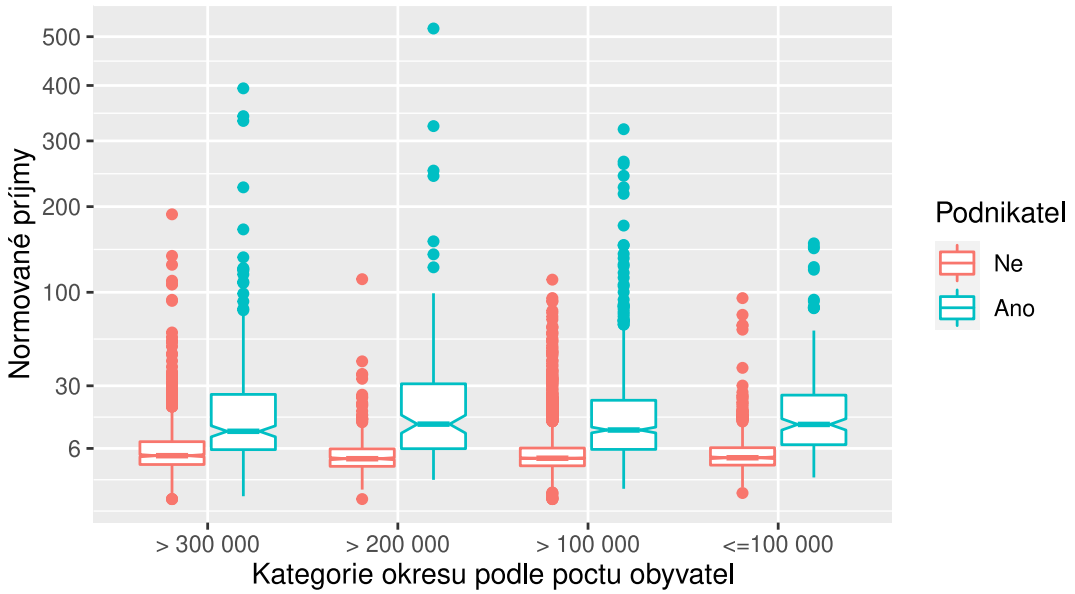
3. je podnikatel – v databázi přiřazeno IČO,
4. má podnikatelský účet – alespoň jeden z jeho účtů je v databázi veden jako podnikatelský (jiné podmínky než pro spotřebitele),
5. má alespoň jeden další úvěr od naší banky,
6. má v bance spořicí účet,
7. alespoň jeden z účtů je notářskou úschovou,
8. má v bance termínovaný vklad,
9. alespoň jeden z účtů klienta je v jiné měně než Kč.

Vlastnosti 3 a 4 chápeme jako oddělení podnikatelů od klasického spotřebitele, lze například očekávat, že aktivita na účtech OSVČ bude vyšší nebo alespoň rozdílná než u spotřebitele, účel úvěru se mezi spotřebiteli a podnikateli taktéž často liší např. podnikatel může chtít pouze vylepšit svou likviditu, naopak spotřebitel si za úvěr něco koupí. Vlastnostem 5-9 zase rozumíme tak, že někteří klienti využívají dalších služeb banky. Rozdělení v našich datech vidíme v tabulce 3.1.

3.1.2 Víceúrovňové regresory

Okresy

Vzhledem k tomu, že počet okresů v ČR je 77 včetně hlavního města Prahy, celkem uvažujeme 78 při započtení kategorie „Slovensko a jiné“. Testem věrohodnostním poměrem (1.22) ověříme, zda je vhodné tento faktor uvažovat. Mějme úplný model a model, z něhož vypouštíme prediktory „okres“. Testová statistika $G = 248{,}13 > 98{,}48 = \chi^2_{77}(1-\alpha)$, tedy na hladině $\alpha = 0{,}05$ zamítáme hypotézu, že by vektor parametrů faktoru okresy byl sdruženě nulový. Vidíme tedy, že uvažovat tento faktor v modelu má význam, ovšem počet úrovní s sebou přináší úskalí v podobě numerické stability modelu. Proto ve prospěch interpretability modelu a zlepšení numerické stability byla proměnná nahrazena kategoriemi okresů podle počtu obyvatel, ty jsou definovány podle 3.2. Okresům v jiných zemích byla přiřazena průměrná hodnota.



Obrázek 3.2: Průměrné celkové půlroční příjmy podle kategorie okresu bydliště.

Státní příslušnost

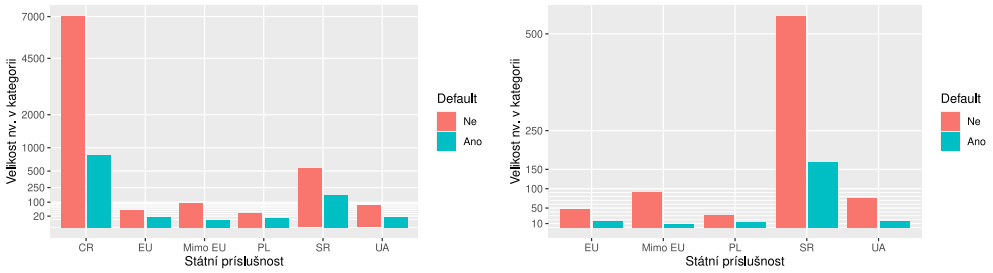
Státní příslušnost bývá v praxi uvažována ve spojení s délkou pobytu v zemi, kde klient o úvěr žádá, a jinými vazbami na ni. Absence takových vazeb může indikovat finanční problémy v zahraničí, případně jiné problémy spojené s AML směrnicemi. Takovou kombinaci dat k dispozici nemáme, proto uvažujeme pouze tuto proměnnou s celkem šesti úrovněmi. Jak jsou úvěry děleny vidíme na 3.3. Počty a jednotlivé úrovně již známe z 2.3.

3.1.3 Diskrétní regresory

Každého klienta charakterizují tyto celočíselné charakteristiky:

1. věk při poskytnutí – vypočteno jako rozdíl mezi rokem poskytnutí a rokem narození klienta, detailnější informaci neznáme,

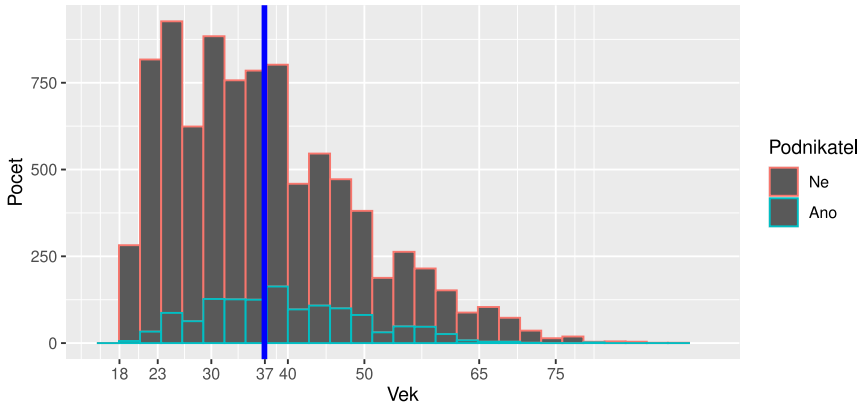
2. dní klientem při poskytnutí – počet dní mezi poskytnutím úvěru a první registrací klienta v bance,



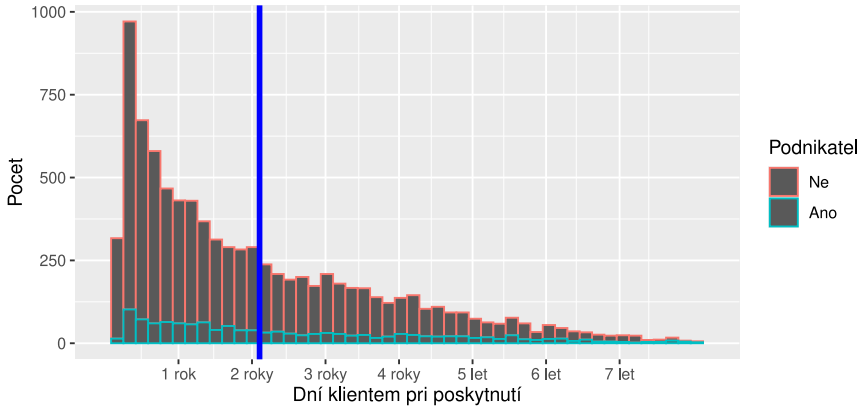
(a) Poskytnuté úvěry podle stát. příslušnosti, kvadratická stupnice.

(b) Poskytnuté úvěry podle stát. příslušnosti bez ČR.

Obrázek 3.3: Počty poskytnutých úvěrů podle státní příslušnosti.



(a) Histogram proměnné věk při poskytnutí.



(b) Histogram proměnné dní klientem banky při poskytnutí.

Obrázek 3.4: Histogramy věku klientů a délky jejich historie v bance s dělením na spotřebitele a podnikatele, modře zvýrazněny výběrové průměry.

3. počet účtů v bance – kolik účtů je k dané osobě v databázi vedeno,
4. počet příchozích plateb na všech účtech,
5. počet odchozích plateb na všech účtech.

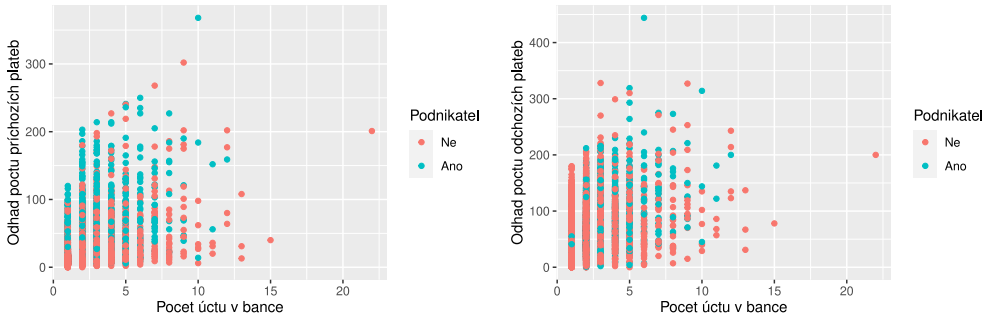
Rozdělení 1 a 2 vidíme v histogramech 3.4. Hodnoty proměnných 4 a 5 jsou kumulativní za šestiměsíční období před poskytnutím úvěru.

3.1.4 Spojité regresory

Další charakteristiky tvoří

1. nezaměstnanost v minulém kvartálu,
2. normované příjmy,
3. normované výdaje.

Hodnoty nezaměstnanosti jsou již dříve zmiňované hodnoty ukazatelů od statistického úřadu, tedy nikoliv původně součástí datového souboru, tyto informace banka ke klientům neeviduje.



(a) Počet příchozích plateb k počtu účtů.

(b) Počet odchozích plateb k počtu účtů.

Obrázek 3.5: Počty evidovaných plateb v závislosti na počtu účtů daného klienta s dělením na podnikatele a spotřebitele.

Regresor	Min.	1. kvartil	Medián	3. kvartil	Max.
Norm. příjmy	0	2,801	4,366	7,449	517,814
Norm. výdaje	−510,018	−7,501	−4,352	−2,723	0

Tabulka 3.2: Výběrové charakteristiky normovaných příjmů, výdajů.

Pod pojmem normované příjmy nebo výdaje rozumíme sumu příchozích plateb respektive výdajů vydělených průměrnou mzdou v předchozím kvartále, než byl poskytnut úvěr. Cílem této normalizace je brát v potaz růst průměrné mzdy v průběhu sedmi sledovaných let. Výběrové charakteristiky pro tyto veličiny najdeme v tabulce 3.2. Jak se normované příjmy liší mezi okresy podle velikosti vidíme na 3.2.

3.2 Tvorba modelu

Podotkněme nejprve, že výsledný model bude vybrán na základě statistik popsaných v kapitole 2.2. Ukážeme si vlastnosti popsané v kapitole 2.3. Pokud nebude možné jednoznačně rozhodnout na základě těchto vlastností, dáme přednost modelu s nižším počtem kovariátů.

Nyní se již můžeme pustit do modelování. Vytvořme tedy úplný model se všemi výše popsanými proměnnými. Podívejme se do tabulky 3.3 na hodnoty koeficientů a jejich signifikantnost. Vidíme, že u mnohých z kovariátů nemůžeme vyloučit, že po jednom nemají vliv na odezvu. V tabulce si též všimněme, že na základě Waldovy statistiky W a p-hodnoty zamítáme hypotézu, že je koeficient kovariátu „Dní klientem při poskytnutí“ nulový, přesnější hodnota jeho odhadu je $-5{,}451 \times 10^{-4}$, regresor totiž nabývá hodnot od jednotek po tisíce, v zájmu zachování přehledného formátu tabulky tuto informaci zmiňujeme zde.

Pokusme se o eliminaci nevýznamných nezávisle proměnných. Proveďme tedy vzestupný a sestupný výběr regresorů. Vzhledem k počtu kovariátů využijeme zabudovanou funkci v R, která pro porovnání modelů v jednotlivých krocích používá statistiku AIC, což nám s ohledem k postupu popsaném v podkapitole 2.2.1 vyhovuje.

```
> step(fit.glm.all, direction = "forward")
> step(fit.glm.all, direction = "backward")
```

Dostáváme dva modely, s nimiž budeme dále pracovat. Vzestupným výběrem dostáváme opět úplný model, sestupný výběr nám ovšem dává již rozdílný model, podívejme se na jejich srovnání do tabulky 3.4. Vidíme, že z modelu byly vypuštěny regresory „Alespoň jeden účet v cizí měně, Jiný úvěr, Kategorie okresu, Normované výdaje, Notářská úschova“ a „Podnikatel, Podnikatelský účet“. Tato informace je překvapivá hlavně u informaci o jiném již poskytnutém úvěru nebo u identifikace podnikatele. Jiné úvěry si lze vysvětlit, že banka, ve snaze udržet si klienta, poskytne další úvěr, který je vyhodnocen jako rizikovější a dojde k delikvencím. Nízkou významnost identifikace podnikatele si můžeme vysvětlit tak, že odlišné chování podnikatele od spotřebitele je dostatečně vysvětleno už jeho nakládáním s účty, tedy příjmy, počty příjmů a počty výdajů. Všimněme si také změny p-hodnot v užším modelu, vidíme, že skoro všechny regresory v modelu sestupného výběru jsou významné, s výjimkou „Termínovaný vklad“ a faktoru státní příslušnosti „mimo EU“, což je samo o sobě velmi široká skupina shrnující země od USA a Kanady až po Bělorusko nebo Srbsko, tedy lze chápat, že chování v rámci této skupiny nebude jednotné.

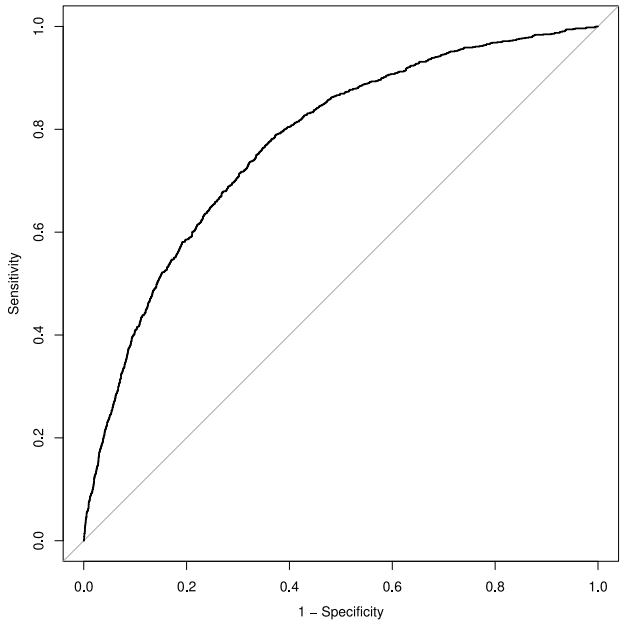
3.2.1 Vyhodnocení modelů

Který z modelů je tedy lepší? Porovnejme jejich vlastnosti v tabulce 3.5. Pozorujeme, že AIC je lepší pro model sestupného výběru, věrohodnost a z ní počítané statistiky deviance a R^2_{gen} jsou lepší pro úplný model. Odhad chybovosti pomocí křížové validace je téměř identický a ani AUC nám neukazuje jasnou převahu jednoho modelu nad druhým, na obrázku 3.6 vidíme ROC křivku modelu sestupného výběru. Poznamenejme, že oba modely podle dříve popsaných pravidel viz (2.8) hodnotíme diskriminaci obou modelů jako akceptovatelná. Zvolme pro další zkoumání model s nižším počtem parametrů, tedy model, který vznikl sestupným výběrem. Pro tento model se rozhodujeme na základě AIC a preference modelu s nižším počtem parametrů.

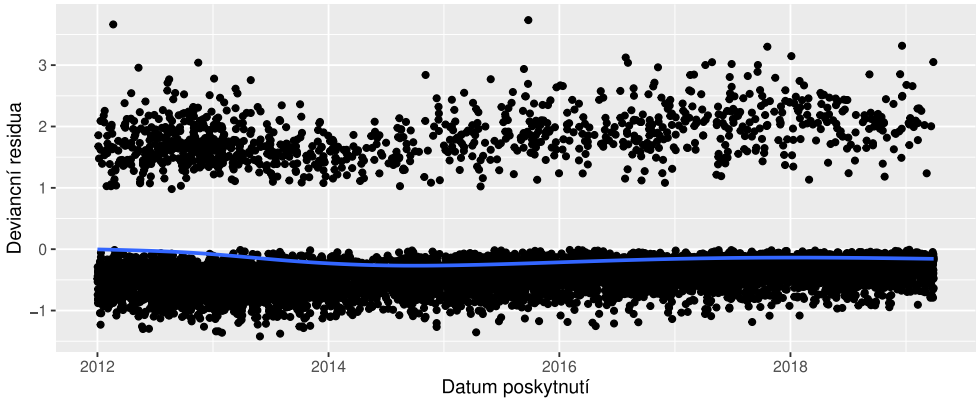
3.2.2 Analýza residuí

Podívejme se na residua vybraného modelu. Nejprve se zaměřme na to, zda se v průběhu času neprojevuje trend, který by bylo potřeba zachytit, to vidíme na grafech 3.7, kde zároveň vidíme rozdíl mezi dříve nadefinovanými Pearsonovými residui (2.6) a deviančními residui (2.7). Modrá křivka vyrovnává data pomocí geom_smooth() knihovny ggplot v R, my chceme, aby byla co nejblíže 0, tedy ose x viz vlastnosti residua (1.14). Pokud by se hodnota významně vychylovala od hodnoty 0, znamenalo by to, že v datech je nějaký časový trend, nebo šok, který se na základě našich prediktorů nepodařilo zachytit. Pod pojmem šok si můžeme představit např. nucené uzavírání podniků kvůli pandemii, jiným projevem by mohly být ekonomické cykly, žádnou takovou informaci ale z grafu residuí nezjišťujeme.

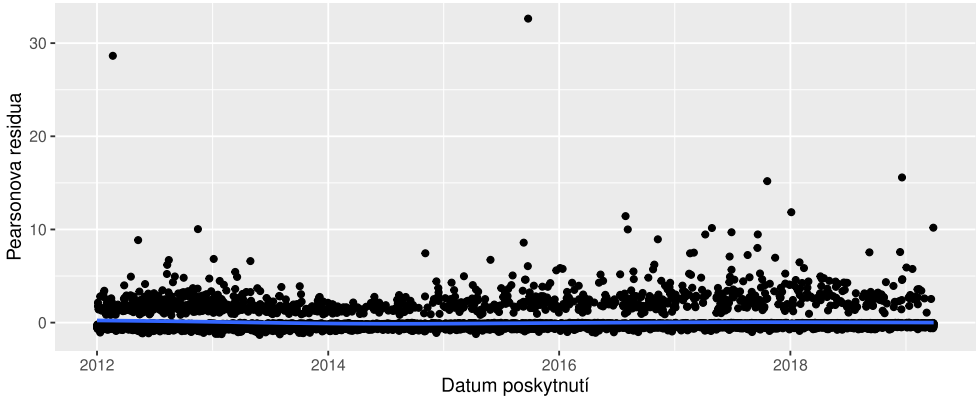
Podívejme se nyní na 3.8a a 3.8b, kde jsou vyobrazeny Pearsonovy residua vzhledem ke dvěma regresorům. Hledáme zde opět jakékoliv vychýlení modré



Obrázek 3.6: ROC pro model vzniklý sestupným výběrem.



(a) Devianční residua.



(b) Pearsonova residua.

Obrázek 3.7: Residua pro model sestupného výběru vzhledem k času.

	Odhad koeficientu	Waldova statistika	p-hodnota
(Intercept)	−0,51	−2,09	0,04
Pohlaví	0,34	4,50	0,00
VŠ vzdělání	−0,82	−4,65	0,00
Věk při poskytnutí	−0,04	−11,14	0,00
Dní klientem při poskytnutí	0,00	−6,18	0,00
Podnikatel	0,03	0,10	0,92
Podnikatelský účet	−0,20	−0,71	0,48
Počet účtů	−0,24	−3,91	0,00
Jiný úvěr	−0,34	−1,24	0,21
Spořicí účet	−0,59	−4,20	0,00
Notářská úschova	−9,83	−0,06	0,95
Termínovaný vklad	−1,05	−1,72	0,09
Alespoň jeden účet v cizí měně	−0,04	−0,26	0,80
Počet příjmů	0,01	4,75	0,00
Počet výdajů	−0,01	−7,85	0,00
Míra nezaměstnanosti	0,17	7,35	0,00
Normované příjmy	−0,02	−1,30	0,19
Normované výdaje	−0,01	−0,72	0,47
Státní příslušnost: EU	1,34	4,33	0,00
Státní příslušnost: mimo EU	0,22	0,61	0,54
Státní příslušnost: PL	1,40	4,11	0,00
Státní příslušnost: SR	0,95	8,74	0,00
Státní příslušnost: UA	0,72	2,39	0,02
Kategorie okresu	−0,02	−0,43	0,67

Tabulka 3.3: Shrnutí úplného modelu.

křivky od osy x, ale nenacházíme jej, což je pozitivní informace. Poslední vyobrazení residuí je graf 3.8c, kde sledujeme devianční residua podle predikované pravděpodobnosti výskytu delikvence dané modelem sestupného výběru. Z tohoto grafu vidíme, že náš model je značně chybový.

3.2.3 Chybovost modelu

Připomeňme, že default rate v celém datovém souboru je 13,3%. Zvolme nyní hodnotu $treshold = 0{,}1$ a podívejme se na confusion matrix. Dostáváme hodnoty

$$TPR = \frac{839}{1048} = 0{,}801,$$
$$TNR = \frac{4775}{7855} = 0{,}608.$$

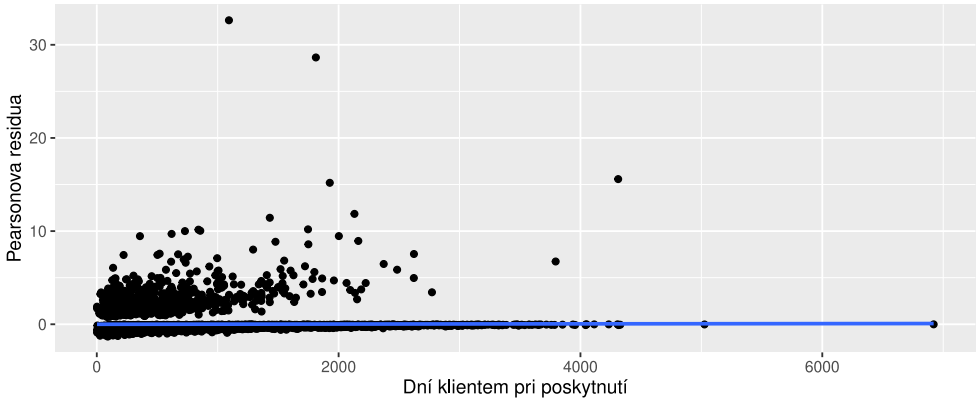
Porovnáme-li tyto hodnoty s modelem založeným pouze na státní příslušnosti viz 2.5, vidíme, že na dané úrovni *treshold* jsme s bohatším modelem odhalili daleko více true positive subjektů. Na úkor senzitivity (TPR) je nižší specificita (TNR),

	Vzestupný výběr		Sestupný výběr	
	Odhad koeficientu	p-hodnota	Odhad koeficientu	p-hodnota
(Intercept)	−0,51	0,04	−0,54	0,01
Alespoň jeden účet v cizí měně	−0,04	0,80		
Dní klientem při poskytnutí	0,00	0,00	0,00	0,00
Jiný úvěr	−0,34	0,21		
Kategorie okresu	−0,02	0,67		
Míra nezaměstnanosti	0,17	0,00	0,17	0,00
Normované příjmy	−0,02	0,19	−0,01	0,02
Normované výdaje	−0,01	0,47		
Notářská úschova	−9,83	0,95		
Počet příjmů	0,01	0,00	0,01	0,00
Počet účtů	−0,24	0,00	−0,27	0,00
Počet výdajů	−0,01	0,00	−0,01	0,00
Podnikatel	0,03	0,92		
Podnikatelský účet	−0,20	0,48		
Pohlaví	0,34	0,00	0,33	0,00
Spořicí účet	−0,59	0,00	−0,55	0,00
Státní příslušnost: EU	1,34	0,00	1,34	0,00
Státní příslušnost: mimo EU	0,22	0,54	0,24	0,51
Státní příslušnost: PL	1,40	0,00	1,40	0,00
Státní příslušnost: SR	0,95	0,00	0,94	0,00
Státní příslušnost: UA	0,72	0,02	0,73	0,01
Termínovaný vklad	−1,05	0,09	−0,98	0,10
Věk při poskytnutí	−0,04	0,00	−0,04	0,00
VŠ vzdělání	−0,82	0,00	−0,82	0,00

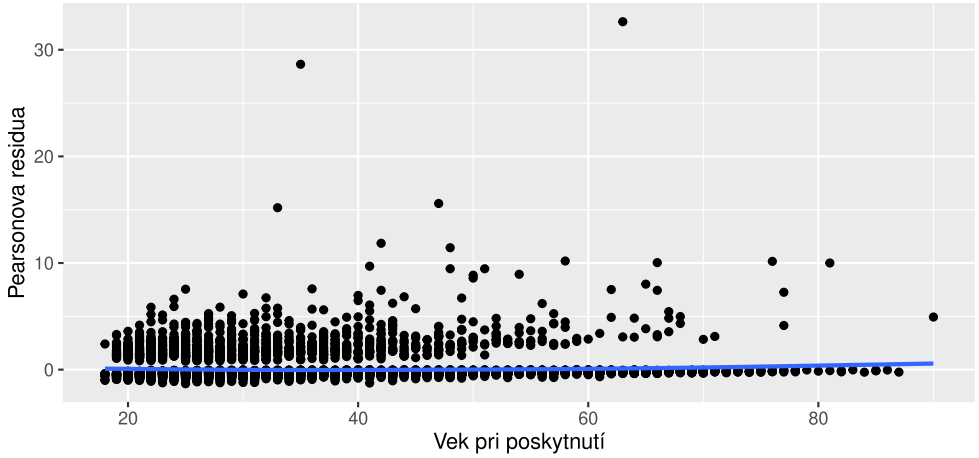
Tabulka 3.4: Porovnání odhadu koeficientů modelů vzniklých různými postupy jejich výběru.

	Plný model	Vzestupný výběr	Sestupný výběr
AIC	5591,376	5591,376	5581,451
Log. věrohodnost	−2771,688	−2771,688	−2773,725
Deviance	5543,376	5543,376	5547,451
R^2_{gen}	0,0970	0,0970	0,0966
CV_{1000}	0,0924	0,0923	0,0924
AUC	0,7729	0,7729	0,7727

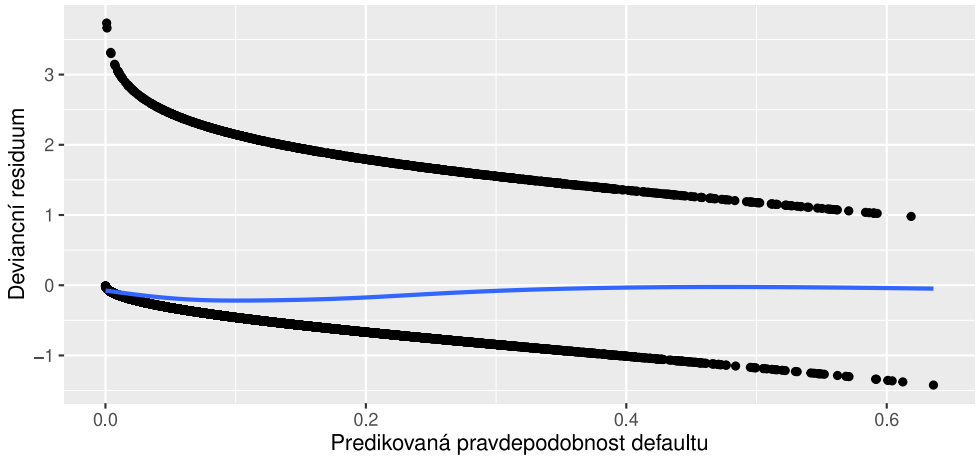
Tabulka 3.5: Porovnání vlastností modelů. Doba výpočtu CV_{1000} byla v průměru 141 vteřin pro plný a vzestupný model, pro sestupný pak 68 vteřin.



(a) Residua podle doby, jakou byl v okamžiku poskytnutí úvěru klientem.



(b) Residua podle věku klienta při poskytnutí úvěru.



(c) Residua podle odhadnuté pravděpodobnosti.

Obrázek 3.8: Residua pro model sestupného výběru vzhledem regresorům a predikované pravděpodobnosti defaultu v modelu sestupného výběru.

		Reálný status		
		0	1	celkem
Predikovaný status	0	4775	209	4984
	1	3080	839	3919
	celkem	7855	1048	8903

Tabulka 3.6: Confusion matrix pro model vzniklý sestupným výběrem, práh 0,1.

vidíme, že více než třetina true negative subjektů by byla modelem odmítnuta. Celkový odhad testové chyby modelu známe z křížové validace v 3.4, jeho hodnota je 9,24%, což je stále nezanedbatelné zlepšení oproti dosavadnímu postupu banky.

3.2.4 Interpretace parametrů

S vybraným modelem a vyhodnocením signifikantnosti jednotlivých kovariátů se pusťme do interpretace odhadnutých parametrů. Baseline tohoto modelu je žena s českým občanstvím, ale bez vysokoškolského vzdělání, nepodniká, nemá v bance zřízené jiné služby (spořicí účet nebo termínovaný vklad). Poměry šancí pro dvouúrovňové, diskrétní a spojité prediktory jsou

$$\begin{aligned}
&OR_{pohlavi}(1,0) && = 1{,}3944147,\\
&OR_{vzdelani}(1,0) && = 0{,}4387678,\\
&OR_{sporeni}(1,0) && = 0{,}5775808,\\
&OR_{terminovanyvklad}(1,0) && = 0{,}3738339,\\
&OR_{dniklientem}(x+180,x) && = 0{,}9065396,\\
&OR_{prijmy}\left(x+\frac{6}{10},x\right) && = 0.9927107,\\
&OR_{pocetprijmu}(x+1,x) && = 1{,}0107185,\\
&OR_{pocetvydaju}(x+5,x) && = 0{,}9462485,\\
&OR_{pocetuctu}(x+1,x) && = 0{,}7615487,\\
&OR_{nezemestnanost}\left(x+\frac{1}{2},x\right) && = 1{,}090557,\\
&OR_{vek}(x+5,x) && = 0{,}8261504.
\end{aligned}$$

Jak bylo zmíněno v (1.20), je nutné vhodně volit volby $c \in \mathbb{R}$, volíme tedy takto:

- dní klientem: $c = 180$, změna poměru šancí s každým dalším půl rokem, kdy je subjekt klientem banky,
- normované příjmy: $c = 0{,}6$, jde o nárůst měsíční mzdy o desetinu průměrné měsíční mzdy v daném období,
- počet příjmů: $c = 1$, každá další příchozí platba,
- počet výdajů: $c = 5$, každých dalších 5 odchozích plateb, hodnota $c = 1$ je v tomto případě nízká, spotřebitel může platit i několikrát denně,

- počet účtů: $c = 1$, každý další účet navíc,
- nezaměstnanost: $c = 0{,}5$, nárůst nezaměstnanosti o 50bp,
- věk: $c = 5$, každých 5 let věku klienta.

Pro víceúrovňový prediktor státní příslušnosti, dostáváme, že v porovnání s osobou o české státní příslušnosti jsou poměry šancí

$$\begin{aligned} OR_{EU}(1{,}0) &= 3{,}825179, \\ OR_{mimoEU}(1{,}0) &= 1{,}268159, \\ OR_{PL}(1{,}0) &= 4{,}074064, \\ OR_{SR}(1{,}0) &= 2{,}555765, \\ OR_{UA}(1{,}0) &= 2{,}077609. \end{aligned}$$

3.3 Shrnutí modelování

V této kapitole jsme se podívali na reálná data českých bankovních klientů, zvážili jsme jejich kategorizaci a provedli jsme logistickou regresi, z níž jsme vytvořili dva modely, ze kterých byl jeden shodný s úplným modelem. Jejich klasifikační vlastnosti byly obdobné, a tak byl pro hlubší analýzu vybrán model s nižším počtem parametrů. Podívali jsme se na chování jeho residuí a pokusili jsme se o interpretaci výsledků.

Zjistili jsme, že odhad pravděpodobnosti selhání významně ovlivňuje vzdělání, dále model zvýhodňuje ty klienty, kteří využívají co nejvíce běžných bankovních služeb co nejčastěji. Vliv výše příjmů na odhad selhání není tak vysoký, jako bychom mohli očekávat, to může být dáno tím, že neznáme příslušné výše úvěrů. Ukázalo se, že muži jsou v porovnání se ženami významně rizikovějšími potenciálními dlužníky.

Dalším významným faktorem je státní příslušnost klienta, zůstává ale otázkou, zda by tento faktor měl být zvažován i v praxi. Výsledkem našeho bádání je, že na základě dat je tento faktor podstatným indikátorem rizikovosti žadatelů o úvěr.

Závěr

Modelování má v prostředí nejen ekonomie a financí své místo, popsali jsme si zběžně základní model lineární regrese, abychom jej zobecnili na v praxi hojně využívaný model logistické regrese, který je pro binární klasifikační problémy na rozdíl od lineární regrese vhodný. S popsaným modelem jsme si ukázali odhad jeho parametrů a nutnost využití numerických metod. Seznámili jsme se s Waldovým a Wilksovým testem statistické významnosti parametrů a jejich podmnožin. Věnovali jsme prostor rozboru kategoriálních proměnných, jejich interpretaci a úskalí. Popsali jsme, jak s danou množinou regresorů vytvořit co nejlepší model vybráním z množiny všech modelů a poté jsme se spokojili s iterativním výběrem proměnných. Pro vyhodnocení modelu jsme zavedli Akaikeho informační kritérium, zobecněný koeficient determinace a ROC křivku pro grafické znázornění diverzifikační schopnosti modelu. V závěrečné kapitole jsme tyto poznatky aplikovali na data o bankovních klientech spolu s daty od Českého statistického úřadu. Překvapením bylo, že se některé z regresorů ukázaly jako nevýznamné. Odhadovaná testová chyba vybraného modelu byla ovšem stále vysoká.

Seznam použité literatury

Cox, D. a Snell, E. J. (1989). *Analysis of Binary Data*. 2nd edition. Chapman and Hall/CRC. ISBN 9780412306204.

David W. Hosmer, J., Stanley Lemeshow a Rodney X. Sturdivant (2013). *Applied Logistic Regression*. 3rd ed. John Wiley & Sons, Inc., Hoboken, New Jersey. ISBN 978-0-470-58247-3.

Fahrmeir, L., Thomas Kneib, Stedan Lang a Brian Marx (2013). *Regression: Models, Methods and Applications*. 1st edition. Springer, Berlín. ISBN 978-3-642-34332-2.

James, G., Witten, D., Hastie, T. a Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. First edition, Corrected at the 8th printing 2017. Springer, New York. ISBN 978-1-4614-7137-0.

Kulich, M. (2014). Přehledový větník. Katedra pravděpodobnosti a matematické statistiky MFF UK. URL `https://www2.karlin.mff.cuni.cz/~pesta/NMFM301/statistika_fm.pdf`. Datum přístupu 18.7.2021.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, **78**(3), 691–692. ISSN 0006-3444. doi: 10.1093/biomet/78.3.691. URL `https://doi.org/10.1093/biomet/78.3.691`.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. a Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77.

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, **9**(1), 60 – 62. doi: 10.1214/aoms/1177732360. URL `https://doi.org/10.1214/aoms/1177732360`.

William, M. (2009). *Introduction to Probability and Statistics*. 13th ed. Brooks/Cole, Belmont, CA. ISBN 978-0-495-38953-8.

Zvára, K. (2019). *Regrese*. 2. vydání. MatfyzPress, nakladatelství Matematicko-fyzikální fakulty Univerzity Karlovy v Praze, Praha, CZ. ISBN 978-80-7378-406-5.

Seznam obrázků

Seznam tabulek