České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská

Analýza cen pojištění pomocí strojového učení

Analysis of insurance pricing by machine learning techniques

Diplomová práce

Autor: **Bc. Barbora Pánková**

Vedoucí práce: **Mgr. Tomáš Petr**

Konzultant: **Ing. Petr Svojítka**

Akademický rok: 2019/2020

Poděkování:

Chtěla bych poděkovat především svému školiteli Mgr. Tomáši Petrovi za vedení mé diplomové práce, za jeho čas, ochotu, vstřícnost a odbornou pomoc. Velké díky patří také konzultantovi Ing. Petru Svojítkovi nejen za podnětné myšlenky a rady v oblasti pojišťovnictví, ale také za ochotu a pomoc při výběru samotného tématu práce. V neposlední řadě bych chtěla poděkovat Janu Macháčovi za poskytnutí klidného a tichého místa pro psaní této práce a rodině za psychickou i finanční pomoc po celou dobu studia.

Čestné prohlášení:

Prohlašuji, že jsem tuto práci vypracovala samostatně a uvedla jsem všechnu použitou literaturu.

V Praze dne 29. června 2020

Barbora Pánková

Název práce:

Analýza cen pojištění pomocí strojového učení

Autor: Barbora Pánková

Obor: Matematické inženýrství

Zaměření: Aplikované matematicko - stochastické metody

Druh práce: Diplomová práce

Vedoucí práce: Mgr. Tomáš Petr, Deloitte Advisory s.r.o., Budova Churchill I, Italská 2581/67, Vinohrady 120 00 Praha 2

Konzultant: Ing. Petr Svojítka, Deloitte Advisory s.r.o., Budova Churchill I, Italská 2581/67, Vinohrady 120 00 Praha 2

Abstrakt: Tato diplomová práce se zabývá použitím zobecněných lineárních modelů a modelů neuronových sítí pro odhadnutí cen konkurence a zkoumáním jejich struktury v pojištění odpovědnosti z provozu vozidla. Práce obsahuje potřebné shrnutí teorie z oblasti zobecněných lineárních modelů a neuronových sítí. Na základě této teorie jsou odhadovány modely struktur cen dvou společností s odlišným množstvím pozorování. V práci jsou srovnávány různé postupy založené na zobecněných lineárních modelech a neuronových sítí aplikovaných na vstupní data.

Klíčová slova: neživotní pojištění, zobecněný lineární model, gamma regrese, výše pojistného, srovnání cen konkurence, povinné ručení, neuronové sítě

Title:

Analysis of insurance pricing by machine learning techniques

Author: Barbora Pánková

Abstract: This master thesis deals with the use of generalized linear models and neural network models for estimating competition prices and examining their structure in motor third party liability insurance. The thesis contains the necessary summary of the theory of generalized linear models and neural networks. Based on this theory, models of price structures of two companies with different amounts of observations are estimated. The work compares various methods based on generalized linear models and neural networks applied to input data.

Key words: non-life insurance, generalized linear model, gamma regression, premium, comparation of prices, car insurance, neural network

Katedra: matematiky Akademický rok: 2019/2020

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student:	Bc. Barbora Pánková
Studijní program:	Aplikace přírodních věd
Studijní obor:	Aplikované matematicko-stochastické metody
Název práce (česky):	Analýza cen pojištění pomocí strojového učení
Název práce (anglicky):	Analysis of insurance pricing by machine learning techniques

Pokyny pro vypracování:

1) Seznamte se s obvyklými přístupy strukturování cen pojištění vozidel a s obvykle využívaným matematickým aparátem. Speciální pozornost věnujte GLM modelu.
2) Seznamte se s metodami strojového učení, pozornost věnujte prediktivním metodám, které lze využít k natrénování modelu cen pojištění na datech o pojistnících a jim poskytnutých cenách za pojištění vozidel a následně pro predikci těchto cen.
3) Pomocí metod strojového učení na základě rozsáhlého vzorku cen pojistitele A (vlastní ceny) sestavte model schopný pro každého zákazníka predikovat jeho cenu nabízenou pojistitelem A.
4) Na základě dostupného méně rozsáhlého vzorku cen pojistitele B (ceny konkurence) upravte model tak, aby pro každého zákazníka predikoval jeho cenu nabízenou pojistitelem B.
5) Ověřte, zda daný model predikuje ceny pojistitele B spolehlivěji než model sestavený pouze na základě vzorku cen pojistitele B a srovnejte se stávajícími modely.

Doporučená literatura:

1) D. J. C. MacKay. Information Theory, Inference and Learning Algorithms, Cambridge University Press, 2003.
2) I. Goodfellow and Y. Bengio and A. Courville. Deep Learning, MIT press, 2016.
3) A. Bernal, S. Fok and R. Pidaparthi. Financial Market Time Series Prediction with Recurrent Neural Networks, Stanford University, 2012.
4) P. Jong and G. Z. Heller. Generalized linear models for insurance data, Cambridge University Press, 2008.

Jméno a pracoviště vedoucího diplomové práce:

Mgr. Tomáš Petr
Deloitte Advisory s.r.o. , Budova Churchill I, Italská 2581/67, Vinohrady, 120 00 Praha 2

Jméno a pracoviště konzultanta:

Ing. Petr Svojítka
Deloitte Advisory s.r.o., Budova Churchill I, Italská 2581/67, Vinohrady , 120 00 Praha 2

Datum zadání diplomové práce: 31.10.2019

Datum odevzdání diplomové práce: 4.5.2020

Doba platnosti zadání je dva roky od data zadání.

Obsah

Úvod **10**

1 Princip pojišťovnictví **11**
1.1 Pojištění odpovědnosti z provozu vozidla - POV . . . 11
1.2 Odhadování cen konkurence . . . 12

2 Zobecněný lineární model **14**
2.1 Specifikace modelu . . . 14
2.2 Exponenciální rodina . . . 14
2.3 Zobecnění lineární regrese . . . 15
2.3.1 Koeficient determinace . . . 17
2.4 Korelace . . . 17
2.5 Testování hypotéz . . . 17
2.6 Křížová validace . . . 18

3 Modely odhadu cen konkurence **19**
3.1 Model bílé pojišťovny . . . 20
3.1.1 Specifikace modelu . . . 20
3.1.2 Analýza proměnných . . . 20
3.1.3 Korelace mezi proměnnými . . . 31
3.1.4 Regresní model a odhad parametrů . . . 33
3.2 Odhad ročního pojistného bílé pojišťovny . . . 36
3.3 Validace modelu na testovacích množinách . . . 38
3.4 Model černé pojišťovny . . . 40
3.4.1 Specifikace modelu . . . 40
3.4.2 Analýza proměnných . . . 40
3.4.3 Korelace mezi proměnnými . . . 48
3.4.4 Regresní model a odhad parametrů . . . 49
3.5 Odhad ročního pojistného černé pojišťovny . . . 52
3.6 Vylepšení odhadu ročního pojistného černé pojišťovny . . . 54

4 Umělé neuronové sítě **58**
4.1 Model s využitím neuronové sítě . . . 59
4.2 Odhad cen konkurence pomocí neuronové sítě . . . 60
4.2.1 Vstupní data . . . 60
4.3 Model bílé pojišťovny . . . 61
4.3.1 Trénovací množina o 30 000 vzorcích . . . 61

4.3.2 Trénovací množina o 200 000 vzorcích 63
4.4 Model černé pojišťovny 64
4.4.1 Trénovací množina o 2 000 vzorcích 64
4.4.2 Vylepšení odhadu černé pojišťovny 65
4.5 Výsledky 67

Závěr **68**

Úvod

Tato diplomová práce se zabývá odhadem cen konkurence v pojištění odpovědnosti z provozu vozidla. Pro toto odvětví je charakteristické, že nabízený produkt je vysoce komoditizovaný a cena je tedy pro zákazníkův výběr pojistitele klíčová. I v důsledku toho je struktura cen každé pojišťovny komplexnější než u jiných produktů, a to především ve snaze podchytit rizikovost každého klienta. Pojišťovny v ČR tak kompletní strukturu svých tarifů v rámci konkurenčního boje nezveřejňují. Ačkoli existují různé veřejné srovnávače či on-line nabídky pojišťoven, sběr většího množství dat z těchto zdrojů je problematický. Ceny zde uváděné navíc nemusí být konečné a zároveň se mohou lišit od cen nabízených přes jiné prodejní kanály. V této situaci je pro každou pojišťovnu důležité co nejlépe rozumět struktuře cen konkurentů. Díky tomu může pojišťovna optimalizovat vlastní ceny nejen s ohledem na rizikovost klientů, ale i na to, zda si za takovou cenu produkt mezi konkurenčními nabídkami vybere dostatečný počet zákazníků.

V dnešní době se odhad cen konkurence řeší různými způsoby. Například pojistitel A může získat určité množství cen pro profily zákazníků vzájemně se lišící vždy pouze v jednom parametru (odlišná značka aut, věk, síla motoru, ...) a na jejich základě dopočítat koeficienty sazebníku pojistitele B. Tato metoda mohla fungovat zejména dříve, pokud pojistitel B používal k sazbování velmi jednoduchý multiplikativní tarif s malým množstvím parametrů. Pokud však pojistitel B počítá cenu na základě složitějšího modelu, tento způsob by pro uspokojivé výsledky vyžadoval obrovská množství pozorování. Jinou užívanou metodou je zobecněný lineární model. Pojistitel A sestaví na základě dostupného vzorku cen pojistitele B zobecněný lineární model a pomocí něj odhaduje ceny konkurenta.

Cílem této práce je vyzkoušet různé metody k nabytí znalosti o struktuře cen konkurence v celé jejich šíři, tedy ve větším detailu, než je reálné přímo získat z veřejně dostupných zdrojů. Této znalosti se pokusíme dosáhnout metodou zobecněné lineární regrese a alternativní metodou neuronových sítí. Obě metody budou nejprve aplikovány na rozsáhlejší vzorek pojistitele A a následně na méně rozsáhlý vzorek pojistitele B. Výsledný odhad cen pomocí obou metod bude porovnán se vstupními daty, bude stanovena spolehlivost odhadu a následně porovnána úspěšnost různých přístupů na těchto různě velkých vzorcích dat.

V první kapitole této práce je stručně sepsán princip pojišťovnictví a pojištění odpovědnosti z provozu vozidla a odhadování cen konkurence. V druhé kapitole se nachází základy teorie potřebné k pochopení zobecněné lineární regrese, korelační analýzy a validace modelu. V další kapitole je pak teorie převedena do praxe a je zde popsán model odhadu cen pomocí zobecněné lineární regrese, a to jak pojistitele A, tak pojistitele B. Dále v kapitole testujeme, zda je možné dosáhnout spolehlivější predikce v modelu pojistitele B obohacením o znalost cen pojistitele A. Poslední kapitola pak nabízí úvod do metody neuronových sítí, představuje možný postup sestavení modelu a jeho použití na dostupných datech pojistitele A a B. Na závěr se srovnána metoda neuronových sítí s metodou zobecněné lineární regrese a jsou porovnány výsledky.

Kapitola 1

Princip pojišťovnictví

Součástí života každého jedince jsou náhodné jevy. Ne všechny tyto jevy jsou však pozitivní a žádané. Nástrojem finanční eliminace negativních důsledků nahodilosti je pojištění. Hlavním cílem pojištění je zajistit pojistnou ochranu jedince, bezporuchový chod ekonomiky státu a konkurenci či spolupráci s bankovním sektorem na finančním trhu [1].

Na pojištění lze tedy nahlížet jako na ochranu proti pojistným rizikům. Pojištěný přenese svá rizika, jejichž potenciální škodní důsledky jsou z jeho individuálního hlediska neúnosné, na pojistitele (pojišťovny, penzijní fond, stát). Ten je schopen celkově převzatá rizika s využitím inkasovaného pojistného nejen zvládat, ale případně je učinit předmětem výnosné činnosti. [1]

Pojistná smlouva se uzavírá na dohodnutou pojistnou dobu s určitou pojistnou částkou, což je nejvyšší dohodnuté plnění z pojistné události, eventuálně z několika pojistných událostí za určité časové období. Může být využívána i jako základ pro výpočet pojistného. Pojistné je pak úplata za poskytování pojistné ochrany. Je placeno v předem daných časových intervalech (obvykle ročně) na předem stanovenou dobu. [3]

Pojištění je rozdělováno pomocí mnoha kritérií. Tím prvním je právní hledisko, kde se rozlišuje dobrovolné pojištění a pojištění povinné. Dobrovolné pojištění uzavírá pojistník s pojišťovatelem ze své vlastní vůle formou pojistné smlouvy (např.: pojištění smrti nebo pojištění úrazu). Povinné pojištění je pak dáno buď zákonem (např.: pojištění odpovědnosti z provozu vozidla) nebo právním předpisem, tzv. povinné smluvní pojištění (např.: pojištění odpovědnosti z provozu civilních letadel). [1]

Dále můžeme pojištění dělit na životní (např. pojištění smrti nebo dožití, důchodové pojištění, atd.) a neživotní (např.: pojištění majetku, zemědělské pojištění, atd.).

V této diplomové práci se budeme zabývat neživotním pojištěním, a to pojištěním odpovědnosti z provozu vozidla, tedy lidově řečeno povinným ručením (POV).

1.1 Pojištění odpovědnosti z provozu vozidla - POV

Pojištění odpovědnosti z provozu vozidla neboli „povinné ručení“ je ochrana pojištěného za škodu způsobenou motorovým vozidlem druhé osobě. „Povinné ručení“ nelze použít pro škody na vlastním vozidle. [5] Jelikož každý vlastník motorového vozidla má povinnost POV uzavřít, je toto pojištění na trhu velmi časté. Zároveň je v „povinném ručení“ vysoká cenová konkurence. Parametry tohoto produktu

jsou dané zákonem. Produkt je proto vysoce komoditizovaný, čili cena zůstává jako dominantní rozlišovací faktor nabídek od různých pojišťoven.

Cenová struktura pojišťoven se může lišit v mnoha ohledech. První odlišnost výpočtu výše ročního pojistného může nastat v tzv. tarifních skupinách. Tarifní skupiny jsou homogenní skupiny pojistných smluv, pro něž je pojištěné riziko přibližně stejné, takže v rámci každé tarifní skupiny je možné vyžadovat jednotnou pojistnou sazbu. Každá tarifní skupina odpovídá určité rizikové úrovni. [1] Jednou z proměnných definující tarifní skupiny v případě povinného ručení může být například proměnná velikosti místa bydliště. Zde by tarifní skupiny mohly být: hlavní město, krajská města, okresní města, menší města a vesnice. Předpokládáme, že se riziko nehody liší. Čím vyšší je hustota dopravy a koncentrace aut, tím vyšší je riziko nehody. Příkladem dalších proměnných definující tarifní skupiny mohou být typ vozidla, věk řidiče, apod.

Dalším faktorem, který může způsobit rozdílnost sazebníků jednotlivých pojišťoven je systém bonusů a malusů. Pojišťovny jsou ze zákona č.168/1999 Sb. § 3b povinné zohlednit při výpočtu výše pojistného bezškodný průběh v předchozích pojistných obdobích. Tento systém jednoznačně zvýhodňuje a motivuje řidiče k jízdě bez nehod. Ti, co jezdí beze škod, získávají slevu na pojistném (tzv. bonus). Naopak jestliže řidič nesplní podmínky stanovené ve smlouvě nebo provozem vozidla způsobil škodu, může pojišťovna uplatnit přirážku k pojistnému (tzv. malus). Pojišťovny stanovují podmínky získání bonusu v pojistných podmínkách a mohou se u jednotlivých pojišťoven lišit. [4] Kromě bonusů/malusů daných zákonem mívají pojišťovny také tarifní faktor škodní historie. Ten vstupuje do modelu pojišťovny jako další proměnná, ale pro zákazníka je neznámý.

Taktéž se roční cena pojištění může lišit v závislosti na formě sjednání. Některé pojišťovny mohou mít slevu za sjednání pojištění online, jiné ne. Plus každá pojišťovna má jiné distributory, kteří mohou také nabízet odlišné ceny.

Pro zjednodušení nebudeme v našem případě žádné bonusy resp. malusy do odhadu výše ročního pojistného zahrnovat.

1.2 Odhadování cen konkurence

Znalost cen konkurence je jeden ze základních principů marketingu. Dává to společnostem možnost stanovit jejich vlastní cenu pod cenou konkurence a tím se dostat na lepší pozici na trhu. Představme si zákazníka, který má na výběr ze dvou nabídek pojištění lišících se pouze cenou. Je zřejmé, že si ve většině případů vybere tu levnější.

Ne vždy jsou ale ceny konkurence veřejně známé. To platí i v pojišťovnictví, kde žádná z pojišťoven nezveřejní své ceny, případně sazebníky ostatním pojišťovnám. V dnešní době, kdy je možnost uzavřít pojištění online, by se dalo namítnout, že stačí vyplnit formulář na internetu a cenu konkurence získat. To je sice pravda, ale představme si kolik možností a kombinací by to vyžadovalo. Vždyť jen značek a typů aut je neskutečně mnoho. Navíc takové srovnávače bývají na tyto případy připraveny. Pokud zjistí mnohokrát opakované vypočítávání ceny pro různé kombinace tarifních skupin, mohou stránku pro daného uživatele zablokovat.

Znalost podrobných cen konkurence v celé jejich šíři, tedy ve větším rozsahu než je reálné získat z veřejně dostupných zdrojů a hlavně schopnost tuto znalost získat, může být velmi významná konkurenční výhoda. Abychom mohli tuto znalost cen konkurence získat, pokusíme se sestavit si její model. V případě, že máme vstupních dat dostatek, neměl by být problém, vytvořit takový model, který bude data dobře popisovat. Výsledná odchylka by měla být malá a odhad spolehlivý. Pokud však budeme mít jen nízký počet vstupních dat, právě například ze srovnávače, nebude lehké sestavit přesný model. Může se stát, že díky nízkému počtu vzorků některé tarifní skupiny ve vstupních datech vůbec nebudou zastoupeny. V tomto případě je dobré pokusit se data konkurence něčím obohatit.

V této diplomové práci se nejdříve pokusíme odhadnout ceny konkurence standardním způsobem (zobecněným lineárním modelem), a to jak v případě, kdy máme k dispozici dostatek dat, tak v případě nedostatku. Po zjištění výsledné chybovosti, která jak předpokládáme bude vyšší u modelu s menším počtem vstupních dat, se pokusíme použít alternativní způsoby k vylepšení odhadu. Tyto alternativy mohou spočívat například v přidání vlastní ceny jako dodatečného vstupu při odhadu cen konkurence, a to na základě úvahy, že naše ceny a ceny konkurenta vykazují řadu podobností, jako je vyšší cena pro mladé řidiče nebo pro řidiče z Prahy. Poté se pokusíme ceny konkurence odhadnout metodou neuronových sítí. Jelikož nejsme obeznámeni, že by některá z pojišťoven v ČR tuto metodu používala, pokusíme se odhadnout strukturu cen konkurence tak, abychom docílili alespoň takových výsledků jako u zobecněných lineárních modelů.

Kapitola 2

Zobecněný lineární model

2.1 Specifikace modelu

Data dostupná pro výpočet modelu v neživotním pojištění pocházejí většinou z databází pojišťoven, jejich sazebníků nebo z veřejně dostupných dat. Data se skládají ze dvou typů proměnných. Prvním jsou tzv. spojité proměnné a druhým tzv. kategorické (diskrétní), které se ještě dále dělí na nominální a ordinální kategorické proměnné.

Spojitou proměnnou nazveme takovou proměnnou, jež nabývá libovolných hodnot z množiny $\mathbb{R}$. Kategorická proměnná pak obsahuje konečný počet variant hodnot. Ordinální kategorická proměnná je taková, jejíž kategorie jsou přirozeně uspořádané, lze je určitým způsobem seřadit, např.: objem válců motoru. Nominální kategorické proměnné mají naopak rovnocenné kategorie, které nelze seřadit, např.: pohlaví.

Abychom mohli model formulovat, je důležité si zvolit vysvětlovanou proměnnou a proměnné vysvětlující. V pojišťovnictví je často vysvětlovanou proměnnou např.: počet škod, výše škod nebo výše pojistného. Vysvětlujícími proměnnými pak jsou např.: pohlaví, věk pojistníka, místo bydliště, typ vozidla, atd.

K formulaci modelu nám poslouží regresní analýza. Nejjednodušší regresí je regrese lineární.

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \tag{2.1}$$

kde $x = \left(x_1, \ldots, x_p\right)^T$ je p dimenzionální vektor vysvětlujících proměnných, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ je vektor odhadovaných koeficientů, Y vysvětlovaná proměnná a ε gaussovský šum.

Ta ovšem vyžaduje, aby vysvětlovaná proměnná měla normální rozdělení, tj. $Y_i \sim \mathcal{N}\left(\mu_i, \sigma^2\right)$, což není vždy zaručeno. A tak je potřeba lineární regresi zobecnit.

2.2 Exponenciální rodina

U zobecněných lineárních modelů není potřeba aby $Y_i \sim \mathcal{N}\left(\mu_i, \sigma^2\right)$. Podmínkou zobecněných lineárních modelů je, aby rozdělení Y_i pocházelo z tzv. exponenciální rodiny.

Definice 2.2.1 (Exponenciální rodina)**.** *Nechť $y \in \mathcal{Y} \subset \mathbb{R}$ a $\theta = (\theta_1, ...\theta_m) \in \Theta \subset \mathbb{R}^m$. Řekneme, že hustota $f(y, \theta)$ patří do exponenciální rodiny, pokud ji lze zapsat ve tvaru*

$$f(y, \theta) = s(y)\, r(\theta) \exp\left\{\sum_{j=1}^{m} t_j(y)\, q_j(\theta)\right\},$$

kde $s(.)$, $r(.)$, $t_j(.)$ a $q_j(.)$, $j \in \widehat{m}$, jsou nějaké reálné funkce. [6]

Pro práci v zobecněných lineárních modelech je vhodné uvažovat hustoty tvaru

$$f(y, \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}, \tag{2.2}$$

kde $a(.)$, $b(.)$, $c(.)$ jsou specifické funkce pro dané rozdělení, $\phi \in \mathbb{R}$ a $a(\phi) > 0$.

2.3 Zobecnění lineární regrese

Pro nezávislé náhodné veličiny Y_i z exponenciální rodiny 2.2.1 platí:

- každá veličina Y_i závisí na jednom parametru θ_i
- rozdělení Y_i je stejného tvaru, a to

$$f(y_i, \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}. \tag{2.3}$$

Parametry $(\theta_1, ...\theta_m)$ zredukujeme na $\left(\beta_1, ...\beta_p\right)$, kde $p < m$, tak aby funkce $\mathbb{E}Y_i$ byla lineární kombinací β, tedy

$$g(\mathbb{E}Y_i) = \sum_{j=1}^{p} x_{ij}\beta_j = X_i^T\beta, \tag{2.4}$$

kde $X_i^T = \left(x_{i1}, ..., x_{ip}\right)$ je vektor vysvětlujících proměnných pro i-tého jedince, β vektor parametrů a g je diferencovatelná a monotónní funkce, kterou nazýváme funkcí spojovací (link function).

Gamma rozdělení

V pojišťovnictví se využívá převážně poissonovská regrese a gamma regrese. Poissonovská regrese je vhodná v případech potřeby predikce počtu pojistných událostí nebo škodní frekvence, gamma regrese zase na výpočet výše škodní události nebo ceny pojistného. V této práci byla použita právě gamma regrese.

Ověřme nyní, že gamma rozdělení $Y \sim Gamma(\alpha, \beta)$ spadá do exponenciální rodiny.

$$f(y, \alpha, \beta) = \frac{1}{\Gamma(\alpha)}\frac{1}{\beta^\alpha}y^{\alpha-1}\exp^{-\frac{y}{\beta}} \tag{2.5}$$

Klasický tvar gamma rozdělení (2.5) v zobecněném lineárním modelu nelze použít. Přepíšeme si ho tedy pomocí substituce $\alpha = \nu, \beta = \frac{\mu}{\nu}$, kde ν je známé, na tvar

$$f\left(y, \nu, \frac{\mu}{\nu}\right) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^\nu y^{\nu-1}\exp^{-y\frac{\nu}{\mu}} \tag{2.6}$$

A tento tvar upravujeme, dokud se nedostaneme na tvar (2.2).

$$
\begin{aligned}
f\left(y, \nu, \frac{\mu}{\nu}\right) &= \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} \exp^{-y\frac{\nu}{\mu}} \\
&= \exp\left\{\nu \ln \frac{\nu}{\mu} + \ln\left(\frac{y^{\nu-1}}{\Gamma(\nu)}\right) + y\left(-\frac{\nu}{\mu}\right)\right\} \\
&= \exp\left\{\nu \ln \nu - \nu \ln \mu + (\nu - 1) \ln y - \ln(\Gamma(\nu)) - \frac{y\nu}{\mu}\right\}
\end{aligned}
$$

Nyní už vidíme tvar exponenciální rodiny, kde

$$
\begin{aligned}
a(\phi) &= \frac{1}{\nu} \text{ je identita } \Rightarrow \phi = \frac{1}{\nu} \\
b(\theta) &= \ln(\mu) \\
\theta &= -\frac{1}{\mu} \Rightarrow b(\theta) = \ln\left(-\frac{1}{\mu}\right) \\
c(y, \phi) &= \nu \ln \nu + \nu \ln y - \ln y - \ln(\Gamma(\nu))
\end{aligned}
$$

Kanonickou spojovací funkcí bude

$$
\mathbb{E}Y = b'(\theta) = -\frac{1}{\theta} = \mu.
$$

Spojovací funkce (Link function)

Spojovací funkce používané v gamma regresi jsou nejčastěji

- inverzní $g(\theta) = \frac{1}{\theta}$
- logaritmická $g(\theta) = \ln \theta$
- identická $g(\theta) = \theta$

[9] V našem modelu budeme používat spojovací funkci logaritmickou.

Zobecněný lineární model

Souhrnně tedy můžeme říci, že zobecněný lineární model má tři základní složky [10]:

- vysvětlovanou proměnnou Y_i, nezávislou, s hustotami z exponenciální rodiny, 2.2.1.
- vektor parametrů $\beta = \left(\beta_1, ..., \beta_p\right)$ a p vysvětlujících proměnných $\mathbb{X} = \left(X_1, ..., X_p\right)$
- spojovací funkci g, která je diferencovatelná a monotónní, taková, že

$$
\eta_i = g(\eta_i) = X_i^T \beta \; a \; \eta_i = (\mathbb{E}Y_i)
$$

2.3.1 Koeficient determinace

Koeficientem determinace nazveme číslo

$$R^2 = 1 - \frac{S_{res}}{S_T}, \tag{2.7}$$

kde S_{res} je reziduální součet čtverců a S_T je kvadratická chyba vysvětlované proměnné.
Koeficient determinace nám tedy říká, kolik procent závislosti vysvětlované proměnné je model schopen objasnit. Koeficient determinace má však tendenci růst s počtem vysvětlujících proměnných. Proto se častěji užívá tzv. adjustovaný koeficient determinace. Ten upravuje koeficient determinace tak, aby výsledek nezávisel na počtu proměnných.

$$\overline{R^2} = 1 - \left(1 - R^2\right) \frac{N-1}{N-p-1}, \tag{2.8}$$

kde N je počet pozorování a p je počet proměnných v modelu. [11]

2.4 Korelace

Korelací nazveme závislost mezi dvěma proměnnými. Velikost korelace (míru závislosti) nám říká tzv. korelační koeficient.

Definice 2.4.1 (Koeficient korelace)**.** *Nechť X, Y jsou náhodné veličiny. Potom číslo*

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{VarX \cdot VarY}} \tag{2.9}$$

nazýváme koeficient korelace náhodných veličin X,Y.

Pro koeficient korelace platí, že

$$-1 \leq r(X,Y) \leq 1.$$

Pokud

$$r \in \{-1, 1\} \Rightarrow \text{Veličiny jsou plně korelované}$$
$$r = 0 \Rightarrow \text{Veličiny jsou zcela nekorelované.}$$

Každý korelační koeficient je třeba doplnit o tzv. interval spolehlivosti $(100 \cdot (1-\alpha))\%$. Ten nám udává, jak moc je náš odhad variabilní. α se pak nazývá koeficientem spolehlivosti a obvykle se užívá $\alpha = 0,05$ nebo $\alpha = 0,01$.

2.5 Testování hypotéz

Předpokládejme, že existují dvě vzájemně si konkurující hypotézy, hypotéza nulová H_0 a hypotéza alternativní H_A. Testem nulové hypotézy H_0 proti hypotéze alternativní H_A rozumíme rozhodovací proces založený na náhodném výběru $X_1, ...X_n$, na jehož základě zamítáme nebo nezamítáme nulovou hypotézu. Mohou nastat 4 možnosti:

- H_0 platí a nezamítáme H_0

- H_0 platí a zamítáme H_0 (chyba 1. druhu)
- H_A platí, H_0 neplatí a nezamítáme H_0 (chyba 2. druhu)
- H_A platí, H_0 neplatí a zamítáme H_0

Postavení H_0 a H_A není symetrické. Za H_0 tu hypotézu, jejíž neoprávněné zamítnutí je závažnější. Pokud totiž H_0 zamítneme uděláme chybu maximálně rovnu α (hladina významnosti testu). Pokud bychom H_0 přijali, může být chyba velká. Volíme tedy ze dvou výsledků:

- zamítáme H_0
- nezamítáme H_0, ale ani nepřijímáme

p-hodnotou nazveme minimální hodnotu α, pro kterou stále zamítáme hypotézu H_0.

2.6 Křížová validace

Křížová validace je metoda validace modelu, která slouží k ověření, zda daný matematický model dobře předpovídá data v praxi. Metoda je složena ze dvou fází, a to z fáze trénovací a testovací.
Nejdříve si veškeré dostupné sady dat rozdělíme do dvou skupin. Jednu skupinu nazveme množinou trénovací a druhou množinou testovací. [12]

Trénovací množiny

Pomocí trénovací množiny dat odhadujeme parametry matematického modelu, dokud optimálně neřeší daný problém, a to právě na množině trénovacích dat. Ve chvíli, kdy jsme nalezli optimální řešení na trénovací množině, vyzkoušíme model na množinách testovacích.

Testovací množiny

Námi zvolený matematický model s odhadnutými parametry spustíme na zbylé množiny dat a pozorujeme chyby predikce. Cílem je nalézt optimální model, který dobře popisuje veškerá data. Pokud se výsledky výrazně liší, obměníme trénovací množinu a opakujeme celý postup, dokud není model optimální na všech množinách.

Kapitola 3

Modely odhadu cen konkurence

Vstupní data

Data využívaná v této diplomové práci vycházejí z historických analýz struktur cen a škodních portfólií dvou středně velkých pojišťoven. Na základě těchto analýz byla synteticky vygenerována data, která jsou v práci použita, a to tak, aby zachovávala v maximální míře strukturu profilů klientů na trhu a obvyklou strukturu cen a zároveň neumožňovala identifikaci konkrétní pojišťovny. Pro větší přehlednost nazývejme tyto pojišťovny bílou a černou. Data z bílé pojišťovny jsou objemnější, obsahují několik sad dat, každá s desítkami až stovkami tisíc smluv. Naopak data z černé pojišťovny obsahují pouze jednu sadu o 2 000 smlouvách.

Pokusme se nejdříve standardním způsobem (zobecněný lineární model) nasimulovat model bílé pojišťovny. U bílé pojišťovny je k dispozici dostatek dat, tedy námi vytvořený model by mohl dobře odhadovat cenu konkurence.

Představme si ale také situaci, že se nacházíme v bílé pojišťovně a sjednáváme pojištění se zákazníkem. Známe jeho požadavky a parametry. Máme k dispozici náš model a sadu dat z konkurenční (černé) pojišťovny. Naši cenu můžeme vygenerovat kdykoli pro kohokoli, ale cenu černé pojišťovny známe jen pro konkrétních 2 000 příkladů. Pokusme se tedy i přes to sestavit obecný model černé pojišťovny pro odhad cen a stanovme chybu tohoto odhadu. Předpokládejme, že nedostatkem dat se zvýší výsledná odchylka od původních hodnot a tedy model nebude tak spolehlivý. Pokusme se následně prozkoumat možné kroky vedoucí k vylepšení odhadu, a to znalostí cen bílé pojišťovny.

Data z obou pojišťoven obsahují 11 proměnných. Jedná se o pohlaví a věk pojištěného, okres, přesné souřadnice bydliště a velikost jeho populace, frekvence placení pojistného, značku a typ vozidla, objem válců, výkon motoru a cenu pojistného. Přičemž za vysvětlovanou proměnnou volíme průměrné roční pojistné. Zbylé proměnné stanovíme proměnnými vysvětlujícími.

Aby bylo možné výsledky naší analýzy dobře porovnat, výše pojistného je přeškálována na průměrné roční pojistné v České republice v roce 2019, které bylo ve výši 2696 Kč [2].

3.1 Model bílé pojišťovny

3.1.1 Specifikace modelu

Jako trénovací množinu dat si vybereme sadu o 30 000 uzavřených smlouvách. Nejprve ověříme, že data neobsahují žádné chybějící hodnoty.

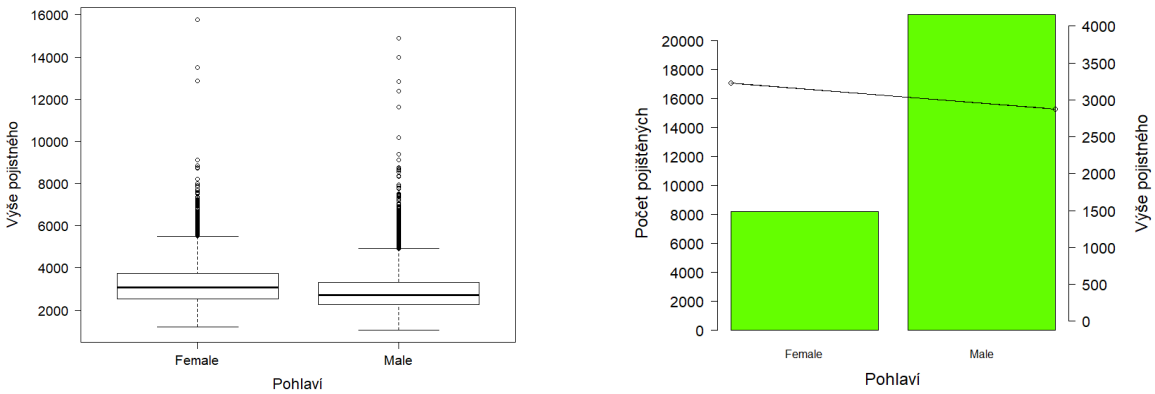
Proměnná přesných souřadnic polohy bydliště nám může posloužit pro vykreslení četnosti uzavřených pojistných smluv na mapě České republiky, ale do modelu ji zahrnovat nebudeme. V době, kdy data vznikala, bylo známo, že pojišťovny vytvářejí tarifní skupiny například podle okresů, velikostí místa bydliště a podobných méně granulárních kritérií. A tak i v našem případě budou použity ty namísto přesných souřadnic polohy bydliště. Proměnné kromě přesných souřadnic zvolíme jako vysvětlující a analyzujeme je.

Již dopředu víme, že některé proměnné bude potřeba převézt na kategorické (tarifní) proměnné. Kategorie (tarifní skupiny) budeme volit tak, aby byly v modelu stále významné a zároveň aby jich nebylo příliš mnoho. To by pak mohlo způsobit přeparametrizování modelu. Prvotní kategorizaci provedeme pomocí vykreslení hodnot. Finální kategorizaci provedeme pomocí významnosti/nevýznamnosti v zobecněné lineární regresi.

3.1.2 Analýza proměnných

Pohlaví

První proměnnou, kterou data obsahují je pohlaví pojištěného. Zde není žádný větší problém, vytvoříme dvě kategorie - muže a ženy. Při vykreslení box plotů a histogramu závislosti ceny na pohlaví zjišťujeme, že je více pojištěných mužů než žen, ale průměrné roční pojistné je vyšší pro ženy. Viz obrázek 3.1. Ač-

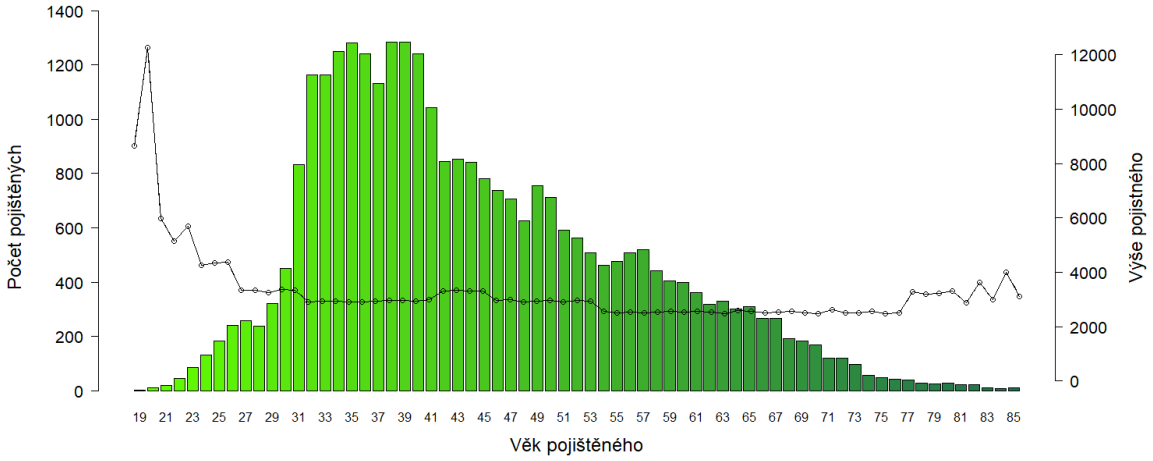


Obrázek 3.1: Grafy závislosti ceny na pohlaví pojištěného

koli je od konce roku 2012 je zakázáno v pojišťovnictví rozlišovat klienty podle pohlaví a je povinnost vyměřovat mužům i ženám stejnou sazbu, budeme se v našem případě držet dat, která mohou pocházet z dřívějších let a proměnnou do modelu zahrneme.

Věk pojištěného

Druhou a důležitou proměnnou je věk pojištěného. Histogram věku v závislosti na pojistném můžeme vidět na obrázku 3.2.



Obrázek 3.2: Graf závislosti ceny na věku pojištěného

Vidíme, že mladí pojištění mají vyšší cenu ročního pojistného a stejně tak řidiči v důchodovém věku. Předpokládáme, že je u mladých řidičů je vyšší pravděpodobnost nehod, a tím i vyšší pojistné, způsobená tím, že čerstvě získali řidičský průkaz a nemají tedy tolik zkušeností. U řidičů důchodového věku je vyšší nehodovost způsobena možným zhoršením pozornosti nebo pomalejšími reakcemi. Nejnižší cenu mají překvapivě řidiči ve věku kolem 60 let.

Proměnnou věku je třeba zkategorizovat. To provedeme po intervalech odpovídajícím skokům ceny z obrázku 3.2, tedy např.: 18 - 23 let, 24 - 26 let, atd. a znovu vykreslíme, obrázek 3.3.
Četnost nejmladší skupiny je sice malá, ale rozdíl ceny oproti ostatním kategoriím je významný, téměř 2 000 Kč, proto kategorii 18 - 23 let ponecháme. Podobně tomu je ve skupině 24 - 26 let, kde je rozdíl kolem 1 000 Kč.

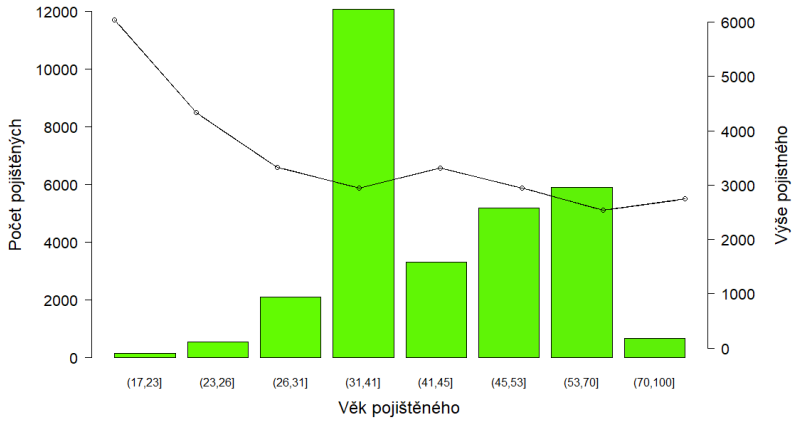
Dále vidíme, že pojištění ve věku 27 - 31 let a 42 - 45 let mají podobnou jak četnost, tak průměrnou výši ročního pojistného, proto tyto kategorie můžeme sloučit. Podobně tomu je u kategorií 32 - 41 let a 46 - 53 let. V těchto kategoriích je také nejvíce pojištěných, což se dá předpokládat, protože právě v tomto věku si nejčastěji lidé pořizují vlastní motorové vozidlo.

Ačkoli průměrné roční pojistné pro věk 71 - 100 let se podobá ceně pro 53 - 70 let a vzorků v této kategorii je málo, tyto kategorie zatím slučovat nebudeme. V reálném životě totiž nemůžeme říci, že 53-ti letý člověk řídí stejně jako 80-ti letý.

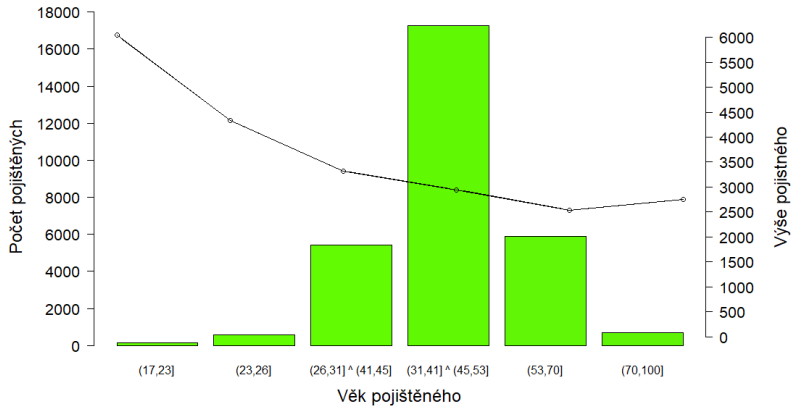
Výsledné kategorie věku z prvotní analýzy vidíme na obrázku 3.4.

Frekvence placení pojistného

Proměnná frekvence placení pojistného nám říká, v jakých intervalech pojištěný platí pojistné. Nejčastěji se sjednává roční placení pojistného, ale existuje i půlroční, čtvrtletní nebo měsíční. Proměnná je



Obrázek 3.3: Graf závislosti ceny na intervalech věku pojištěného



Obrázek 3.4: Graf závislosti ceny na upravených intervalech věku pojištěného

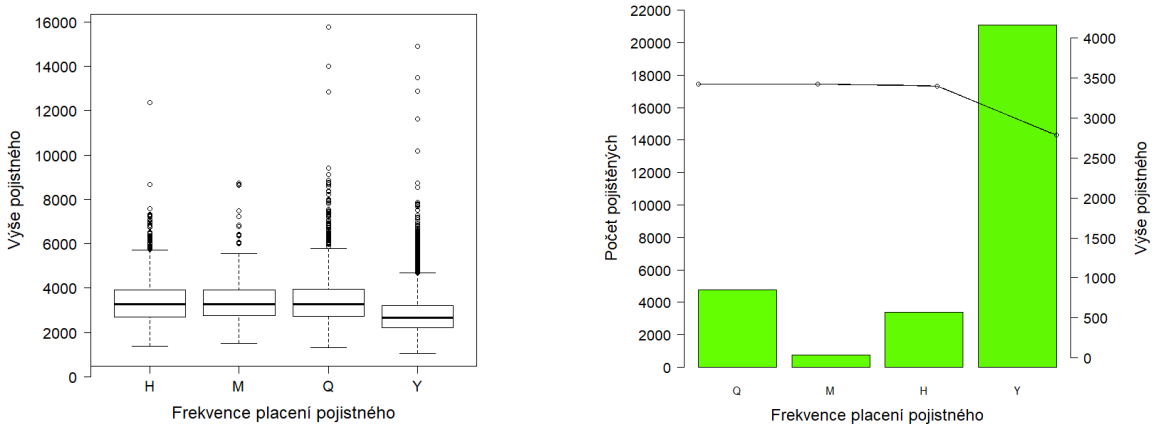
vykreslená na obrázku 3.5.

Vidíme, že nejvíce máme opravdu smluv s frekvencí ročního placení. Zároveň tyto smlouvy jsou přibližně o 20% levnější než smlouvy uzavírané půl ročně (H), čtvrtletně (Q) a měsíčně (M). Smlouvy uzavírané s častějším placením pojistného než je jeden rok, sloučíme, jelikož mají jak podobný počet vzorků, tak výši pojistného, obrázek 3.6.

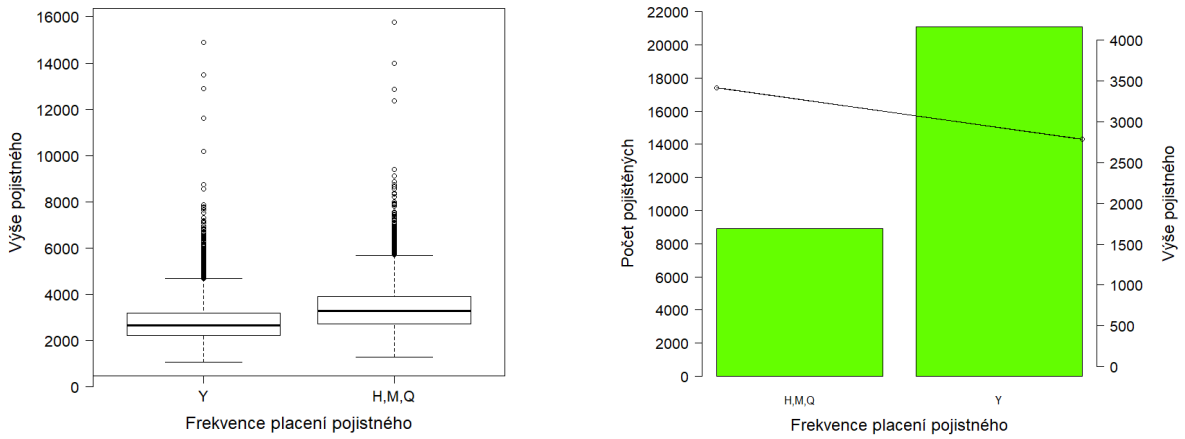
Objem válců

Pro větší přehlednost si proměnnou objemu válců nejprve převedeme do jednotek dm^3. Dopředu předpokládáme, že motorová vozidla s větším objemem válců budou mít vyšší pojistné, laicky řečeno čím silnější auto, tím vyšší pojistné.

Po vykreslení proměnné zjišťujeme, že neexistují přirozené kategorie, a proto kategorie (tarifní skupiny) musíme vytvořit. Nejdříve si tedy proměnnou rozdělíme na menší intervaly po 0,1 dm^3, abychom



Obrázek 3.5: Grafy závislosti ceny na frekvenci placení pojistného

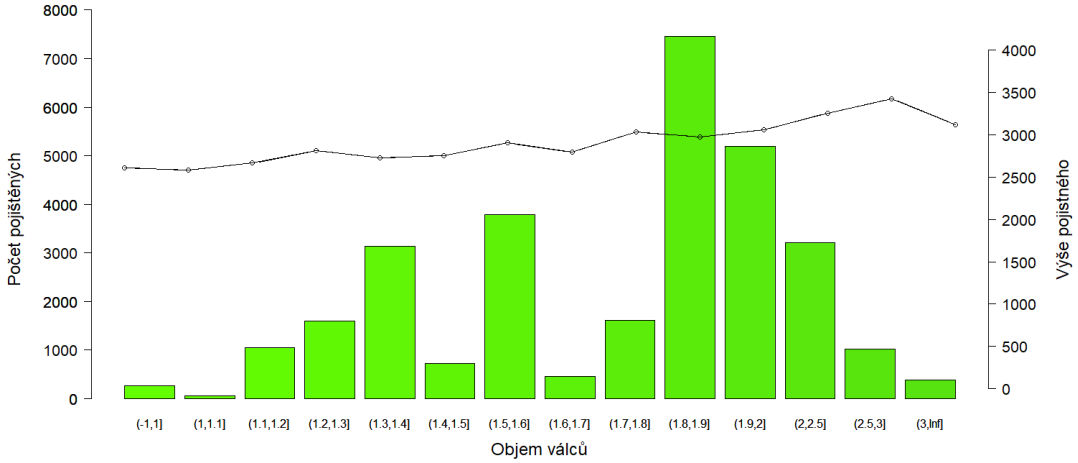


Obrázek 3.6: Grafy závislosti ceny na frekvenci placení pojistného po sloučení

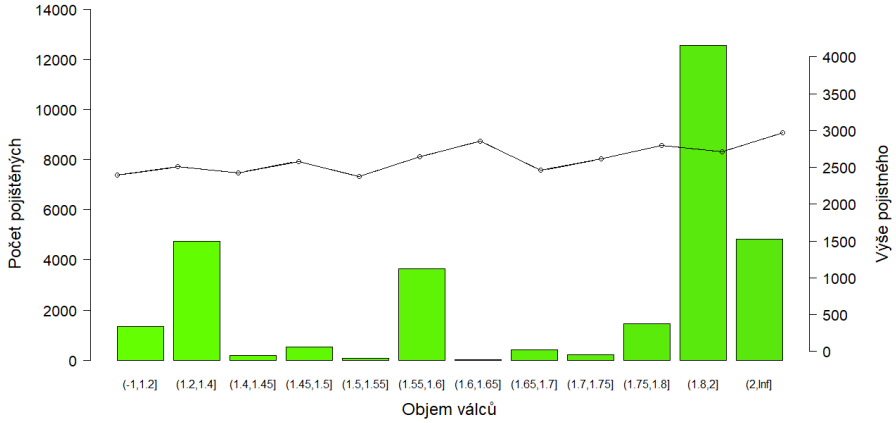
mohli zaznamenat cenové skoky a následně některé kategorie sloučit. Podrobnější intervaly se nachází na obrázku 3.7.

Z obrázku je vidět, že skoky nastávají u hodnot 1.2 a 2.0. Můžeme tedy říci, že motorová vozidla s objemem válců motoru do 1.2 dm^3 sloučíme do jedné kategorie a stejně tak motorová vozidla o objemu válců nad 2.0 dm^3. Dále vidíme pokles ceny pojistného v intervalu $(1.6, 1.7]$. Nic nenaznačuje tomu, že by tato kategorie měla mít nižší pojistné, proto tento interval a intervaly s ním sousedící rozdělíme jemněji, abychom zjistili, čím je onen pokles způsoben, obrázek 3.8.

Zmenšením intervalů vidíme, že skok se nachází v hodnotách 1.55 a 1.75. Tyto hodnoty tedy budou hraniční pro intervaly. Celkově nám tak vznikne 5 kategorií, obr. 3.9, které budeme považovat za finální vstup do modelu.



Obrázek 3.7: Graf závislosti ceny na intervalech objemu válců motoru



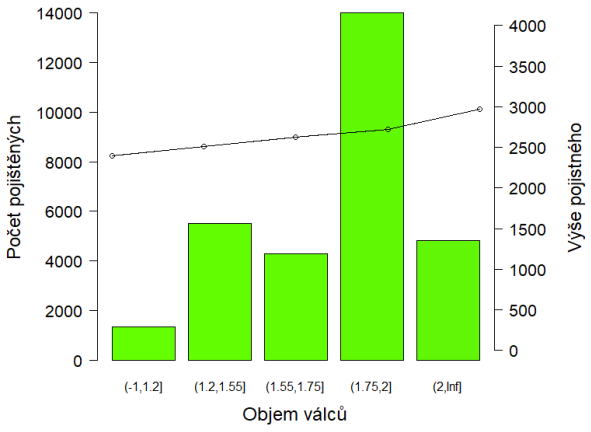
Obrázek 3.8: Graf závislosti ceny na upravených intervalech objemu válců motoru

Výkon motoru

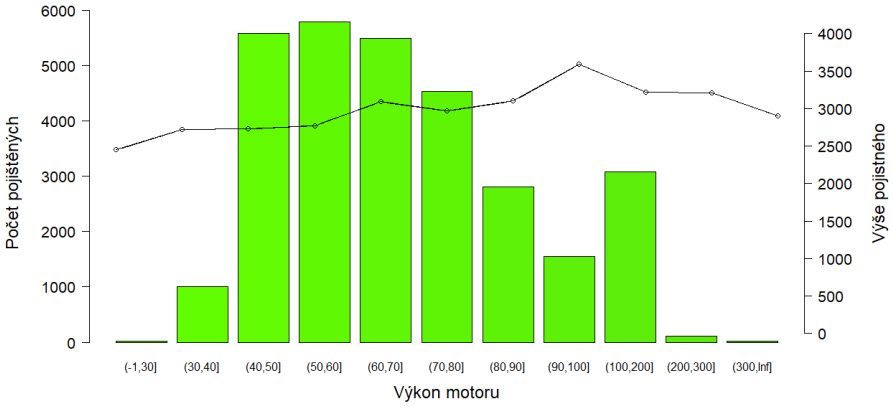
Z praxe víme, že výkon motoru úzce souvisí s objemem válců. Obecně se dá říci, že čím vyšší výkon motoru, tím větší objem válců. Je tedy nutné počítat s možnou korelací mezi těmito proměnnými, a tu také ověřit. Nejdříve ale proměnnou opět zkategorizujme a ke korelaci se vrátíme později.

Z obrázku 3.10 Vidíme, že v 60 kW máme první výrazný skok průměrného pojistného, a to i s ohledem na četnost. Po 60 kW máme lehký pokles. Nejvyšší cenu mají vozidla s výkonem motoru mezi $90 - 100kW$. Opět se podíváme na interval $(60, 90]$ kW podrobněji, 3.11.

Vidíme, že nejvýraznější hranice je mezi intervaly $(60, 85]$ a $(85, 90]$. Výsledné vstupní intervaly do modelu budou následovné, obrázek 3.12.



Obrázek 3.9: Graf závislosti ceny na intervalech objemu válců motoru po sloučení

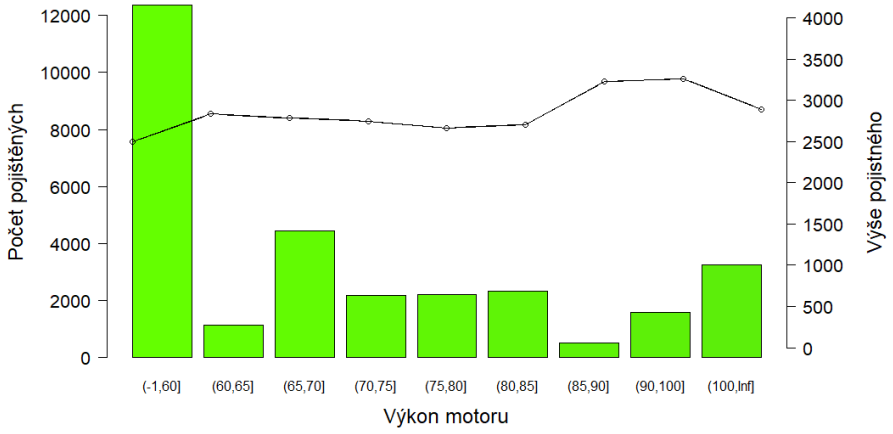


Obrázek 3.10: Graf závislosti ceny na intervalech výkonu motoru

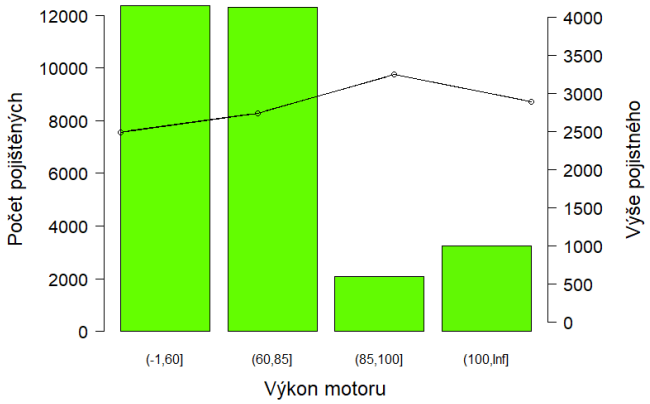
Velikost místa bydliště

Tato proměnná nám ukazuje, z jak velké obce či města pojištěný pochází. Výše pojistného se totiž jistě bude lišit pro řidiče žijící a jezdící v Praze a pro řidiče pocházející z malé vesnice. Statisticky je ve městech vyšší pravděpodobnost nehod než na vesnici a tedy i cena pojištění bude s vyšším počtem obyvatel růst. Zároveň ale nepředpokládáme, že by proměnná velikosti místa bydliště mohla být závislá na okresu, jelikož v okresech se nachází jak větší města, tak malé vesnice. Jedinou možnou závislostí je okres Praha, případně okresy větších měst jako je Brno, Ostrava či Plzeň. Proměnnou si vykreslíme a zkategorizujeme podle skoků, obrázek 3.13.

Na obrázku 3.13 je vidět, že v některých případech s nárůstem počtu obyvatel pojistné klesá. Nic ovšem nenasvědčuje tomu, že by v menších městech mělo být vyšší pojistné. Rozdělíme tedy data podle skoků a také podle geografického rozdělení na: vesnice (do 3 000 obyvatel), menší města (do 25 000 obyvatel),



Obrázek 3.11: Graf závislosti ceny na upravených intervalech výkonu motoru

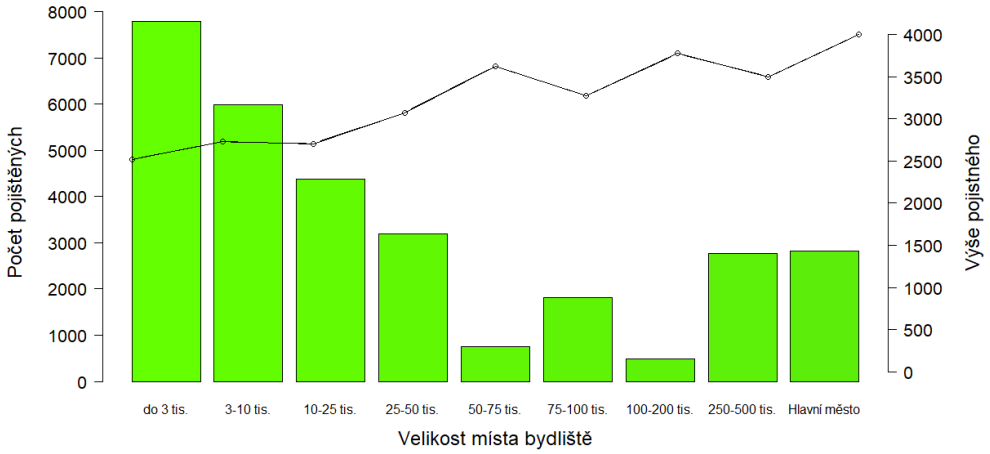


Obrázek 3.12: Graf závislosti ceny na intervalech výkonu motoru po sloučení

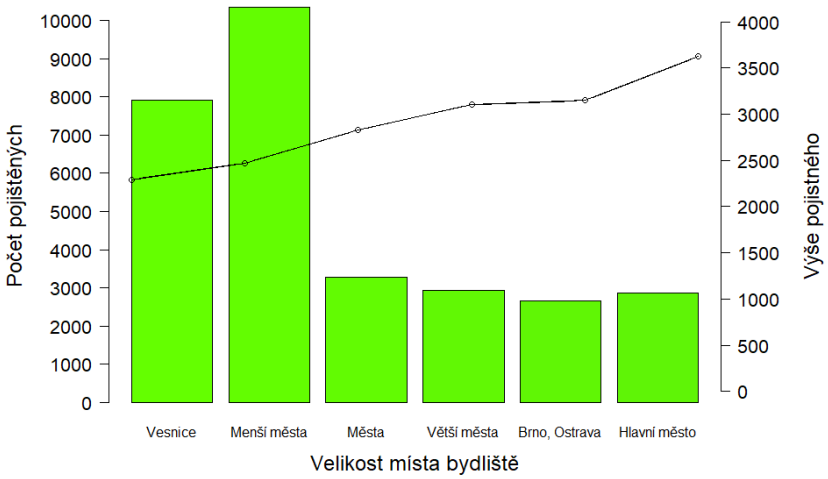
města (do 50 000 obyvatel), větší města (do 250 000 obyvatel), Brno a Ostrava (do 500 000 obyvatel) a Praha (nad 500 000 obyvatel), obrázek 3.14.

Okres bydliště

Další důležitou proměnnou pro odhad pojistného je okres bydliště pojištěného. Je zřejmé, stejně jako tomu bylo i u velikosti místa bydliště, že některé okresy (např.: Praha-střed, Praha-západ, ...) budou mít vyšší pravděpodobnost nehod, a tedy i vyšší pojistné než jiné. Jednotlivé okresy a jejich průměrné roční pojistné si vykreslíme na mapě České republiky, obrázek 3.15.



Obrázek 3.13: Graf závislosti ceny na velikosti místa bydliště

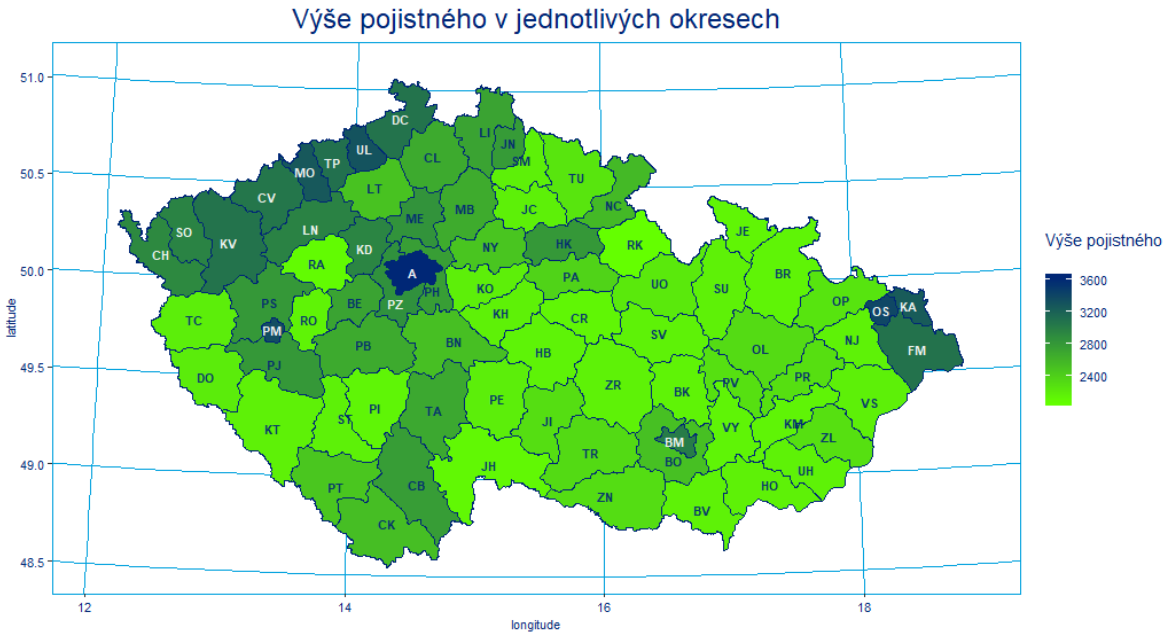


Obrázek 3.14: Graf závislosti ceny na velikosti místa bydliště po sloučení

Předpokládali jsme, že ceny budou vyšší v okresech větších měst. Co je ale zajímavé, je zvýšení ceny v okresech, jimiž prochází nějaká významná řeka České republiky (České Budějovice, Plzeň, . . .).

Tato skutečnost je dána historicky. V 18. století byly tyto cesty hojně využívány, jednalo se o hlavní tahy z Prahy do tehdejšího Rakouska-Uherska a Německa. Což platí i dnes, cesta přes Tábor a České Budějovice je hlavním tahem do Lince, přes Plzeň a Rozvadov se jezdí na jih Německa a z Prahy na sever do Drážďan či Berlína. Vyšší pojistné můžeme vidět také na Karvinsku a v Krušných horách.

Pokud bychom nechali každý okres v modelu samostatně, nejen že by byl model přeparametrizován,



Obrázek 3.15: Mapa průměrného pojistného v jednotlivých okresech ČR

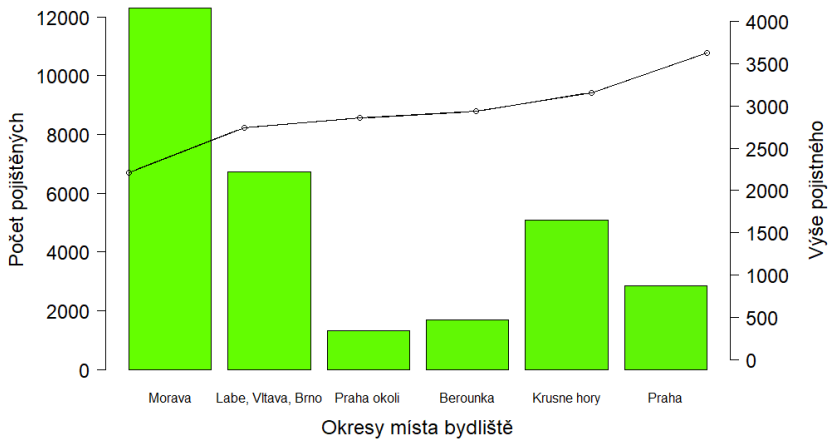
ale taktéž by byly některé okresy vůči sobě nevýznamné. Proto jednoduchým zobecněným modelem (pojistné $\sim$ *glm*(okres)) postupně přeskupíme a pospojujeme okresy do větších oblastí. Vznikne nám tak 6 oblastí:

- **Praha** - hl. město Praha
- **Praha okolí** - Kladno, Mělník, Praha západ a Praha východ
- **Krušné hory** - všechny okresy od Chebu až po Děčín, plus Ostrava, Karviná a Frýdek Místek
- **Berounka** - Plzeň a její okolí, Příbram, Beroun
- **Labe, Vltava a Brno** - od Náchoda před Nymburk do Liberce, z Prachatic přes České Budějovice do Benešova, Brno okolí a Brno město
- **Morava** - zbytek okresů - od Jižní Moravy do Jeseníků, Vysočina, Šumava, Krkonoše, Orlické hory, Rakovník a Rokycany

Průměrné ceny a četnost pojištěných v jednotlivých oblastech můžeme vidět na obrázku 3.16. Tyto oblasti budou vstupními tarifními skupinami proměnné okresu do modelu.

Značka a typ vozidla

Posledními proměnnými, které v našem modelu využijeme, jsou značka a typ vozidla. Tyto dvě proměnné spolu samozřejmě souvisí, každá značka má své originální vozy, které se liší právě typem. Ačkoli bychom mohli vzít proměnnou značky vozidla a udělat analýzu a kategorizaci pouze pomocí ní, v praxi to tak nefunguje. Vždyť každá značka vyrábí naprosto odlišné typy vozidel.



Obrázek 3.16: Graf závislosti ceny na oblastech okresů pojištěného

Vezměme si například Volkswagen. Volkswagen Polo je malé levnější auto, které si spíše koupí mladí lidi, Volkswagen Golf je pak větší auto vhodné pro rodiny, Volkswagen Passat je rozšířený mezi manažery a často je využíváno pro firemní účely, Volkswagen Transporter je dodávka nejčastěji užívaná na přepravu materiálu, atd. Všechny tyto auta jsou pod jednou značkou, ale rozhodně nemůžeme říci, že pojistné bude stejné.

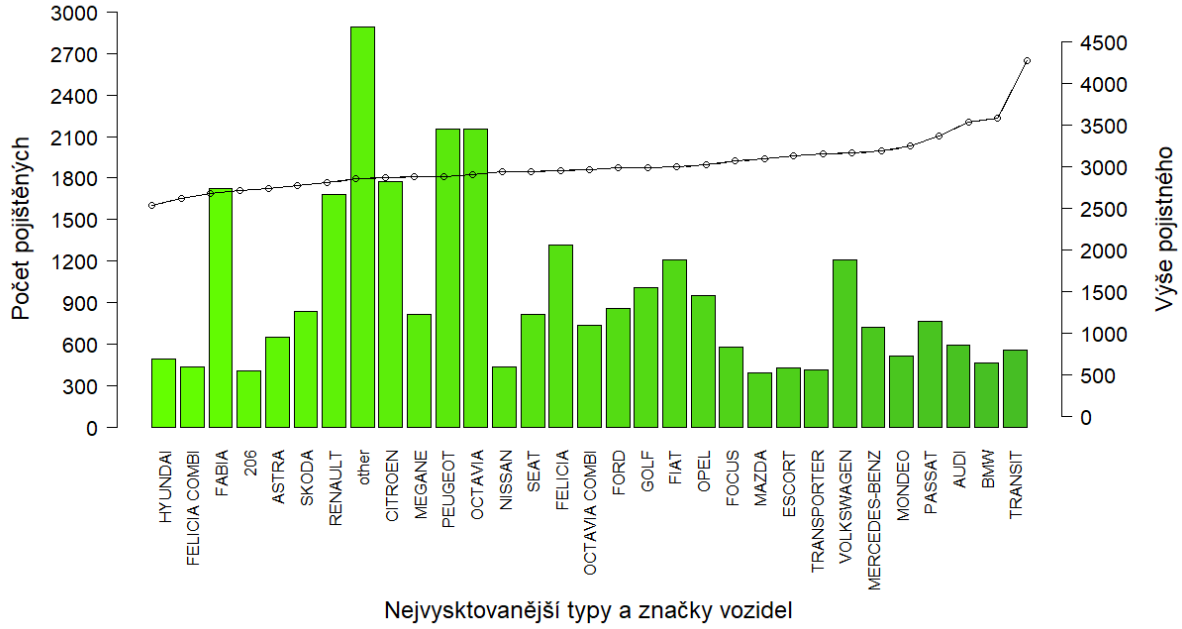
Pokud bychom udělali tarifní skupiny pomocí typů vozidla nikdy bychom se nedostali ke zdárnému výsledku. Jen naše data o 30 000 vzorcích obsahují přibližně 700 různých typů značek. Některé značky a typy mají navíc velmi malé procento zastoupení (např.: Fiat Ulysse), takže je nemá smysl považovat za samostatnou tarifní skupinu. Naopak jiné (např.: Škoda Octavia) plnohodnotně naplní samostatnou tarifní skupinu.

Budeme muset udělat nějaký kompromis. Tedy z dat nejprve určíme 15 nejvyskytovanějších typů vozidel a 15 nejvyskytovanějších značek. Jejich četnosti a průměrné pojistné můžeme vidět na obrázku 3.17. Zbylá vozidla, které nemají tak častý výskyt roztřídíme do 8 kategorií podle typu karoserie, tabulka 3.1.

1	Liftback
2	Sports, Roadster, Coupé
3	SUV
4	Sedan, Combi
5	Hatchback
6	Pick-up, Off-road
7	MPV
8	Dodávky

Tabulka 3.1: Kategorie vozidel podle karoserie

Toto rozdělení zbývajících značek a typů vozidel odpovídá našim potřebám, a to rozřadit auta podle toho



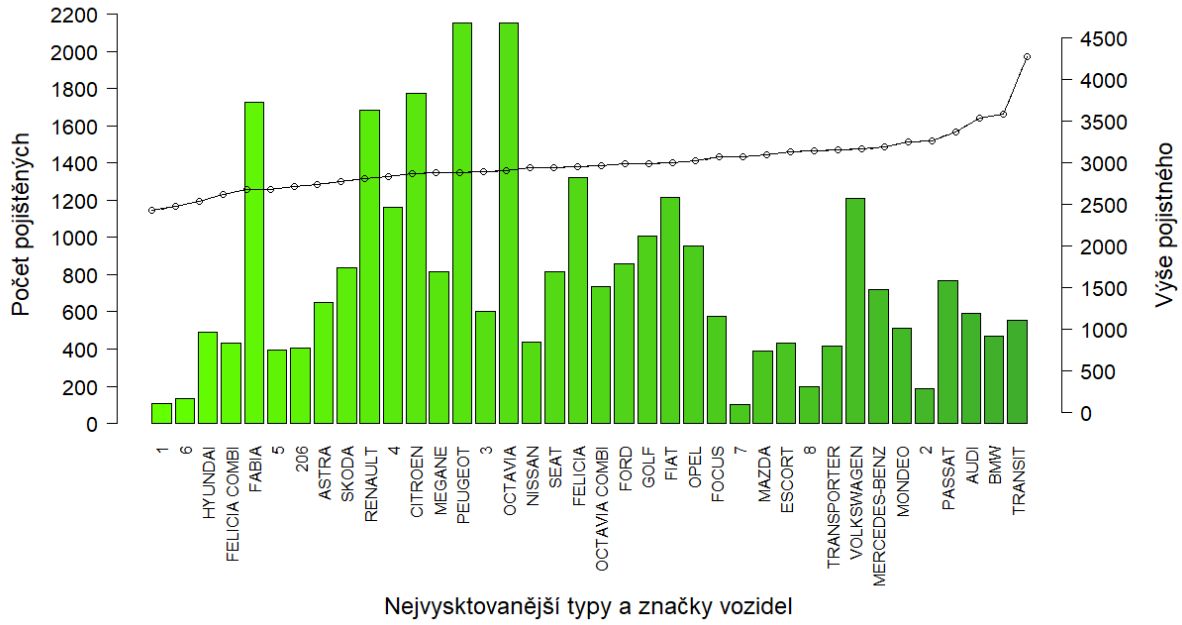
Obrázek 3.17: Graf závislosti ceny na 15-ti nejvyskytovanějších značkách a typech vozidla

k čemu jsou užívána a jaký typ člověka si dané vozidlo pořídí. Mladý svobodný řidič ho bude jistě užívat pro jiné účely než otec od rodiny. Vybere si tedy spíše hatchback, otec rodiny pak sáhne pro sedan, combi, či větší MPV.

Nyní obě rozdělení sloučíme, čímž nám vznikne 38 kategorií, obrázek 3.18. Vidíme, že nejnižší roční pojistné mají vozy s karoserií liftback, naopak nejvyšší cenu má Ford Transit. Vysokou cenu mají také luxusní vozy jako Audi, BMW nebo vozy sportovní. Opět pomocí jednoduchého zobecněného modelu se pokusíme kategorie zredukovat na menší počet.

Postupně slučujeme jednotlivé kategorie, jak podle významnosti v modelu, tak podle typu auta. Dostáváme 11 kategorií, které zahrneme do celkového modelu.

a) Transit
b) Audi, BMW
c) Mazda, Focus, Volkswagen, Mercedes-Benz, Mondeo, Escort, Passat a Sportovní auta
d) Dodávky, MPV
e) Ford, Fiat, Golf, Opel, Felicia
f) SUV
g) Citroen, Nissan, Peugeot, Octavia Combi, Megane, Seat a Sedan
h) Renault, Škoda, Fabia, Astra, 206 a Hatchback
i) Hyundai, Felicia Combi
j) Pick-up, Off-road
k) Liftback, starší auta



Obrázek 3.18: Graf závislosti ceny na 38 kategoriích vozidel

Nyní máme všechny proměnné zkategorizované a převedené na tarifní skupiny. Nicméně ještě před samotnou regresí je nutní ověřit, zda nejsou některé proměnné závislé.

3.1.3 Korelace mezi proměnnými

Korelační analýzu provedeme s proměnnými věku, velikosti místa bydliště, výkonu motoru a objemu válců motoru. Koeficienty korelace vidíme v tabulce 3.2. Věk a místo bydliště jsou zcela nezávislé vůči

	Věk	Velikost místa bydliště	Výkon motoru	Objem válců
Věk	1,00	0,00	0,00	0,00
Velikost místa bydliště	0,00	1,00	0,00	0,00
Výkon motoru	0,00	0,00	1,00	0,66
Objem válců	0,00	0,00	0,66	1,00

Tabulka 3.2: Tabulka koeficientů korelace pro vybrané proměnné bílé pojišťovny

ostatním proměnným. Výkon motoru a objem válců vykazují mezi sebou závislost s korelačním koeficientem 0,66. Je třeba ověřit, zda jsou obě proměnné v celkovém modelu významné. Pokud by nebyly, jednu proměnnou bychom do modelu nezahrnuli.

To provedeme deviačním testem. Porovnáme model, který obsahuje jen proměnnou objemu válců s modelem, který obsahuje jak objem válců, tak výkon motoru.

Z výsledku deviační analýzy , výpis 3.1, můžeme říci, že je lepší model s oběma proměnnými. Je tomu tak díky p - hodnotě ($<$ 2e-16), která je menší než hladina významnosti (α = 0,05) a zbytkové deviaci, která je u modelu s oběma proměnnými nižší (178,96) než u modelu pouze s objemem válců (294,19).

Listing 3.1: Deviační analýza modelů bílé pojišťovny

```
Analysis of Deviance Table

Model 1: price ~ policyholder_m + age_m + inhab_num_city_m +  payment_frequency_m +
cubic_capacity_m + distr_m + car_segm
Model 2: price ~ policyholder_m + age_m + inhab_num_city_m +  payment_frequency_m +
cubic_capacity_m + engine_power_m + distr_m + car_segm
  Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
1     29942     294.19
2     29939     178.96  3   115.22 6547.9 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '␣' 1
```

Další problém, který by mohl nastat je singularita kategorie Praha v proměnné okres bydliště a Hlavního města v proměnné velikost místa bydliště. Vytvoříme tedy zobecněný lineární model závislosti pojistného na těchto dvou proměnných, abychom se přesvědčili o této domněnce, výpis 3.2.

Listing 3.2: Výsledky gamma regrese pro proměnné okres bydliště a velikost místa bydliště

```
Coefficients: (1 not defined because of singularities)
                                        Estimate Std. Error  t value Pr(>|t|)
(Intercept)                             7.976589   0.005913 1349.043  < 2e-16 ***
distr_mLabe, Vltava, Brno              -0.083657   0.006249  -13.388  < 2e-16 ***
distr_mMorava                          -0.248266   0.005882  -42.206  < 2e-16 ***
distr_mPraha okoli                      0.023180   0.008116    2.856  0.00429 **
distr_mKrusne hory                      0.042874   0.006359    6.742 1.59e-11 ***
distr_mPraha                            0.317680   0.007226   43.963  < 2e-16 ***
inhab_num_city_mMensi mesta             0.065622   0.003320   19.764  < 2e-16 ***
inhab_num_city_mMesta                   0.171341   0.004713   36.352  < 2e-16 ***
inhab_num_city_mVetsi mesta             0.230114   0.004834   47.604  < 2e-16 ***
inhab_num_city_mBrno, Ostrava           0.224080   0.005322   42.104  < 2e-16 ***
inhab_num_city_mHlavni mesto                  NA         NA       NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '␣' 1

(Dispersion parameter for Gamma family taken to be 0.04883339)

    Null deviance: 2373.3  on 29999  degrees of freedom
Residual deviance: 1299.0  on 29990  degrees of freedom
AIC: 468622

Number of Fisher Scoring iterations: 4
```

Jak jsme předpokládali, koeficient u kategorie Hlavního města nemohl být odhadnut kvůli singularitě. Kategorii Hlavní město sloučíme s Brnem a Ostravou a ponecháme kategorii Praha v proměnné okresech. To by nám mělo zajistit, že singularita vymizí a přitom zůstane významnost Prahy v modelu, kterou zajistí právě tarifní skupina Praha u proměnné okresů. Ověříme, zda jsou nyní všechny kategorie významné a jestli není potřeba sloučit některé další, výpis 3.3.

Vše je nyní v pořádku a můžeme spustit celý model.

Listing 3.3: Výsledky gamma regrese pro proměnné okres bydliště a velikost místa bydliště po sloučení Hlavního města s Brnem a Ostravou

```
Coefficients:
                                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)                               7.976589   0.005913 1349.043  < 2e-16 ***
distr_mLabe, Vltava, Brno                -0.083657   0.006249  -13.388  < 2e-16 ***
distr_mMorava                            -0.248266   0.005882  -42.206  < 2e-16 ***
distr_mPraha okoli                        0.023180   0.008116    2.856  0.00429 **
distr_mKrusne hory                        0.042874   0.006359    6.742 1.59e-11 ***
distr_mPraha                              0.093600   0.008400   11.143  < 2e-16 ***
inhab_num_city_mMensi mesta               0.065622   0.003320   19.764  < 2e-16 ***
inhab_num_city_mMesta                     0.171341   0.004713   36.352  < 2e-16 ***
inhab_num_city_mVetsi mesta               0.230114   0.004834   47.604  < 2e-16 ***
inhab_num_city_mBrno, Ostrava a
Hlavni mesto                              0.224080   0.005322   42.104  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '␣' 1

(Dispersion parameter for Gamma family taken to be 0.04883339)

    Null deviance: 2373.3  on 29999  degrees of freedom
Residual deviance: 1299.0  on 29990  degrees of freedom
AIC: 468622

Number of Fisher Scoring iterations: 4
```

3.1.4 Regresní model a odhad parametrů

K odhadu parametrů použijeme zobecněný lineární model s gamma rozdělením a logaritmickou spojovací funkcí. Parametry β odhadujeme na námi zvolené trénovací množině. Jako intercept volíme:

- Pohlaví: Muž
- Věk: 32 - 41 let a 46 - 53 let
- Frekvence placení pojistného: ročně
- Výkon motoru: 60-85 kW
- Objem válců motoru: 1.55 - 1.75 dm^3
- Velikost místa bydliště: Větší města
- Místo bydliště: Labe, Vltava, Brno
- Vozidlo: Citroen, Nissan Peugeot, Octavia, Octavia Combi, 4, Megane, Seat

Regresní model bílé pojišťovny s odhadnutými koeficienty můžeme vidět na výpise 3.4.

U proměnné objemu válců vidíme, že kategorie (-1, 1.2] je nevýznamná vůči interceptu, a tedy ji v modelu sloučíme s kategorií (1.55, 1.75]. Finální model s p-hodnotami a odhady koeficientů β je znázorněn na výpise 3.5.

Listing 3.4: Gamma regrese bílé pojišťovny

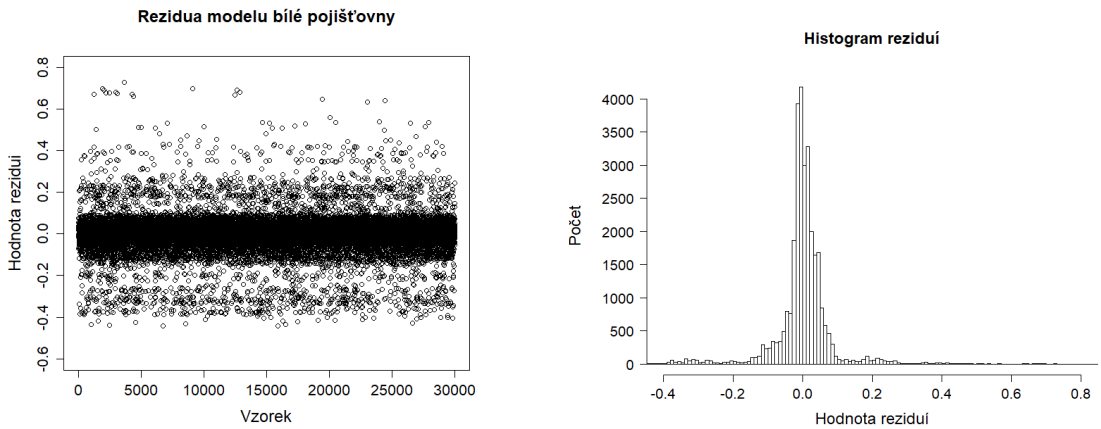
```
Coefficients:
                                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                                     7.986608   0.002249 3551.799  < 2e-16 ***
policyholder_mFemale                            0.116844   0.001038  112.544  < 2e-16 ***
age_m(17,23]                                    0.706350   0.006320  111.770  < 2e-16 ***
age_m(23,26]                                    0.380054   0.003453  110.076  < 2e-16 ***
age_m(53,70]                                   -0.144010   0.001207 -119.285  < 2e-16 ***
age_m(70,100]                                  -0.067676   0.003137  -21.571  < 2e-16 ***
age_m(26,31] ^ (41,45]                          0.115602   0.001246   92.776  < 2e-16 ***
inhab_num_city_mVesnice                        -0.228168   0.001751 -130.319  < 2e-16 ***
inhab_num_city_mMensi mesta                    -0.160613   0.001684  -95.402  < 2e-16 ***
inhab_num_city_mMesta                          -0.047376   0.002056  -23.046  < 2e-16 ***
inhab_num_city_mBrno, Ostrava a Hlavni mesto   -0.005034   0.002185   -2.304   0.0212 *
payment_frequency_mH,M,Q                        0.202187   0.001011  199.933  < 2e-16 ***
cubic_capacity_m(-1,1.2]                       -0.002395   0.002580   -0.928   0.3534
cubic_capacity_m(1.2,1.55]                      0.011823   0.001696    6.972 3.19e-12 ***
cubic_capacity_m(1.75,2]                        0.011273   0.001412    7.982 1.49e-15 ***
cubic_capacity_m(2,Inf]                         0.023055   0.001975   11.676  < 2e-16 ***
engine_power_m(-1,60]                          -0.089791   0.001179  -76.173  < 2e-16 ***
engine_power_m(85,100]                          0.156421   0.001969   79.448  < 2e-16 ***
engine_power_m(100,Inf]                         0.026663   0.001855   14.372  < 2e-16 ***
distr_mBerounka                                 0.085584   0.002264   37.805  < 2e-16 ***
distr_mMorava                                  -0.164517   0.001319 -124.764  < 2e-16 ***
distr_mPraha okoli                              0.105729   0.002429   43.531  < 2e-16 ***
distr_mKrusne hory                              0.128228   0.001509   84.953  < 2e-16 ***
distr_mPraha                                    0.178483   0.002195   81.310  < 2e-16 ***
car_segm1                                      -0.161836   0.007738  -20.915  < 2e-16 ***
car_segm3                                      -0.069312   0.003442  -20.134  < 2e-16 ***
car_segm6                                      -0.165301   0.006983  -23.672  < 2e-16 ***
car_segmTRANSIT                                 0.412061   0.003661  112.554  < 2e-16 ***
car_segmAUDI, BMW                               0.157651   0.002745   57.434  < 2e-16 ***
car_segmRENAULT, SKODA, FABIA, ASTRA, 206, 5   -0.009690   0.001423   -6.809 1.00e-11 ***
car_segmFORD, FIAT, GOLF, OPEL, FELICIA         0.065923   0.001402   47.024  < 2e-16 ***
car_segmVANS                                    0.070957   0.003230   21.970  < 2e-16 ***
car_segmHYUNDAI, FELICIA COMBI                 -0.081562   0.002784  -29.294  < 2e-16 ***
car_segmMAZDA, FOCUS, 2, VOLKSWAGEN, M-B,
MONDEO, ESCORT, PASSAT                          0.085865   0.001422   60.377  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '␣' 1

(Dispersion parameter for Gamma family taken to be 0.006402701)

    Null deviance: 2373.3  on 29999  degrees of freedom
Residual deviance:  201.6  on 29966  degrees of freedom
AIC: 412594

Number of Fisher Scoring iterations: 4
```

Listing 3.5: Gamma regrese bílé pojišťovny bez nevýznamných kategorií

```
Coefficients:
                                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                                     7.986213   0.002208 3617.157  < 2e-16 ***
policyholder_mFemale                            0.116851   0.001038  112.554  < 2e-16 ***
age_m(17,23]                                    0.706369   0.006320  111.774  < 2e-16 ***
age_m(23,26]                                    0.380061   0.003453  110.078  < 2e-16 ***
age_m(53,70]                                   -0.144012   0.001207 -119.287  < 2e-16 ***
age_m(70,100]                                  -0.067654   0.003137  -21.564  < 2e-16 ***
age_m(26,31] ^ (41,45]                          0.115602   0.001246   92.775  < 2e-16 ***
inhab_num_city_mVesnice                        -0.228180   0.001751 -130.329  < 2e-16 ***
inhab_num_city_mMensi mesta                    -0.160616   0.001684  -95.404  < 2e-16 ***
inhab_num_city_mMesta                          -0.047394   0.002056  -23.056  < 2e-16 ***
inhab_num_city_mBrno, Ostrava a Hlavni mesto   -0.005039   0.002185   -2.306   0.0211 *
payment_frequency_mH,M,Q                        0.202183   0.001011  199.931  < 2e-16 ***
cubic_capacity_m(1.2,1.55]                      0.012480   0.001540    8.101 5.64e-16 ***
cubic_capacity_m(1.75,2]                        0.011763   0.001310    8.982  < 2e-16 ***
cubic_capacity_m(2,Inf]                         0.023534   0.001906   12.350  < 2e-16 ***
engine_power_m(-1,60]                          -0.090030   0.001150  -78.272  < 2e-16 ***
engine_power_m(85,100]                          0.156353   0.001967   79.468  < 2e-16 ***
engine_power_m(100,Inf]                         0.026579   0.001853   14.344  < 2e-16 ***
distr_mBerounka                                 0.085589   0.002264   37.807  < 2e-16 ***
distr_mMorava                                  -0.164521   0.001319 -124.768  < 2e-16 ***
distr_mPraha okoli                              0.105729   0.002429   43.531  < 2e-16 ***
distr_mKrusne hory                              0.128220   0.001509   84.950  < 2e-16 ***
distr_mPraha                                    0.178494   0.002195   81.316  < 2e-16 ***
car_segm1                                      -0.162114   0.007730  -20.973  < 2e-16 ***
car_segm3                                      -0.069283   0.003442  -20.128  < 2e-16 ***
car_segm6                                      -0.165251   0.006983  -23.666  < 2e-16 ***
car_segmTRANSIT                                 0.412111   0.003661  112.580  < 2e-16 ***
car_segmAUDI, BMW                               0.157666   0.002745   57.440  < 2e-16 ***
car_segmRENAULT, SKODA, FABIA, ASTRA, 206, 5   -0.009873   0.001410   -7.004 2.54e-12 ***
car_segmFORD, FIAT, GOLF, OPEL, FELICIA         0.065959   0.001401   47.066  < 2e-16 ***
car_segmVANS                                    0.070995   0.003229   21.984  < 2e-16 ***
car_segmHYUNDAI, FELICIA COMBI                 -0.081505   0.002784  -29.278  < 2e-16 ***
car_segmMAZDA, FOCUS, 2, VOLKSWAGEN, M-B,
MONDEO, ESCORT, PASSAT                          0.085895   0.001422   60.413  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '␣' 1

(Dispersion parameter for Gamma family taken to be 0.00640271)

    Null deviance: 2373.3  on 29999  degrees of freedom
Residual deviance:  201.6  on 29967  degrees of freedom
AIC: 412593

Number of Fisher Scoring iterations: 4
```

Zbývá ověřit, zda náš model dobře popisuje data. Z gamma regrese, výpis 3.5, můžeme říci, že model s námi vytvořenými tarifními skupinami daleko lépe popisuje daná data, zbytková odchylka (201,6) je totiž výrazně nižší než odchylka nulová (2373,3).
Pokud se podíváme na $\chi^2_{0,95}$, zjistíme, že $\chi^2_{0,95}$ už pro 200 stupňů volnosti ($\chi^2_{0,95}(200) = 233,99$) je větší než naše deviační statistika (201,6), tedy $\chi^2_{0,95}$ pro 25 860 stupňů volnosti bude jistě větší. Hypotézu o správnosti modelu nezamítáme a můžeme říci, že náš model dobře popisuje data.

Dále si vykreslíme rezidua, obrázek 3.19 a pomocí Lillieforsova testu ověříme jejich normalitu. Na základě pohledu na 3.19 můžeme říci, že rezidua vypadají, že pocházejí z normálního rozdělení a nevykazují žádnou závislost. Testy normality však normalitu reziduí zamítají, a to díky p-hodnotě, která je nižší



Obrázek 3.19: Rezidua modelu bílé pojišťovny

Lillieforsův test	
D = 0,19282	p-hodnota < 2e-16

Tabulka 3.3: Lillieforsův test normality reziduí

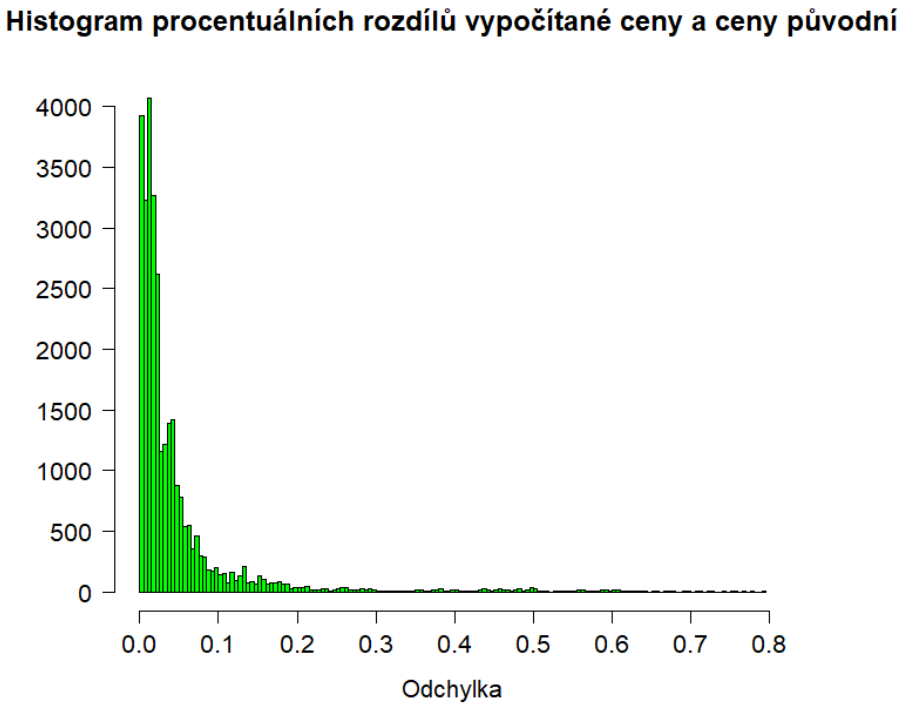
než hladina významnosti ($\alpha = 0,05$). Neznamená to nutně, že rezidua z normálního rozdělení opravdu nepocházejí, testy pro více než 1 000 vzorků jsou velmi citlivé.

3.2 Odhad ročního pojistného bílé pojišťovny

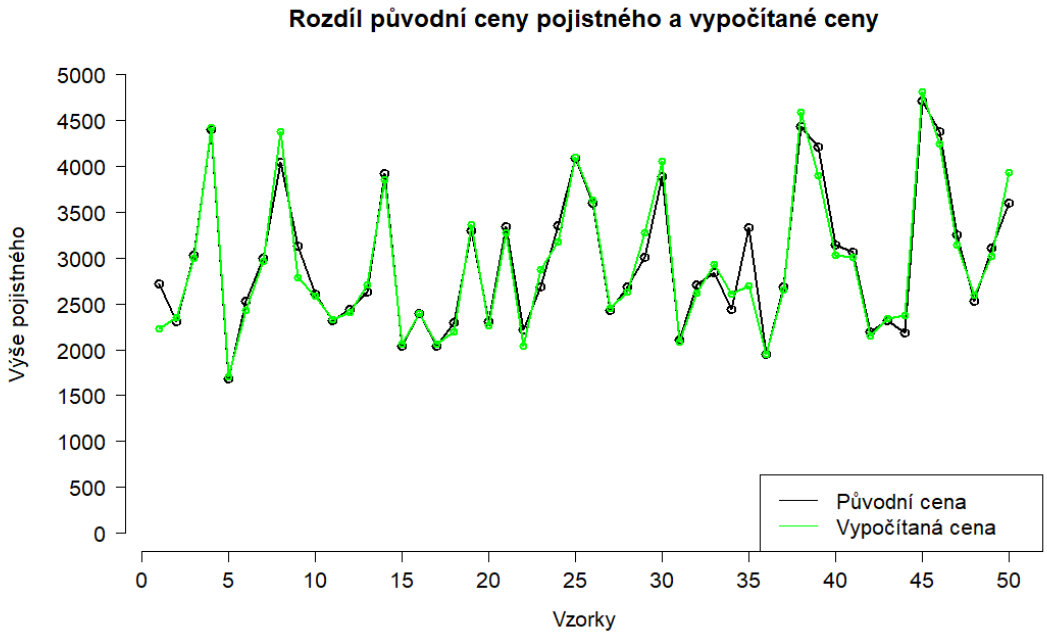
Koeficienty β, které jsme získali z našeho modelu použijeme k odhadu ročního pojistného bílé pojišťovny. Námi vypočítanou cenu porovnáme s cenou skutečnou. A to tak, že vypočítáme absolutní procentuální odchylku od původní ceny. Tedy

$$rozdil = abs\left(1 - \frac{\text{cena vypočítaná}}{\text{cena původní}}\right).$$

Díky tomu zjistíme, jak dobře jsme odhadli model bílé pojišťovny a jak dobře dokážeme určit její cenu. Histogram těchto rozdílů je znázorněn na obrázku 3.20.



Obrázek 3.20: Histogram procentuálních odchylek od původní ceny



Obrázek 3.21: Původní cena vs. cena vypočítaná

Na obrázku 3.21 je pak černě znázorněna originální cena a zeleně námi vypočítaná. Pro lepší názornost je vyňato pouze 50 vzorků. Můžeme vidět, že ve většině případů hezky kopírujeme původní cenu. Celkové výsledky pak nalezneme v tabulce 3.4, kde vidíme kolika procenty se náš odhad trefil do dané odchylky od originální ceny, např.: 77% našich odhadů se trefilo do odchylky 5% od originální ceny bílé pojišťovny.

Odchylka	Vzorky	% vzorků
do 5%	23 176	77,3%
do 10%	3 811	12,7%
do 15%	1 172	3,9%
do 20%	715	2,4%
do 25%	221	0,7%
do 30%	218	0,7%
do 35%	67	0,2%
do 40%	132	0,4%
do 45%	105	0,4%
do 50%	159	0,5%
nad 50%	224	0,7%

Tabulka 3.4: Celková chybovost modelu bílé pojišťovny na trénovací množině

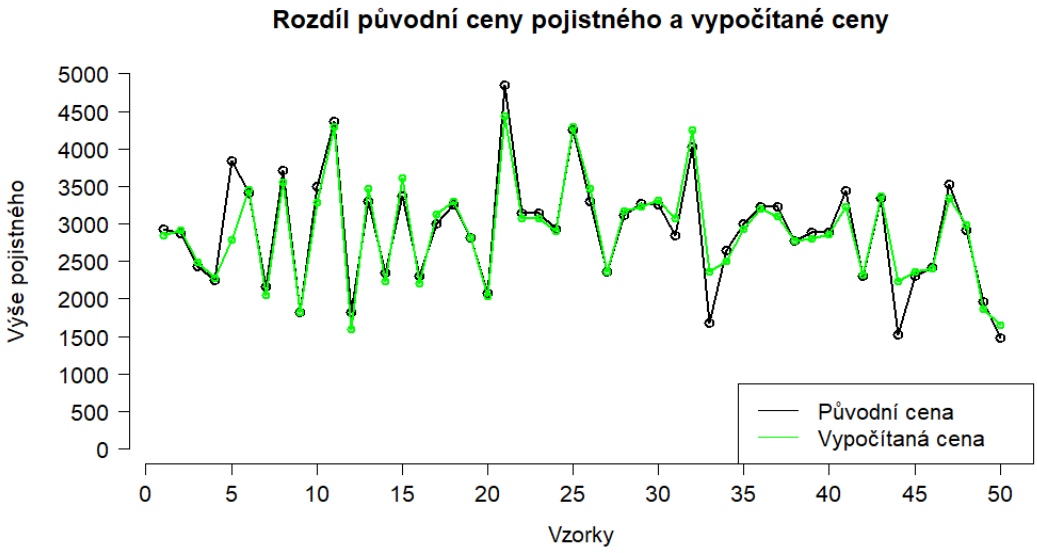
3.3 Validace modelu na testovacích množinách

3.3.0.1 Testovací množiny

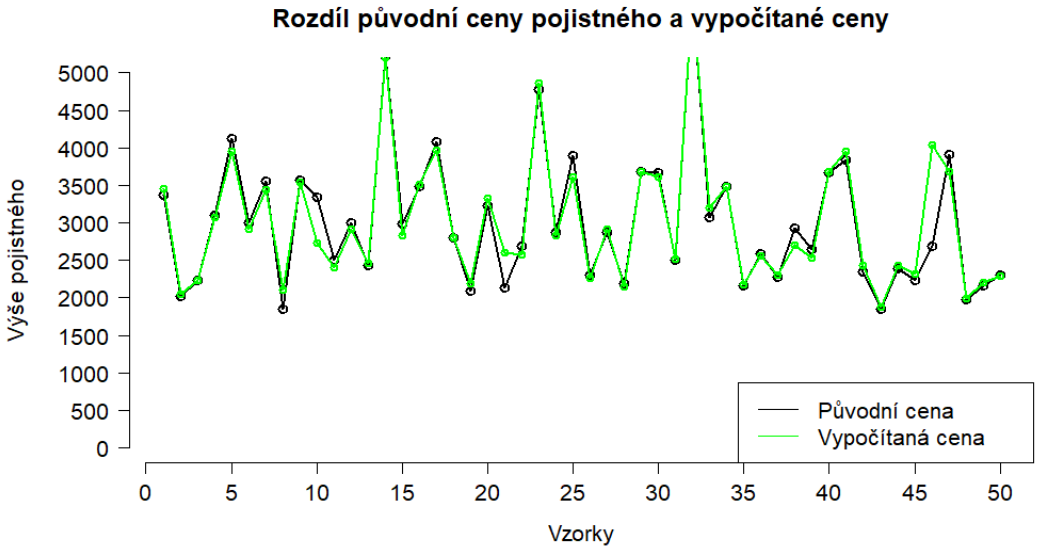
Koeficienty z modelu, výpis 3.5, nyní použijeme k výpočtu ročního pojistného na testovacích množinách. První testovací množina obsahuje taktéž 30 000 pozorování. Druhá množina, na které budeme model testovat, obsahuje 200 000 vzorků. Zjistíme tak, jestli počet vzorků v množinách má vliv na výslednou chybu odhadu. Námi vypočítanou cenu vs. cenu původní vidíme na 3.22, resp. 3.23. Tabulka 3.5 obsahuje celkové výsledky odchylek odhadů na testovacích množinách.

Odchylka	Vzorky test I	% vzorků test I	Vzorky test II	% vzorků test II
do 5%	23 304	77,7%	154 912	77,5%
do 10%	3 481	11,6%	23 876	11,9%
do 15%	1 331	4,5%	8 973	4,5%
do 20%	701	2,3%	4 681	2,3%
do 25%	216	0,7%	1 398	0,7%
do 30%	238	0,8%	1 414	0,7%
do 35%	85	0,3%	413	0,2%
do 40%	149	0,5%	909	0,5%
do 45%	63	0,2%	579	0,3%
do 50%	154	0,5%	1 047	0,5%
nad 50%	278	0,9%	1 798	0,9%

Tabulka 3.5: Celková chybovost modelu bílé pojišťovny na I. a II. testovací množině



Obrázek 3.22: Původní cena vs. cena vypočítaná - testovací množina I



Obrázek 3.23: Původní cena vs. cena vypočítaná - testovací množina II

Vidíme, že úspěšnost odhadu ceny je srovnatelná jako u trénovací množiny, a to jak u první testovací množiny, tak u druhé. Model prohlásíme za zvalidovaný, přičemž nebylo potřeba užít křížovou validaci.

Lze říci, že pokud máme dostatek vstupních dat konkurence, jsme schopni vytvořit model, který spolehlivě odhaduje strukturu cen bílé pojišťovny. Otázkou je, jak tomu bude v případě, že vstupních dat nemáme dostatek. Podívejme se nyní na druhou konkurenční společnost, kterou je černá pojišťovna.

3.4 Model černé pojišťovny

3.4.1 Specifikace modelu

Pokusme se nejdříve odhadnout strukturu cen černé pojišťovny pouze pomocí dostupných dat. Ty obsahují jednu sadu o 2 000 vzorcích, se stejnými proměnnými jako bílá pojišťovna. Tedy provedeme prvotní analýzu proměnných a vytvoření tarifních skupin, ověříme korelaci, vypočteme koeficienty a následně odhadneme ceny konkurence. Porovnáme s původními cenami a stanovíme spolehlivost odhadu.

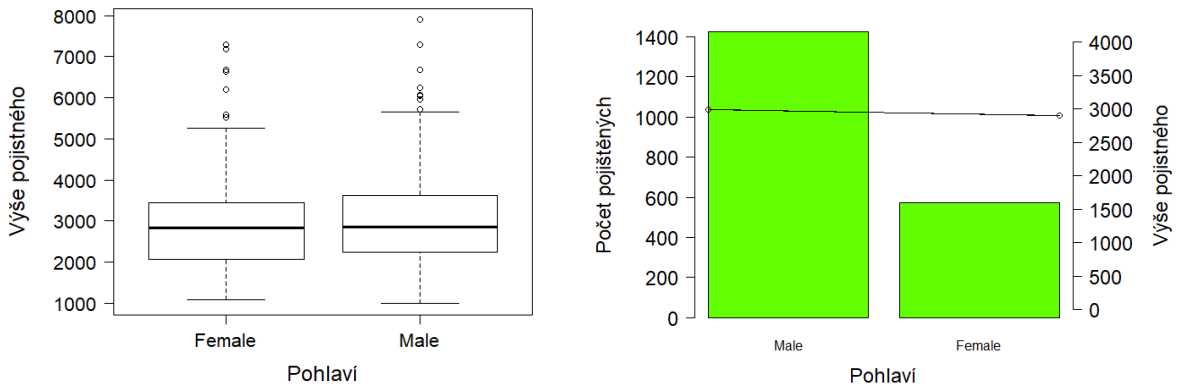
Předpokládáme, že kvůli nedostatku dat nebude možné odhadnout strukturu cen konkurence s takovou přesností, jakou bychom si představovali. V tomto případě bychom se pokusili námi vytvořený model vylepšit pomocí znalosti nějaké originální ceny, v našem případě originální ceny bílé pojišťovny.

Vysvětlovanou proměnnou zůstává výše pojistného, vysvětlující proměnné zvolíme stejně jako u bílé pojišťovny pohlaví, věk pojištěného, velikost místa bydliště, okres a informace o vozidle - síla motoru, objem válců motoru, značka a typ vozidla. Tyto proměnné zanalyzujeme.

3.4.2 Analýza proměnných

Pohlaví

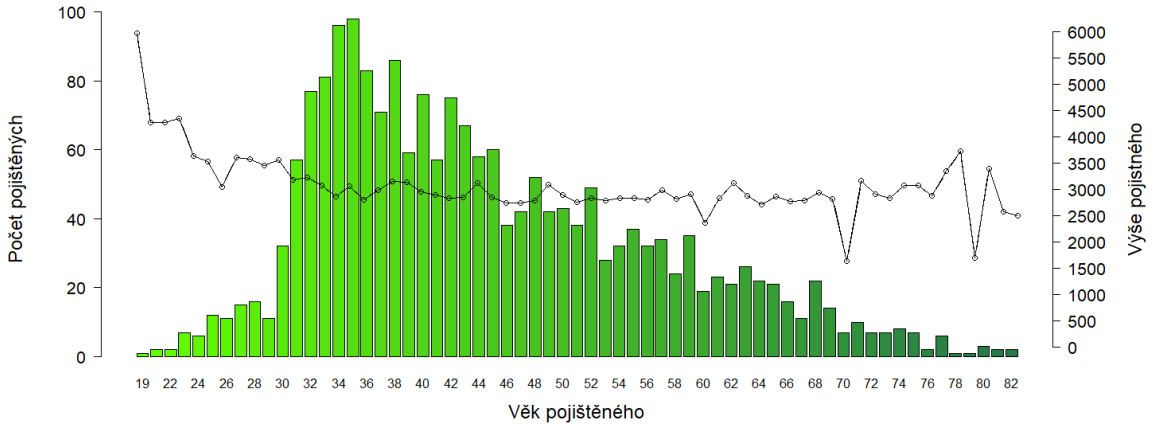
Proměnná obsahuje dva faktory - muže a ženy. Při vykreslení box plotů a histogramu, obrázek 3.24, zjišťujeme, že je opět sice více pojištěných mužů, ale na rozdíl od bílé pojišťovny je výše pojistného pro obě pohlaví srovnatelná. Ačkoli mají ženy ve vzorku významné zastoupení, můžeme říci, že výše pojistného na pohlaví nezávisí. Tuto proměnnou do modelu nezahrneme.



Obrázek 3.24: Grafy závislosti pojistného na pohlaví

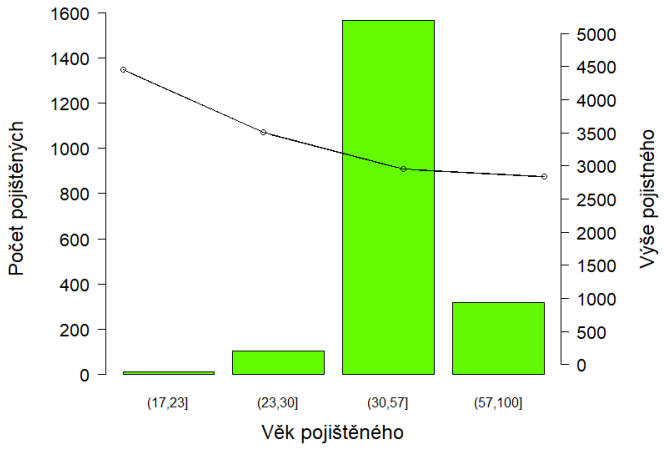
Věk pojištěného

Proměnnou věku pojištěného si vykreslíme po jednotlivých rocích v závislosti na pojistném, obrázek 3.25. Hranice tarifních skupin zde nejsou tak očividné, jako u bílé pojišťovny, přesto se pokusíme proměnnou zkategorizovat.



Obrázek 3.25: Graf závislosti ceny na věku pojištěného

Uvažujeme stejně jako u bílého modelu. Mladší a starší lidé budou mít průměrné pojistné vyšší než střední skupina, uděláme tedy hranice u 23. a 57. roku. Výraznější skok můžeme vidět ještě mezi lety 30 a 31. Tarifní skupiny 18 - 23 let, 24 - 30 let, 31 - 57 let a 58 - 100 let, které budeme považovat za finální vstup do modelu jsou zobrazeny na 3.26.

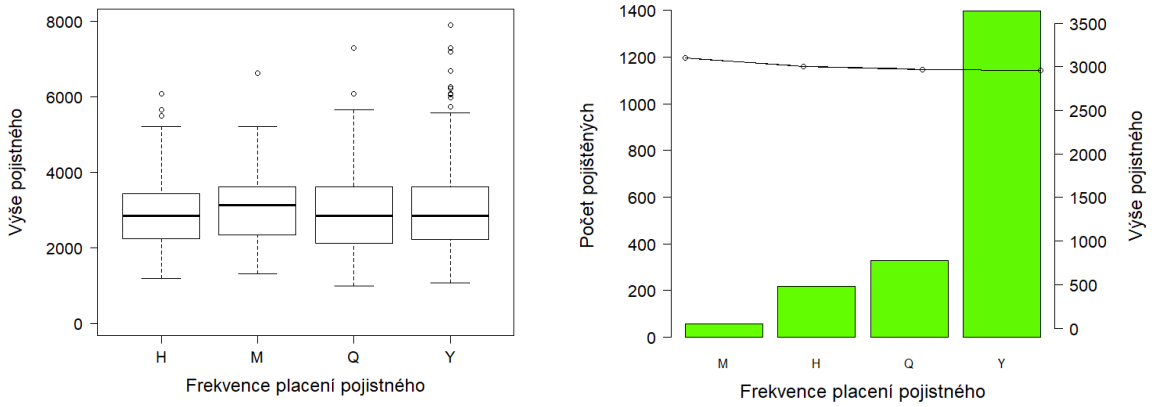


Obrázek 3.26: Graf závislosti ceny na věku pojištěného po intervalech

Frekvence placení pojistného

Proměnná obsahuje opět čtyři kategorie - roční placení pojistného, půlroční, čtvrtroční a měsíční. Při vykreslení box plotů a histogramu, obrázek 3.27, zjišťujeme, že měsíční placení pojistného má průměrné pojistné vyšší než ostatní kategorie, ale liší se jen nepatrně. Pokud se navíc podíváme na četnost, tak

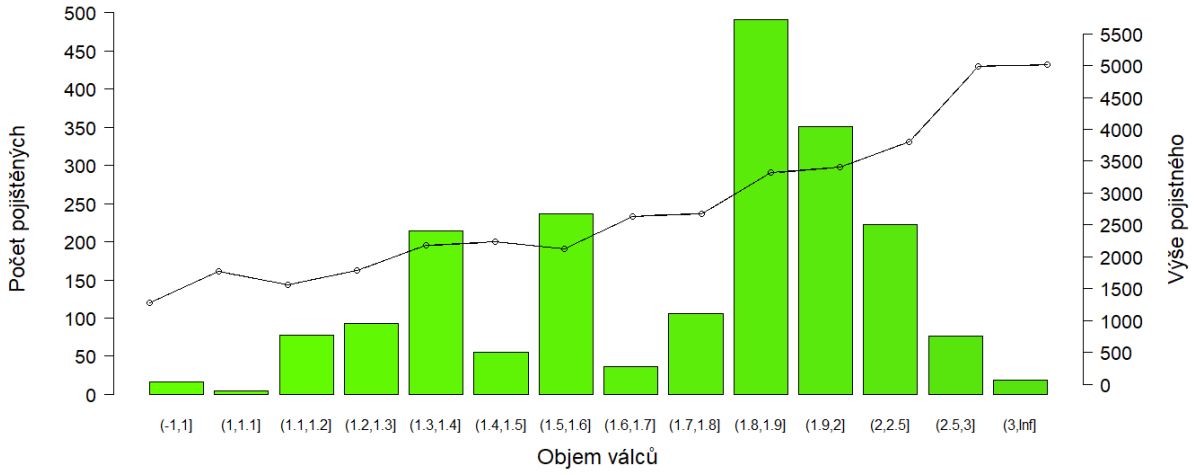
těchto případů je výrazně méně, něž případů ostatních. Nebudeme tedy faktory, ani proměnnou do modelu zahrnovat.



Obrázek 3.27: Grafy závislosti ceny na frekvenci placení pojistného

Objem válců

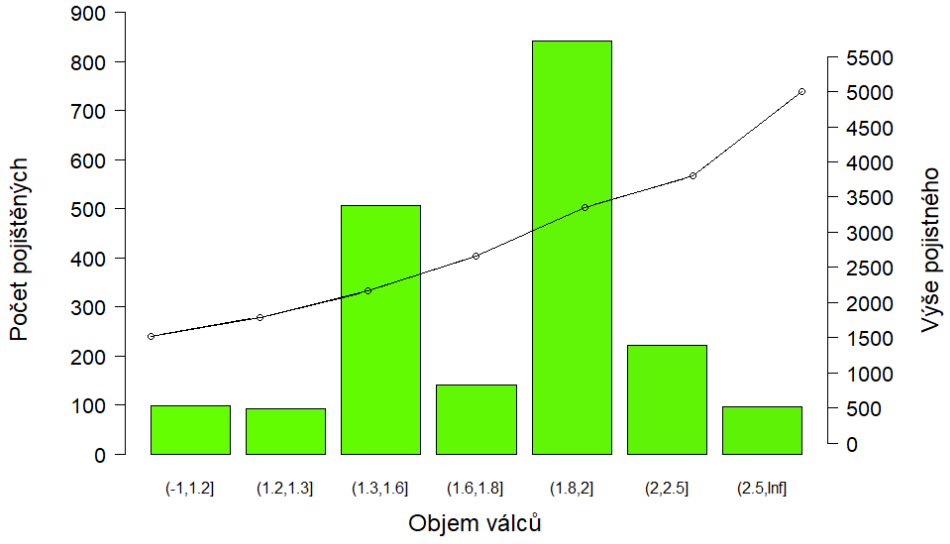
Převedeme proměnnou do jednotek dm^3 a vykreslíme intervaly po 0,1 dm^3, obrázek 3.28. Průměrné roční pojistné se nám u této proměnné pohybuje od 1 500 Kč do 5 000 Kč.



Obrázek 3.28: Graf závislosti ceny na objemu válců motoru

Z obrázku vyplývá, že první skok pojistného nastává u intervalu (1,1.1], není ale žádný zásadní důvod, proč by auta s objemem válců 1.1dm^3 měly mít vyšší pojistné než auta s 1.2 dm^3. Navíc četnost této

kategorie nedosahuje ani 20 vzorků, nebudeme ji tedy přikládat velkou váhu na samostatnou tarifní skupinu a auta do objemu válců 1.2 dm^3 sloučíme.
Dále vidíme podobnost ceny u vozidel s objemem válců od 1.3 dm^3 do 1.6 dm^3, od 1.6 dm^3 do 1.8 dm^3, od 1.8 dm^3 do 2.0 dm^3 a 2.5 dm^3 a výš. Výsledné tarifní skupiny jsou na obrázku 3.29.



Obrázek 3.29: Graf závislosti ceny na objemu válců motoru pro sloučené intervaly

Kategorií objemů válců v černé pojišťovně je více než u pojišťovny bílé, což může značit, že černá pojišťovna klade větší váhu na tuto proměnnou ve výpočtu pojistného pro své zákazníky.

Výkon motoru

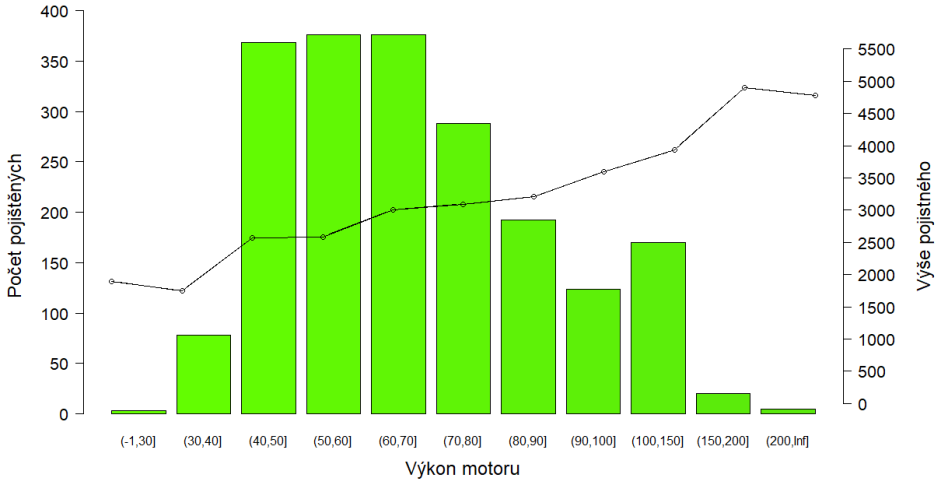
Stejně jako u objemu válců postupujeme i u výkonu motoru. Prvotní menší intervaly si zobrazíme na obrázeku 3.30. Vidíme skoky u 40 kW, 60 kW, 90kW a 150kW. Kategorie sloučíme, obrázek 3.31.

Opět je důležité si uvědomit, že proměnná výkonu motoru může souviset s objemem válců, a tedy je nutné počítat s možnou korelací. Korelaci proto před spuštěním finálního modelu ověříme.

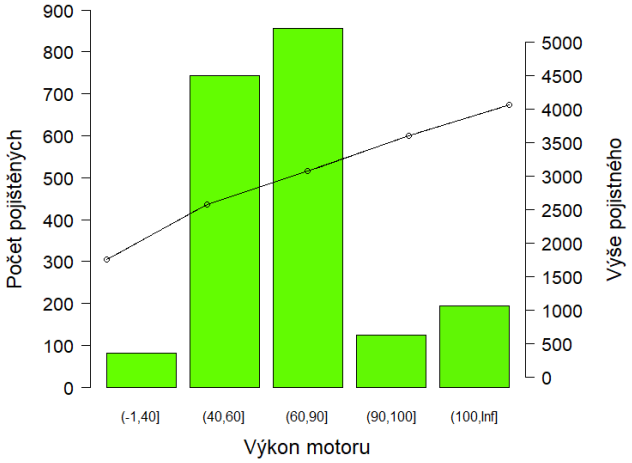
Velikost místa bydliště

Proměnnou velikosti místa bydliště si vykreslíme v menších intervalech, abychom správně rozpoznali změny výše pojistného. Opět předpokládáme, že se výše pojistného bude lišit pro zákazníka z vesnice a z města, ale i pro zákazníka většího města a menšího, obrázek 3.32.

Vidíme, že hranice nejsou stejné jako u bílé pojišťovny. Vesnice do 3 tis. obyvatel mají téměř stejnou výši průměrného pojistného jako menší města, a to až do 10 tis. Další tarifní skupina pak bude do 25 tis. obyvatel, do 75 tis, do 250 tis., do 500 tis. a hlavní město Praha, které má jediné v České republice nad půl milionu obyvatel. Brno a Ostravu necháváme zatím jako samostatnou kategorii (do 500 tis.). Námi vytvořené tarifní skupiny jsou zobrazeny níže na obrázku 3.33.

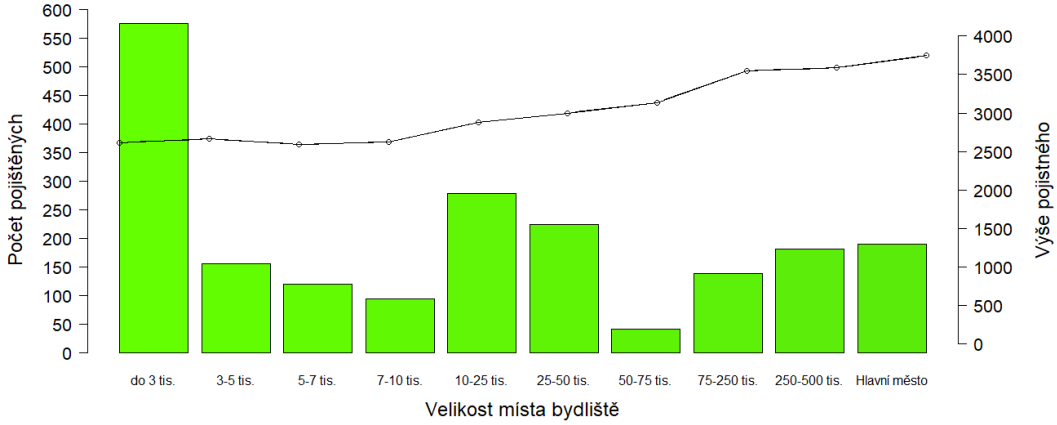


Obrázek 3.30: Graf závislosti ceny na výkonu motoru

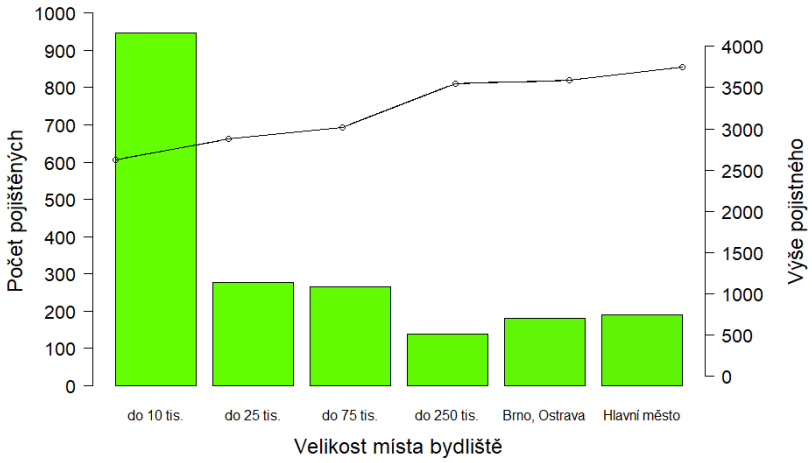


Obrázek 3.31: Graf závislosti ceny na výkonu motoru pro sloučené intervaly

Kategorie hlavního města by mohla být opět singulární vůči kategorii Prahy v proměnné okresů. Jakmile tedy proměnnou okresů zkategorizujeme, ověříme singularitu a případně kategorii sloučíme.



Obrázek 3.32: Graf závislosti ceny na velikosti místa bydliště



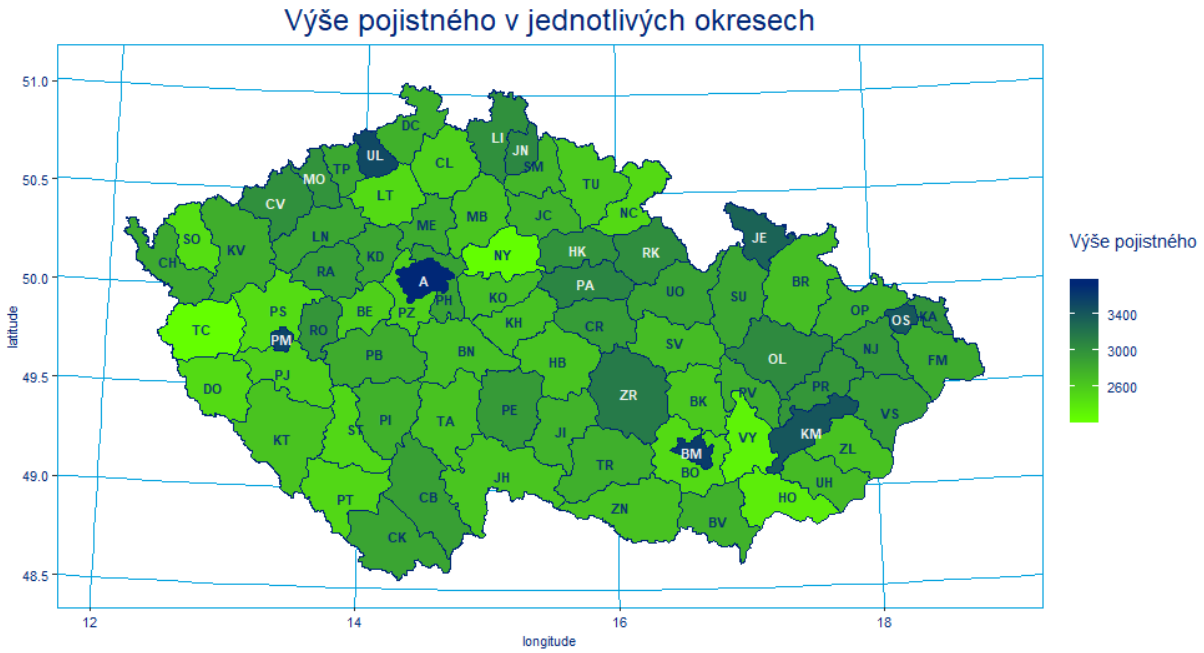
Obrázek 3.33: Graf závislosti ceny na velikosti místa bydliště pro sloučené intervaly

Okres bydliště

Závislost výše pojistného na okrese místa bydliště je vidět na mapě České republiky 3.34.

Oproti bílé pojišťovně nevidíme tak jednoznačné linie okresů s vyšším průměrným pojistným. Zůstává však Praha, kde je průměrné pojistné nejvyšší a větší města jako Brno, Ostrava, Plzeň nebo Ústí nad Labem. Zajímavé jsou okresy Kroměříže a Jeseníku, kde je pojistné také vyšší, ale přesný důvod k tomu neznáme. Stejně tak naopak okresy s nízkým průměrným pojistným jako Nymburk nebo Vyškov. Je možné, že je to způsobeno malým počtem vzorků, nicméně, je to pouze domněnka. Po sloučení okresů nám vznikne 6 oblastí:

- **Praha** - hl. město Praha



Obrázek 3.34: Mapa průměrného ceny v jednotlivých okresech ČR

- **Větší města** - Brno, Ostrava, Plzeň, Ústí nad Labem, Liberec, Jablonec nad Nisou, Pardubice, Hradec Králové,...
- **Z a J Čechy a Ostravsko** - okresy Krušných hor, Rakovnicko, České Budějovice a kolem Ostravy
- **Jih, Vysočina a Krkonoše** - od Jindřichova Hradce po Zlín, okolí Žďáru nad Sázavou a Krkonoše
- **Beroun a Šumava** - Beroun a Praha - západ, Šumava
- **V Morava a Nymburk** - Brno - venkov, Vyškov, Hodonín a Nymburk

Graf výše pojistného na četnosti výskytu vzorků v daných oblastech vidíme na obrázku 3.35.

Značka a typ vozidla

V případě značky a typu vozidla postupujeme stejným způsobem jako u bílé pojišťovny. Nejdříve vyjmeme 15 nejčastějších značek a 15 nejčastějších typů vozidel. Zbytek rozřadíme do kategorií podle karoserie 3.1. Vznikne opět 38 skupin, obrázek 3.36, které sloučíme podle podobnosti výše pojistného.

Výsledkem je 10 tarifních skupin:

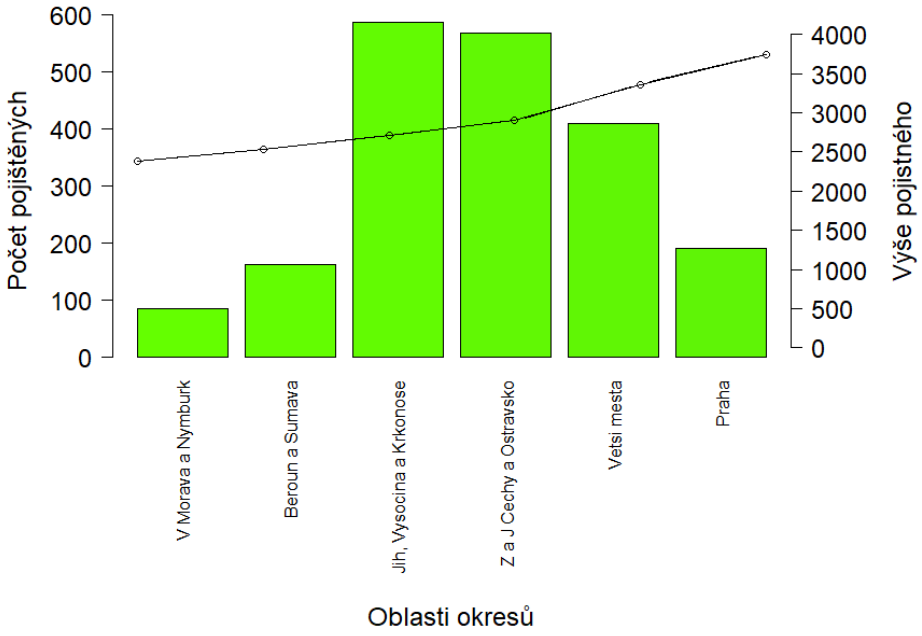
a) Audi, BMW, Mercedes-Benz

b) Dodávky, MPV

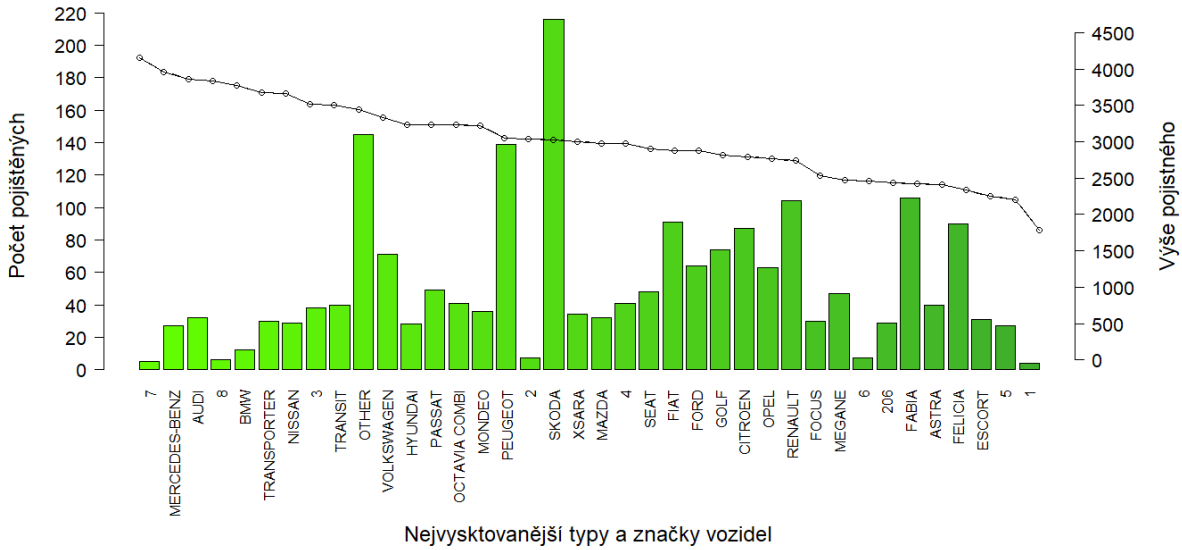
c) Ostatní

d) SUV

e) Passat, Mondeo, Octavia Combi, Hyundai, Volkswagen, Nissan a sportovní vozy



Obrázek 3.35: Graf závislosti ceny na oblastech bydliště



Obrázek 3.36: Graf závislosti ceny na 38 kategoriích vozidel

f) Peugeot, Škoda, Xsara, Mazda

g) Ford, Seat, Fiat, Sedany

h) Golf, Citroen, Opel, Renault

i) Escort, Fabia, 206, Astra, Felicia, Megane, Focus

j) Hatchbacky, Liftbacky a starší auta

Můžeme vidět, že některé skupiny jsou totožné jako u bílé pojišťovny, některé se liší. Těchto 10 skupin budeme považovat za finální vstup do modelu.

3.4.3 Korelace mezi proměnnými

Již z prvního modelu víme, že mezi výkonem motoru a objemem válců může existovat korelace. Ověříme tedy, jestli proměnné korelují i v datech černé pojišťovny a jak vysoká jejich závislost je. Vypočítáme si tedy Pearsonovy korelační koeficienty pro proměnné věku, velikost místa bydliště, výkonu motoru a objemu válců. Výsledky vidíme v tabulce 3.6.

	Věk	Velikost místa bydliště	Výkon motoru	Objem válců
Věk	1,00	0,00	0,01	0,04
Velikost místa bydliště	0,00	1,00	-0,01	0,00
Výkon motoru	0,01	-0,01	1,00	0,65
Objem válců	0,04	0,00	0,65	1,00

Tabulka 3.6: Tabulka koeficientů korelace pro proměnné černé pojišťovny

V případě černé pojišťovny vidíme lehkou korelaci i mezi proměnnými věku a objemem válců, věkem a výkonem motoru nebo velikostí místa bydliště a výkonem motoru. Tato závislost je však tak nízká, že ji vůbec uvažovat nemusíme. Mezi výkonem motoru a objemem válců je ale korelace s koeficientem 0,65. Pomocí deviačního testu ověříme, zda jsou obě proměnné v modelu významné, analýza 3.6. Pokud by nebyly, jednu proměnnou bychom do modelu nezahrnuli.

Listing 3.6: Deviační analýza modelů černé pojišťovny

```
Analysis of Deviance Table

Model 1: prices ~ age_m + cubic_capacity_m + distr_m + inhab_num_city_m +
    car_segm
Model 2: prices ~ age_m + cubic_capacity_m + engine_power_m + distr_m +
    inhab_num_city_m + car_segm
  Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
1      1971     13.593
2      1967     13.201  4  0.39131 14.555 1.017e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '␣' 1
```

Podle p-hodnoty a zbytkové deviaci můžeme říci, že finální model černé pojišťovny bude obsahovat obě proměnné. Obě proměnné jsou totiž v modelu významné.

Druhým problém, který jsme řešili u modelu bílé pojišťovny byla singularita u kategorií Hlavního města ve velikosti místa bydliště a Prahy v okresech. Tento problém nacházíme i zde, výpis 3.7, a stejně jako v předchozím případě, sloučíme Prahu s Brnem a Ostravou v proměnné velikosti místa bydliště.

Listing 3.7: Výsledky gamma regrese pro proměnné okres bydliště a velikost místa bydliště

```
Coefficients: (1 not defined because of singularities)
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                7.93786    0.02400 330.812  < 2e-16 ***
distr_mPraha                               0.29048    0.03225   9.006  < 2e-16 ***
distr_mV Morava a Nymburk                 -0.18341    0.03992  -4.594 4.62e-06 ***
distr_mZ a J Cechy a Ostravsko            -0.02841    0.02594  -1.095 0.273660
distr_mBeroun a Sumava                    -0.12648    0.03335  -3.792 0.000154 ***
distr_mJih, Vysocina a Krkonose           -0.07652    0.02636  -2.903 0.003737 **
inhab_num_city_mdo 25 tis.                 0.09050    0.02031   4.455 8.85e-06 ***
inhab_num_city_mdo 75 tis.                 0.12530    0.02102   5.962 2.95e-09 ***
inhab_num_city_mdo 250 tis.                0.24875    0.03083   8.069 1.22e-15 ***
inhab_num_city_mBrno, Ostrava              0.24670    0.03261   7.565 5.88e-14 ***
inhab_num_city_mHlavni mesto                    NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '␣' 1

(Dispersion parameter for Gamma family taken to be 0.08826883)

    Null deviance: 220.76  on 1999  degrees of freedom
Residual deviance: 180.54  on 1990  degrees of freedom
AIC: 32681

Number of Fisher Scoring iterations: 4
```

3.4.4 Regresní model a odhad parametrů

Do modelu nám tedy vstupuje o dvě vysvětlující proměnné méně než v případě bílé pojišťovny, a to o pohlaví pojištěného a frekvenci placení pojistného. Ostatní proměnné jsme zkategorizovali podle stejného principu, avšak do jiných intervalů než u bílé pojišťovny, např.: objem válců jsme rozdělili na více skupin než v případu bílé pojišťovny, značky a typy aut jsme lehce přeskupili.

Pomocí zobecněného lineárního modelu s gamma regresí a logaritmickou spojovací funkcí odhadneme koeficienty β. Jako intercept volíme:

- Věk: 24 - 30 let
- Výkon motoru: 100 kW a výš
- Objem válců motoru: 1.6 dm^3 - 1.8 dm^3
- Velikost místa bydliště: do 10 tis. obyvatel
- Místo bydliště: Jih, Vysočina a Krkonoše
- Vozidlo: Escort, Fabia, 206, Astra, Felicia, Megane, Focus

Odhadnuté koeficienty jednotlivých faktorů proměnných se nachází v výpisu 3.8.

Zjišťujeme, že i přes prvotní analýzu a zkategorizování dat je spousta faktorů v modelu nevýznamná. Jejich p-hodnota přesahuje koeficient spolehlivosti ($\alpha = 0,05$). Postupným slučováním nevýznamných kategorií (od nejvyšší p-hodnoty) docílíme výsledného modelu, výpis 3.9. Model bez nevýznamných proměnných je mírně lepší i podle Akaikova informačního kritéria, které u původního modelu činilo 24 468 a nyní 24 466. Není to však nijak významný rozdíl, a tak modely můžeme považovat za srovnatelné.

Listing 3.8: Gamma regrese černé pojišťovny

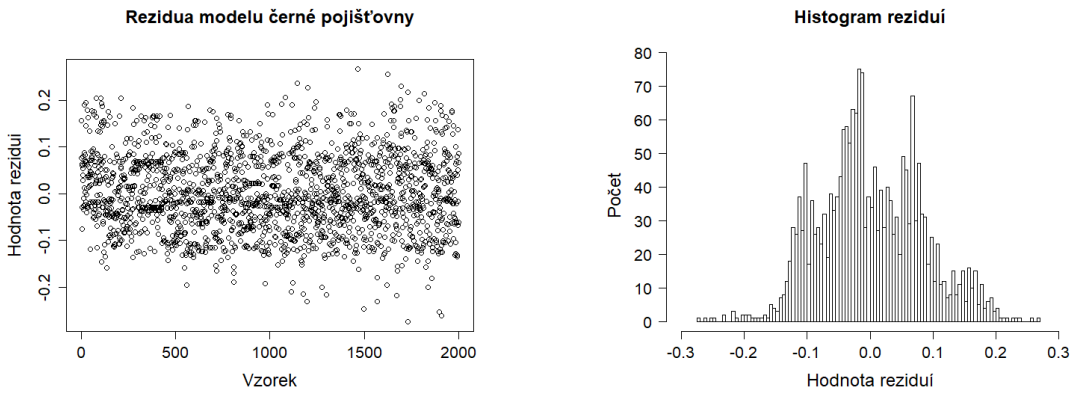
```
Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     7.999217   0.013894 575.748  < 2e-16 ***
age_m(17,23]                                    0.152549   0.025177   6.059 1.64e-09 ***
age_m(30,57]                                   -0.195589   0.008387 -23.321  < 2e-16 ***
age_m(57,100]                                  -0.241992   0.009360 -25.853  < 2e-16 ***
engine_power_m(-1,40]                          -0.083755   0.013718  -6.105 1.23e-09 ***
engine_power_m(40,60]                          -0.051484   0.008188  -6.288 3.95e-10 ***
engine_power_m(60,90]                          -0.034076   0.007603  -4.482 7.82e-06 ***
engine_power_m(90,100]                         -0.013377   0.009881  -1.354 0.175964
cubic_capacity_m(-1,1.2]                       -0.540267   0.012358 -43.717  < 2e-16 ***
cubic_capacity_m(1.2,1.3]                      -0.388369   0.011518 -33.720  < 2e-16 ***
cubic_capacity_m(1.3,1.6]                      -0.196564   0.007936 -24.768  < 2e-16 ***
cubic_capacity_m(1.8,2]                         0.221831   0.007551  29.379  < 2e-16 ***
cubic_capacity_m(2,2.5]                         0.367347   0.009489  38.711  < 2e-16 ***
cubic_capacity_m(2.5,Inf]                       0.623249   0.011997  51.951  < 2e-16 ***
distr_mVetsi mesta                              0.017613   0.007331   2.403 0.016375 *
distr_mPraha                                    0.022294   0.011257   1.980 0.047787 *
distr_mV Morava a Nymburk                      -0.009150   0.009634  -0.950 0.342381
distr_mZ a J Cechy a Ostravsko                  0.008834   0.004882   1.809 0.070528 .
distr_mBeroun a Sumava                          0.009682   0.007449   1.300 0.193803
inhab_num_city_mdo 25 tis.                      0.095093   0.005629  16.893  < 2e-16 ***
inhab_num_city_mdo 75 tis.                      0.128216   0.005827  22.003  < 2e-16 ***
inhab_num_city_mdo 250 tis.                     0.310214   0.008567  36.210  < 2e-16 ***
inhab_num_city_mBrno, Ostrava a Hlavni mesto    0.322616   0.009033  35.716  < 2e-16 ***
car_segm1                                      -0.081064   0.041450  -1.956 0.050644 .
car_segm3                                      -0.010461   0.014640  -0.715 0.474978
car_segm5                                      -0.008214   0.016576  -0.496 0.620262
car_segmOTHER                                  -0.029058   0.008691  -3.343 0.000843 ***
car_segmM-B, AUDI, BMW                         -0.012653   0.011693  -1.082 0.279346
car_segmVANS                                   -0.039439   0.011322  -3.484 0.000506 ***
car_segmFORD, SEAT, FIAT, 4                    -0.019010   0.006991  -2.719 0.006600 **
car_segmPASSAT, MONDEO, OCTAVIA COMBI,
HYUNDAI, VOLKSWAGEN, 2, NISSAN                 -0.026613   0.007224  -3.684 0.000236 ***
car_segmPEUGEOT, SKODA, XSARA, MAZDA           -0.007488   0.006305  -1.188 0.235118
car_segmGOLF, CITROEN, OPEL, RENAULT           -0.026489   0.006461  -4.100 4.31e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '␣' 1

(Dispersion parameter for Gamma family taken to be 0.006721101)

    Null deviance: 220.759  on 1999  degrees of freedom
Residual deviance:  13.201  on 1967  degrees of freedom
AIC: 27468

Number of Fisher Scoring iterations: 4
```

Listing 3.9: Gamma regrese černé pojišťovny bez nevýznamných kategorií

```
Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     7.996079   0.012966 616.686  < 2e-16 ***
age_m(17,23]                                    0.154057   0.025176   6.119 1.13e-09 ***
age_m(30,57]                                   -0.195969   0.008386 -23.369  < 2e-16 ***
age_m(57,100]                                  -0.242484   0.009364 -25.896  < 2e-16 ***
engine_power_m(-1,40]                          -0.081281   0.012712  -6.394 2.01e-10 ***
engine_power_m(40,60]                          -0.048167   0.006520  -7.387 2.20e-13 ***
engine_power_m(60,90]                          -0.029142   0.006008  -4.850 1.33e-06 ***
cubic_capacity_m(-1,1.2]                       -0.540316   0.012314 -43.878  < 2e-16 ***
cubic_capacity_m(1.2,1.3]                      -0.387775   0.011481 -33.776  < 2e-16 ***
cubic_capacity_m(1.3,1.6]                      -0.197238   0.007928 -24.877  < 2e-16 ***
cubic_capacity_m(1.8,2]                         0.221943   0.007518  29.521  < 2e-16 ***
cubic_capacity_m(2,2.5]                         0.365731   0.009290  39.367  < 2e-16 ***
cubic_capacity_m(2.5,Inf]                       0.625549   0.011836  52.851  < 2e-16 ***
distr_mVetsi mesta                              0.016710   0.007109   2.350 0.018853 *
distr_mPraha                                    0.020672   0.011134   1.857 0.063509 .
distr_mZ a J Cechy a Ostravsko                  0.008017   0.004547   1.763 0.078000 .
inhab_num_city_mdo 25 tis.                      0.095694   0.005631  16.995  < 2e-16 ***
inhab_num_city_mdo 75 tis.                      0.127570   0.005766  22.125  < 2e-16 ***
inhab_num_city_mdo 250 tis.                     0.309527   0.008546  36.218  < 2e-16 ***
inhab_num_city_mBrno, Ostrava a Hlavni mesto    0.322561   0.009048  35.652  < 2e-16 ***
car_segmFORD, SEAT, FIAT, 4, PEUGEOT,
SKODA, XSARA, MAZDA                            -0.011440   0.005558  -2.058 0.039694 *
car_segmM-B, AUDI, BMW, VANS, 3, OTHER         -0.025349   0.007140  -3.550 0.000394 ***
car_segmPASSAT, MONDEO, OCTAVIA C.,
HYUNDAI, VW, 2, NISSAN, GOLF, CITROEN,
OPEL, RENAULT                                  -0.026619   0.005716  -4.657 3.42e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '␣' 1

(Dispersion parameter for Gamma family taken to be 0.006748471)

    Null deviance: 220.759  on 1999  degrees of freedom
Residual deviance:  13.324  on 1977  degrees of freedom
AIC: 27466

Number of Fisher Scoring iterations: 4
```

Můžeme říci, že náš model dobře popisuje data, jelikož naše deviační statistika je rovna 13,324 a $\chi^2_{0,95}$ pro 1 011 stupňů volnosti bude jistě větší. Tedy hypotézu o správnosti modelu nezamítneme.

Vykreslíme si rezidua, obrázek 3.37 a pomocí Lillieforsova testu ověříme jejich normalitu, tabulka 3.7. Podle 3.37 rezidua z normálního rozdělení pocházet můžou a nevykazují žádnou závislost.



Obrázek 3.37: Rezidua modelu černé pojišťovny

Lillieforsův test	
D = 0,05964	p-hodnota < 2e-16

Tabulka 3.7: Lillieforsův test normality reziduí

Lillieforsův test však normalitu reziduí zamítá, a to díky p-hodnotě, která je nižší než hladina významnosti ($\alpha = 0{,}05$). Neznamená to nutně, že rezidua z normálního rozdělení opravdu nepocházejí, testy pro více než 1 000 vzorků jsou velmi citlivé.

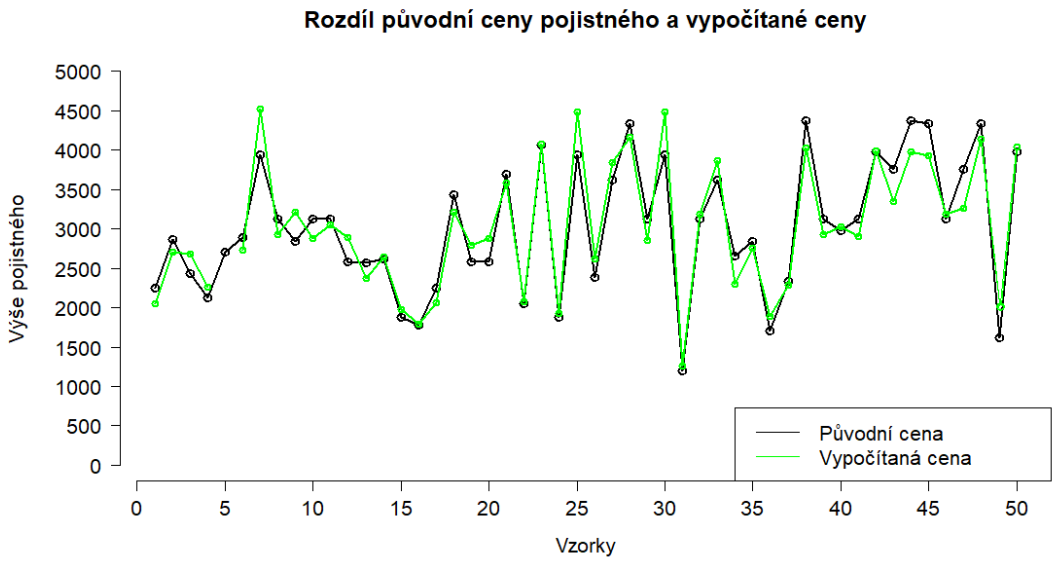
3.5 Odhad ročního pojistného černé pojišťovny

Odhadnuté koeficienty β z výsledného modelu 3.9 použijeme pro odhad výše pojistného černé pojišťovny. Námi vypočítanou cenu porovnáme s původní cenou a zobrazíme si histogram odchylek. Názorná ukázka ceny původní a námi vypočítané ceny pro 50 vzorků je vykreslena na 3.38, absolutní procentuální odchylka od původní ceny je vidět na obrázku 3.39.

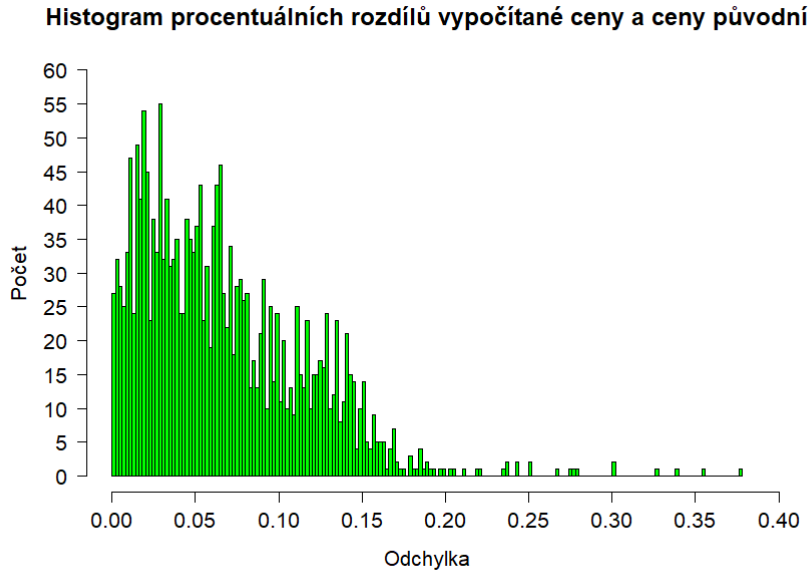
Abychom mohli srovnat náš model, odhadněme cenu pomocí průměru. Tedy řekněme, že odhadovaná cena je průměrná cena pojistného v České republice za rok 2019 (2 696 Kč), a to pro všechny pojistky. Tento odhad porovnejme se skutečnou cenou a vypočítejme si procentuální úspěšnost (3.1).

$$\text{usp_odhadu} = \text{abs}\left(1 - \frac{\text{průměrná cena}}{\text{cena původní}}\right). \tag{3.1}$$

Tabulka s procentuální úspěšností odhadů našeho modelu a odhad pomocí průměru je v tabulce 3.8.



Obrázek 3.38: Rozdíl ceny vypočítané a původní ceny černé pojišťovny



Obrázek 3.39: Histogram procentuálních odchylek od původní ceny

Pokud bychom uvažovali pouze průměrnou cenu trefujeme se do ceny téměř náhodně. Je tedy vidět, že cena u černé pojišťovny je odhadovaná složitějším výpočtem než pouze průměrem.

Námi sestavený model černé pojišťovny dává lepší výsledky než model s průměrem. Nicméně celkový odhad ceny máme výrazně horší než u bílé pojišťovny. Do odchylky 10% jsme odhadli stejný procentuální počet vzorků jako v bílé pojišťovně do 5%. Tím se i potvrdil náš předpoklad, že výsledný odhad bude horší, protože máme k dispozici méně obsáhlá data.

Odchylka	Vzorky glm model	% vzorků glm	Vzorky modelu s průměrem	% vzorků průměr
do 5%	879	44,0%	331	17,0%
do 10%	656	32,8%	170	9,0%
do 15%	364	18,2%	292	15,0%
do 20%	79	4,0%	146	7,0%
do 25%	10	0,5%	86	4,0%
do 30%	6	0,3%	204	10,0%
do 35%	4	0,2%	161	8,0%
do 35%	2	0,1%	66	3,0%
nad 40%	0	0,0%	544	27,0%

Tabulka 3.8: Celková chybovost modelu černé pojišťovny ve srovnání s jednoduchým modelem založeným na průměru

Jelikož nemáme možnost natrénovat model na jiné sadě dat a následně model zvalidovat, pokusíme se využít znalosti originální ceny z bílé pojišťovny a odhad pojistného černé pojišťovny vylepšit.

3.6 Vylepšení odhadu ročního pojistného černé pojišťovny

Cena bílé pojišťovny jako vysvětlující proměnná

Prvním způsobem, jak se pokusíme odhad pojistného černé pojišťovny vylepšit je, že přidáme do modelu originální cenu bílé pojišťovny jako jednu z vysvětlujících proměnných. Náš model černé pojišťovny pak bude vypadat následovně:

$$\text{pojistne}_{cerna} \sim glm(\text{vek}, \text{okres}, ..., \text{pojistne}_{orig_bila}) \tag{3.2}$$

Model spustíme. Jelikož jsou některé faktory nevýznamné, budeme je eliminovat postupným slučováním (od nejvyšší p-hodnoty) z modelu. Po sloučení několika kategorií zjišťujeme, že proměnná originálního pojistného bílé pojišťovny je nevýznamná, a to s aktuálně nejvyšší p-hodnotou, a tedy bychom ji měli z modelu odstranit, výpis 3.10.

To nám však dává původní model s původními odhadnutými koeficienty. Odhad pojistného tedy zůstává stejný a tak tímto způsobem bohužel vylepšení nedosáhneme.

Podíl cen pojišťoven

Druhým způsobem jakým se pokusíme model vylepšit je vzít za vysvětlovanou proměnnou podíl pojistného černé pojišťovny a pojišťovny bílé. Tento podíl pak vysvětlit stejnými proměnnými jako v klasickém modelu černé pojišťovny. Tedy

$$\frac{\text{pojistne}_{cerna}}{\text{pojistne}_{orig_bila}} \sim glm(\text{vek}, \text{okres}, ...). \tag{3.3}$$

Model spustíme. Nevýznamné kategorie sloučíme. Vidíme, že svou nevýznamnost zvyšuje intercept, což je v pořádku, jelikož výslednou cenu budeme počítat jako

$$\text{pojistne}_{cerna} = \exp(\beta_{intercept} + \beta_{pohlavi} + \cdots + \beta_{okres}) \cdot \text{pojistne}_{orig_bila}. \tag{3.4}$$

Listing 3.10: Model černé pojišťovny s proměnnou ceny bílé pojišťovny

```
Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     7.899e+00  1.723e-02 458.542  < 2e-16 ***
age_m(30,57]                                   -1.951e-01  8.522e-03 -22.895  < 2e-16 ***
age_m(17,23]                                    1.487e-01  2.627e-02   5.658 1.75e-08 ***
age_m(57,100]                                  -2.410e-01  9.754e-03 -24.709  < 2e-16 ***
engine_power_m(-1,40]                          -8.357e-02  1.378e-02  -6.067 1.56e-09 ***
engine_power_m(40,60]                          -5.132e-02  8.328e-03  -6.162 8.68e-10 ***
engine_power_m(60,90]                          -3.383e-02  7.627e-03  -4.436 9.65e-06 ***
engine_power_m(90,100]                         -1.418e-02  9.989e-03  -1.420 0.155738
cubic_capacity_m(1.2,1.3]                      -3.875e-01  1.152e-02 -33.647  < 2e-16 ***
cubic_capacity_m(-1,1.2]                       -5.394e-01  1.238e-02 -43.584  < 2e-16 ***
cubic_capacity_m(1.3,1.6]                      -1.966e-01  7.950e-03 -24.724  < 2e-16 ***
cubic_capacity_m(1.8,2]                         2.220e-01  7.562e-03  29.357  < 2e-16 ***
cubic_capacity_m(2,2.5]                         3.674e-01  9.501e-03  38.670  < 2e-16 ***
cubic_capacity_m(2.5,Inf]                       6.233e-01  1.201e-02  51.879  < 2e-16 ***
distr_mVetsi mesta                              1.737e-02  7.368e-03   2.357 0.018504 *
distr_mPraha                                    2.131e-02  1.151e-02   1.852 0.064233 .
distr_mV Morava a Nymburk                      -9.901e-03  9.643e-03  -1.027 0.304641
distr_mZ a J Cechy a Ostravsko                  8.305e-03  5.082e-03   1.634 0.102413
distr_mBeroun a Sumava                          8.571e-03  7.617e-03   1.125 0.260619
inhab_num_city_mdo 25 tis.                      9.508e-02  5.634e-03  16.874  < 2e-16 ***
inhab_num_city_mdo 75 tis.                      1.275e-01  6.078e-03  20.971  < 2e-16 ***
inhab_num_city_mdo 250 tis.                     3.093e-01  8.890e-03  34.789  < 2e-16 ***
inhab_num_city_mBrno, Ostrava a Hlavni mesto    3.215e-01  9.379e-03  34.279  < 2e-16 ***
car_segm3                                      -1.010e-02  1.470e-02  -0.687 0.492229
car_segmOTHER                                  -2.891e-02  8.715e-03  -3.318 0.000925 ***
car_segmM-B, AUDI, BMW                         -1.323e-02  1.173e-02  -1.128 0.259536
car_segmVANS                                   -4.075e-02  1.168e-02  -3.490 0.000494 ***
car_segmFORD, SEAT, FIAT, 4                    -1.912e-02  6.997e-03  -2.733 0.006323 **
car_segmPASSAT, MONDEO, OCTAVIA COMBI,
HYUNDAI, VOLKSWAGEN, 2, NISSAN                 -2.688e-02  7.232e-03  -3.716 0.000208 ***
car_segmPEUGEOT, SKODA, XSARA, MAZDA           -7.487e-03  6.321e-03  -1.184 0.236361
car_segmGOLF, CITROEN, OPEL, RENAULT           -2.647e-02  6.468e-03  -4.092 4.45e-05 ***
car_segm5, 1                                   -1.683e-02  1.560e-02  -1.079 0.280664
cena_bila                                       1.602e-06  3.472e-06   0.461 0.644680
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '␣' 1

(Dispersion parameter for Gamma family taken to be 0.006733206)

    Null deviance: 220.759  on 1999  degrees of freedom
Residual deviance:  13.218  on 1967  degrees of freedom
AIC: 27085

Number of Fisher Scoring iterations: 4
```

Listing 3.11: Model černé pojišťovny s podílem cen černé a bílé pojišťovny

```
Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     0.001574   0.028392   0.055 0.955788
age_m(17,23]                                   -0.342566   0.062198  -5.508 4.11e-08 ***
age_m(30,57]                                   -0.046791   0.020728  -2.257 0.024095 *
age_m(57,100]                                   0.048401   0.023129   2.093 0.036503 *
engine_power_m(40,60]                           0.081465   0.010357   7.865 6.01e-15 ***
engine_power_m(90,100]                         -0.152622   0.019484  -7.833 7.70e-15 ***
cubic_capacity_m(-1,1.2]                       -0.470847   0.026737 -17.611  < 2e-16 ***
cubic_capacity_m(1.2,1.3]                      -0.404702   0.027833 -14.541  < 2e-16 ***
cubic_capacity_m(1.3,1.6]                      -0.158490   0.019520  -8.119 8.16e-16 ***
cubic_capacity_m(1.8,2]                         0.257142   0.018504  13.897  < 2e-16 ***
cubic_capacity_m(2,2.5]                         0.400839   0.023004  17.425  < 2e-16 ***
cubic_capacity_m(2.5,Inf]                       0.661274   0.027522  24.027  < 2e-16 ***
inhab_num_city_mdo 25 tis.                      0.092810   0.013913   6.671 3.29e-11 ***
inhab_num_city_mdo 75 tis.                     -0.051291   0.014347  -3.575 0.000359 ***
inhab_num_city_mdo 250 tis.                     0.070965   0.021129   3.359 0.000798 ***
inhab_num_city_mBrno, Ostrava a Hlavni mesto    0.047890   0.022326   2.145 0.032075 *
distr_mVetsi mesta                             -0.049783   0.017876  -2.785 0.005405 **
distr_mPraha                                   -0.189088   0.027671  -6.834 1.10e-11 ***
distr_mZ a J Cechy a Ostravsko                 -0.132912   0.011702 -11.358  < 2e-16 ***
distr_mBeroun a Sumava                         -0.162939   0.018114  -8.995  < 2e-16 ***
car_segm3                                       0.129379   0.034156   3.788 0.000157 ***
car_segmM-B, AUDI, BMW                         -0.120477   0.026292  -4.582 4.89e-06 ***
car_segmVANS                                   -0.279474   0.025709 -10.871  < 2e-16 ***
car_segmPASSAT, MONDEO, OCTAVIA COMBI,
HYUNDAI, VOLKSWAGEN, 2, NISSAN                 -0.068400   0.014602  -4.684 3.00e-06 ***
car_segmFORD, SEAT, FIAT, 4, GOLF,
CITROEN, OPEL, RENAULT                         -0.039255   0.010857  -3.615 0.000307 ***
car_segm5, 1                                    0.144493   0.037564   3.847 0.000124 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '␣' 1

(Dispersion parameter for Gamma family taken to be 0.04112208)

    Null deviance: 227.2  on 1999  degrees of freedom
Residual deviance:  79.7  on 1974  degrees of freedom
AIC: -768.12

Number of Fisher Scoring iterations: 4
```

Tedy pokud uzavřená pojistná smlouva bude mít parametry interceptu, odhadovaný koeficient bude roven $0 \Rightarrow \exp(0) = 1$.

Odhadnuté koeficienty pro podíl vidíme ve výpisu 3.11.

Pomocí koeficientů vypočítáme podle rovnice (3.4) pojistné a porovnáme s původní cenou. Procentuální odchylky modelu s podílem nalezneme v tabulce 3.9, kde pro srovnání nalezneme i procentuální odchylky původního modelu bez vylepšení.

Ve srovnání s původním modelem je vidět, že odhadnutí podílů cen černé a bílé pojišťovny nám celkovou chybovost nevylepšuje. Stále tedy považujeme původní model za nejlepší odhad modelu černé pojišťovny.

Odchylka	Vzorky model	% vzorků model	Vzorky model s podílem	% vzorků modelu s podílem
do 5%	879	44,0%	419	21,0%
do 10%	656	32,8%	373	18,7%
do 15%	364	18,2%	338	16,9%
do 20%	79	4,0%	259	13,0%
do 25%	10	0,5%	194	9,7%
do 30%	6	0,3%	127	6,4%
do 35%	4	0,2%	97	4,9%
do 35%	2	0,1%	58	2,9%
nad 40%	0	0,0%	135	6,8%

Tabulka 3.9: Celková chybovost modelu s podílem a původního modelu černé pojišťovny

Kapitola 4

Umělé neuronové sítě

Jako alternativní metodu k zobecněným lineárním modelům volíme umělé neuronové sítě. Statistické modely, jako je například zobecněný lineární model, fungují na principu hledání řešení pomocí nějakého přesně zadaného algoritmu. Tedy my jako uživatelé říkáme počítači, jak a pomocí čeho má model daný problém vyřešit. Naopak umělé neuronové sítě si zakládají na principu učení z dat a vymýšlení vlastního řešení problému, podobně jako to dělá lidský mozek. Cílem umělých neuronových sítí je co nejvíce napodobit lidský mozek co se týče rozhodování. [13]

Struktura neuronové sítě

Základní jednotkou neuronové sítě je neuron neboli percepton. Neurony jsou vzájemně propojeny a předávají si mezi sebou signály. Každý neuron může mít více vstupů, označme $x_1, \ldots, x_N$, ale pouze jeden výstup - y. Vstupem nazýváme informaci z vnějšího okolí nebo z jiného neuronu (u vícevrstvých sítí). Každý vstup nese určitou váhu $w_1 \ldots, w_N$ (tzv. synaptická váha), která vstupuje do těla neuronu společně s prahovou hodnotou h.

V těle jednotlivých neuronů je na základě synaptických vah vytvářen vnitřní potenciál Σ. Pokud tento potenciál překoná hodnotu prahu h, neuron je nabuzen a vysílá signál ve formě přenosové (aktivační) funkce. [14]

V praxi se používá velké množství přenosových funkcí, záleží vždy na typu sítě. Nejjednodušší aktivační funkcí je funkce lineární, kde není užita žádná transformace. Síť složená z lineárních transformací je velmi snadno vycvičitelná, ale nedokáže se naučit složité mapovací funkce. Proto jsou preferovány nelineární aktivační funkce, které právě umožňují uzlům naučit se složitější struktury v datech. [15]

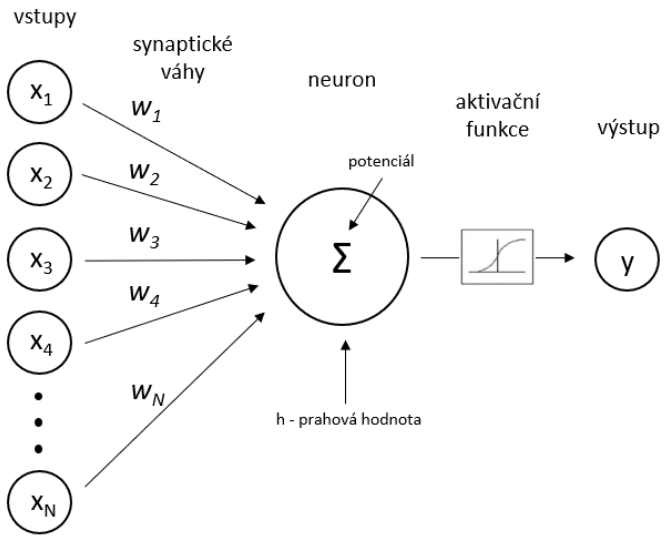
Nejčastěji využívaná přenosová funkce u vícevrstvých neuronových sítí je funkce sigmoidní. Je užitečná zejména proto, že je lze snadno zderivovat a dosazení do již zderivované funkce snižuje výpočetní náročnost algoritmu během učení sítě. [17]

Sigmoidní funkce je definovaná předpisem

$$f_s(x) = \frac{1}{1 + e^{-x}}. \tag{4.1}$$

Strukturu zjednodušené umělé neuronové sítě můžeme vidět na obrázku 4.1.

Levá vrstva se nazývá vrstvou vstupní, pravá vrstva vrstvou výstupní a střední vrstva (příp. vrstvy) vrstvou skrytou. Kolik bude mít síť skrytých vrstev záleží na konkrétním případu. Obecně ale platí, že více skrytých vrstev používají složitější úlohy.



Obrázek 4.1: Schéma umělé neuronové sítě

Učení neuronové sítě

Tak jako se během života učí lidský mozek, je potřeba naučit i umělou neuronovou síť. Učení neuronové sítě znamená nastavit váhy modelu tak, aby vstupní informace byla správně transformována do informace výstupní. K učení neuronové sítě se používají dvě metody: učení s učitelem a učení bez učitele.

Při učení bez učitele nemá síť k dispozici informaci o správnosti výstupu. Síť využívá tzv. shlukovou analýzu, kdy se ve vstupním prostoru snaží najít sobě podobné prvky, které rozdělí do shluků. Nelze však nikde ověřit, že byl prvek zařazen správně.

Učení s učitelem znamená, že síť má k dispozici výstupy odpovídající vstupům, tedy ví, jaké je správné řešení úlohy. Neuronová síť odhadne výstup na základě vstupů. Tento odhadnutý výstup je následně porovnán s výstupem požadovaným, určí se odchylka a provede se korekce (změna vah a prahů), tak aby se odchylka minimalizovala. Optimální váhy se hledají pomocí iteračních numerických metod. Neuronová síť se tak učí ze svých vlastních chyb. Tento optimalizační algoritmus se nazývá backpropagation. [20]

4.1 Model s využitím neuronové sítě

Pro odhad ceny konkurence pomocí neuronových sítí v této diplomové práci je použit programovací jazyk Python a jeho volně dostupné balíčky Tensorflow a Keras. V těchto balíčcích je třeba si model sestavit. Sestavení modelu neuronových sítí se skládá z 5 kroků. [18]

1. Definování struktury sítě

V prvním kroku je třeba definovat síť. Tedy říci, kolik vrstev a kolik neuronů v každé vrstvě má model obsahovat. První vrstva musí vždy definovat počet očekávaných vstupů. Poslední vrstva musí definovat počet očekávaných výstupů. Vrstvy mezi první a poslední definují počet skrytých neuronů v dané vrstvě. Dále všechny skryté vrstvy a vrstva výstupní musí mít definovanou přenosovou funkci. Nejdůležitější je

výběr přenosové funkce u výstupní vrstvy, ta totiž definuje formát, jakým budou predikovány výsledné hodnoty.

2. Kompilace sítě

Po definování sítě je třeba provést kompilaci. Ta transformuje jednoduchou sekvenci definovaných vrstev do vysoce efektivní řady maticových transformací. Kompilace vyžaduje definování dvou parametrů - optimalizačního algoritmu a ztrátovou funkci. Mezi používané optimalizační algoritmy patří Stochastic Gradient Descent (SGD), Adaptive Moment Estimation (adam) nebo Root Mean Square Propagation (RMSProp) [16]. Nejčastěji užívanou ztrátovou funkcí je střední kvadratická chyba (mse).

3. Trénování sítě

Po zkompilování modelu následuje samotné trénování sítě. Tedy přizpůsobování vah na vstupní data. To se u dopředných neuronových sítí provádí pomocí backpropagation algoritmu. Zjednodušeně řečeno, algoritmus pracuje na principu porovnání vlastního výsledku se skutečným a zjištěním o kolik se neuronová síť v predikci spletla. Zpětně se pak vypočítá kde a jak se musí změnit váhy, aby se chyba predikce minimalizovala. Takový jeden průchod všemi trénovacími daty se nazývá epocha. Počet epoch volíme tak, abychom minimalizovali chybu, ale zároveň nepřetrénovali model.

4. Vyhodnocení sítě

Námi sestavenou síť vyhodnotíme na základě ztrátové funkce a přesnosti odhadu. V případě spojité proměnné se model hodnotí na základě koeficientu determinace, případně adjustovaného koeficientu determinace (2.8).

5. Predikce

Jakmile máme model sestavený, natrénovaný a vyhodnocený, přejdeme k predikci. Pomocí vstupních dat a natrénovaného modelu predikujeme výstupní hodnoty. Tyto hodnoty jsou ve tvaru, jaký jsme definovali při vytváření modelu.

4.2 Odhad cen konkurence pomocí neuronové sítě

Jak jsme již zmínili, jako alternativní metodu k zobecněným lineárním modelům použijeme neuronovou síť. Sestavíme model neuronové sítě a porovnáme přesnost odhadu cen konkurence pomocí neuronových sítí s metodou zobecněných lineárních modelů.

Provedeme obdobný postup jako u zobecněných lineárních modelů. Nejdříve se pokusíme odhadnout ceny bílé pojišťovny, zde platí, že máme dostatek trénovacích i testovacích dat. A následně provedeme odhad cen černé pojišťovny, kde nám zůstává pouze jedna sada dat.

4.2.1 Vstupní data

Vysvětlující proměnné je třeba převést na kategorické. Jelikož jsme každou proměnnou již analyzovali, nebudeme postup opakovat a použijeme kategorie z první metody. Zároveň je třeba převést vysvětlovanou proměnnou v trénovací množině dat na hodnoty z intervalu $[0, 1]$, se kterými dokáže neuronová síť

pracovat. A to tak, aby stále vyjadřovaly svoji hodnotu.

Můžeme například využít $\max(y)$ a touto hodnotou vydělit hodnoty ostatní. Transformované hodnoty by sice ležely v $(0, 1]$, ale prvky by nebyly v intervalu rovnoměrně rozdělené, jelikož ani cena pojistného není rovnoměrně rozdělena. Model neuronové sítě by tak mohl nepřesně odhadovat ceny blízké průměrnému pojistnému (2 696 Kč), kterých je v datech nejvíce. Proto transformaci dat na interval $[0, 1]$ provedeme pomocí empirické distribuční funkce.

Definice 4.2.1 (Empirická distribuční funkce)**.** *Nechť* $(X_1, ...X_n)$ *jsou nezávislé, stejně rozdělené reálné náhodné veličiny se společnou distribuční funkcí* $F(x)$*. Potom empirickou distribuční funkce definujeme jako*

$$\widehat{F}_n(x) = \frac{\text{počet prvků} \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leq x},$$

kde $\mathbf{1}_{X_i \leq x}$ *je tzv. indikátorová funkce. [19]*

Díky této transformaci se nám hodnoty rozmístí rovnoměrně v intervalu $[0, 1]$. Takto upravené hodnoty považujeme za vstup do trénovací fáze modelu neuronové sítě.

Abychom následně mohli porovnat náš odhad ceny s cenu skutečnou, tedy získat zpět cenu v korunách, použijeme kvantilovou funkci.

Definice 4.2.2 (Kvantilová funkce)**.** *Nechť* $F(x)$ *je distribuční funkce náhodné veličiny* X*. Pak funkce*

$$\widehat{F}^{-1}(u) = inf\{x \in \mathbb{R}; F(x) \geq u\},$$

kde $0 < u < 1$ *se nazývá kvantilová funkce náhodné veličiny* X*. Pro* $0 < \alpha < 1$ *se hodnota* $F^{-1}(\alpha)$ *nazývá* α*- kvantil. [19]*

4.3 Model bílé pojišťovny

Pro model odhadu cen konkurence použijeme vícevrstvou síť, konkrétně síť s jednou skrytou vrstvou. Do vstupní vrstvy volíme 41 neuronů (což odpovídá počtu kategorií ve vstupních datech), do skryté vrstvy taktéž 41 neuronů a ve výstupní vrstvě bude 1 neuron. Jako přenosovou funkci jsme vyzkoušeli funkci ReLu (rectified linear unit [15]) a funkci sigmoidální. Funkce sigmoidální dávala opakovaně lepší výsledky a tedy budeme používat tu, a to jak u skryté vrstvy tak u výstupní. Ztrátová funkce je zvolena jako střední kvadratická chyba a optimalizační algoritmus adam. Při volbě epoch začneme s menším počtem a postupně, pokud bude chyba klesat, budeme přidávat. Začneme na 10 epochách.

4.3.1 Trénovací množina o 30 000 vzorcích

Model nejprve zkusíme natrénovat na množině dat o 30 000 vzorcích, otestovat ho na této množině a následně zvalidovat pomocí jiné množiny. Po trénovací fázi vidíme, že přesnost odhadu se pohybuje řádově kolem e-05. Ačkoli by se mohlo zdát, že díky ukazateli přesnosti je náš model špatný, je třeba si uvědomit, že odhadujeme velmi malé a hlavně konkrétní desetinná čísla. Nemůžeme předpokládat, že se model trefí zcela přesně. Ztrátová funkce se pak pohybuje od 0,02 do 0,007.

Pomocí vytvořené a natrénované neuronové sítě odhadneme hodnoty trénovací množiny a následně množiny testovací. Námi odhadnuté ceny si pomocí kvantilové funkce převedeme na hodnoty v Kč a porovnáme s originálními. Výsledky můžeme vidět v tabulce 4.1.

Odchylka	Vzorky test. mn.	% vzorků test. mn.	Vzorky valid. mn.	% vzorků valid. mn.
do 5%	20 617	68,7%	20 784	69,3%
do 10%	5 506	18,4%	5 279	17,6%
do 15%	1 736	5,8%	1 882	6,3%
do 20%	600	2,0%	669	2,2%
do 25%	458	1,5%	384	1,3%
do 30%	337	1,1%	277	0,9%
do 35%	338	1,1%	327	1,1%
do 40%	233	0,8%	256	0,9%
do 45%	75	0,3%	75	0,3%
do 50%	32	0,1%	26	0,1%
nad 50%	60	0,2%	43	0,1%

Tabulka 4.1: Celková chybovost modelu neuronové sítě bílé pojišťovny s 10 epochami

Z výsledků lze vidět, že se nacházíme kolem úspěšnosti odhadu 70% s odchylkou do 5% od originální ceny. Pokud bychom tolerovali odchylku do 10%, pak je náš odhad dobrý v 88% případů. Adjustovaný koeficient korelace má hodnotu $\overline{R^2} = 0,8435$. Náš model objasňuje 84,3% celkové variability vysvětlované proměnné.

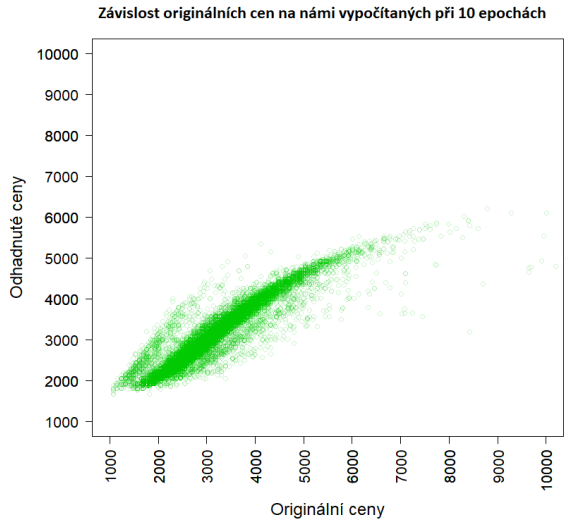
Zkusme nyní zvýšit počet epoch při trénování dat z 10 na 100. Znovu spustíme trénování modelu. Ztrátová funkce se nyní pohybuje od 0,02 do 0,006. Výsledné odchylky odhadu od originálních dat pro testovací a validační množinu nalezneme v tabulce 4.2.

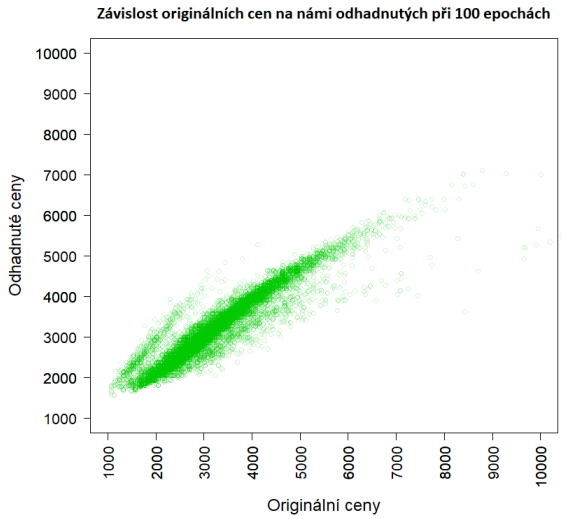
Odchylka	Vzorky test. mn.	% vzorků test. mn.	Vzorky valid. mn.	% vzorků valid. mn.
do 5%	22 167	73,9%	22 194	74,0%
do 10%	4 535	15,1%	4 378	14,6%
do 15%	1 459	4,9%	1 538	5,1%
do 20%	542	1,8%	584	2,0%
do 25%	392	1,3%	376	1,3%
do 30%	257	0,9%	256	0,9%
do 35%	360	1,2%	386	1,3%
do 40%	172	0,6%	165	0,6%
do 45%	45	0,2%	54	0,2%
do 50%	23	0,1%	27	0,1%
nad 50%	48	0,2%	43	0,1%

Tabulka 4.2: Celková chybovost modelu neuronové sítě bílé pojišťovny se 100 epochami

Zvýšením počtu epoch z 10 na 100 se nám odhad vylepšil přibližně o 5%. Odhady transformovaných cen pomocí empirické distribuční funkce a jejich originální hodnoty můžeme vidět na obrázku 4.2, kde vidíme i rozdíl mezi modelem s 10 epochami a modelem se 100 epochami. Bohužel žádné další zvýšení počtu epoch už nepřineslo zlepšení výsledků a tak budeme považovat 100 epoch pro náš model za nejlepší a nadále s ním pracovat. Stále se ale nedostáváme na úspěšnost odhadu zobecněného lineárního

modelu. Jelikož od bílé pojišťovny máme dat dostatek zkusme nyní k trénování použít sadu dat o 200 000 vzorcích a k validaci sadu o 30 000 vzorcích.





Obrázek 4.2: Odhady cen a jejich skutečná hodnota pro model s 10 a 100 epochami

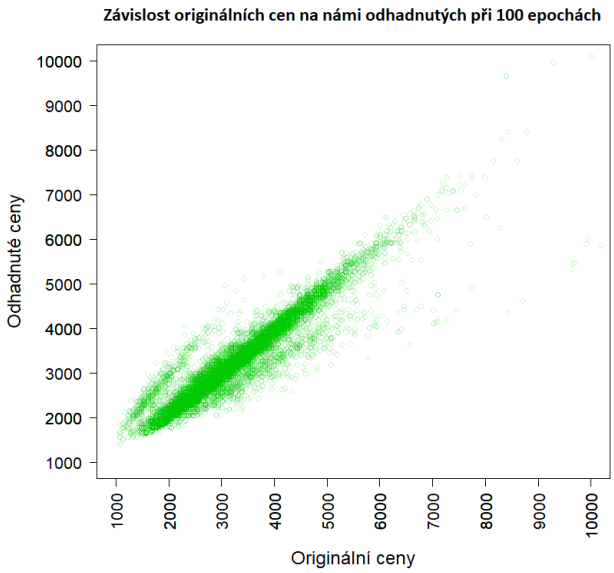
4.3.2 Trénovací množina o 200 000 vzorcích

Předpokládáme, že navýšením počtu trénovacích dat bychom mohli dosáhnout lepších výsledků odhadu cen konkurence. Natrénujeme nyní model na datové sadě o 200 000 vzorcích se stejnými parametry a se 100 epochami. Výsledky v tabulce 4.3, ajdustovaný koeficient korelace je roven $\overline{R^2} = 0,9255$.

Odchylka	Vzorky valid. mn.	% vzorků valid. mn.
do 5%	23 684	79,0%
do 10%	3 436	11,5%
do 15%	1 163	3,9%
do 20%	481	1,6%
do 25%	351	1,2%
do 30%	221	0,7%
do 35%	374	1,3%
do 40%	173	0,6%
do 45%	49	0,2%
do 50%	23	0,1%
nad 50%	45	0,2%

Tabulka 4.3: Celková chybovost modelu neuronové sítě bílé pojišťovny s trénovací množinou o 200 000 vzorcích

Náš předpoklad se potvrdil, pokud natrénujeme model na větším množství dat, odhad se vylepší. S tímto modelem jsme schopni dosáhnout minimálně stejně tak dobré výsledky, jako se zobecněným lineárním modelem. Graficky vidíme původní cenu a cenu odhadnutou na obrázku 4.3.



Obrázek 4.3: Odhady cen a jejich skutečná hodnota pro model s trénovací množinou o 200 000 vzorcích a 100 epochami

4.4 Model černé pojišťovny

Nyní se podívejme, jak na tom bude odhad cen černé pojišťovny. Zda odhad pomocí neuronových sítí bude přesnější než pomocí zobecněného lineárního modelu. Opět převezmeme kategorie jednotlivých proměnných ze zobecněného lineárního modelu. Proměnné pohlaví a frekvence placení pojistného do modelu pro jejich nevýznamnost zahrnovat nebudeme. Nastavení neuronové sítě necháme stejné jako u bílé pojišťovny, jediné co se nám změní, tak počet neuronů ve vstupní a skryté vrstvě. U černé pojišťovny máme dohromady ve všech vstupních proměnných 38 kategorií, a tedy i neuronů zvolíme v prvních dvou vrstvách 38, ve výstupní vrstvě zůstává 1.

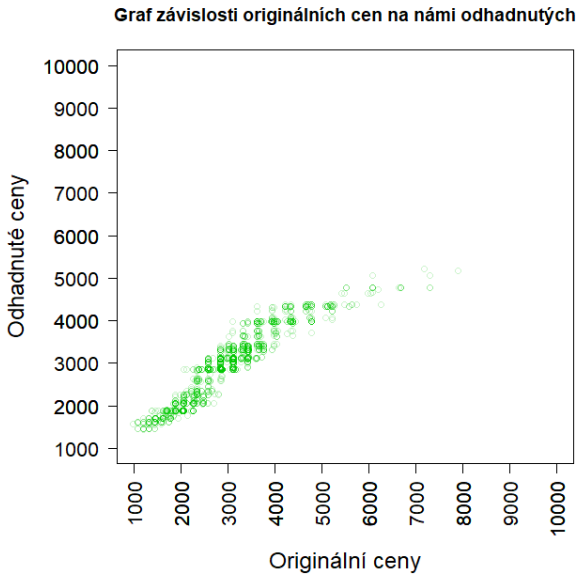
4.4.1 Trénovací množina o 2 000 vzorcích

Na množině o 2 000 vzorcích budeme model jak trénovat, tak testovat. Při trénování nám ztrátová funkce klesá od 0,15 do 0,005. Přesnost je opět velmi nízká, ale platí co jsme již zmínili. Pro spojitý případ to není dobrý ukazatel. Po natrénování přejdeme k odhadu cen. Procentuální přesnost můžeme vidět v tab. 4.4, na obrázku 4.4 pak závislost původních cen a cen námi odhadnutých. Adjustovaný koeficient determinace je v tomto případě $\overline{R^2} = 0,9334$, tedy model černé pojišťovny lépe objasňuje variabilitu vysvětlované proměnné než model bílé pojišťovny.

S procentuální úspěšností do 5% jsme se nedostali blíže než u zobecněného lineárního modelu. Pokud se díváme na úspěšný odhad do 10%, pak se dostáváme naším odhadem na úspěšnost odhadu zobecněným lineárním modelem. Nelze však říci, že by základní model s využitím neuronové sítě byl lepší.

Odchylka	Vzorky	% vzorků
do 5%	777	38,9%
do 10%	745	37,3%
do 15%	242	12,1%
do 20%	144	7,2%
do 25%	47	2,4%
do 30%	23	1,2%
do 35%	10	0,5%
do 40%	7	0,4%
do 45%	2	0,1%
do 50%	0	0,0%
nad 50%	3	0,2%

Tabulka 4.4: Celková chybovost modelu neuronové sítě černé pojišťovny



Obrázek 4.4: Odhady cen a jejich skutečná hodnota pro model černé pojišťovny

4.4.2 Vylepšení odhadu černé pojišťovny

Jelikož stále nejsme spokojeni s výsledky pro černou pojišťovnu, pokusíme se model vylepšit, a to dvěma způsoby.

4.4.2.1 Přidání proměnné pohlaví a frekvenci placení pojistného

Ačkoli se proměnné pohlaví a frekvence placení pojistného zdály z prvotní analýzy nevýznamné, není vyloučené, že model s neuronovou sítí tyto proměnné nějakým způsobem do odhadu zahrne. Natrénujeme tedy neuronovou síť na vstupních datech s přidáním proměnné pohlaví a frekvenci placení pojistného.

Odchylka	Vzorky	% vzorků
do 5%	778	38,9%
do 10%	752	37,6%
do 15%	244	12,2%
do 20%	131	6,6%
do 25%	59	3,0%
do 30%	20	1,0%
do 35%	9	0,5%
do 40%	3	0,2%
do 45%	4	0,2%
do 50%	0	0,0%
nad 50%	0	0,0%

Tabulka 4.5: Celková chybovost modelu neuronové sítě černé pojišťovny s přidáním proměnných pohlaví a frekvencí placení pojistného

Z odhadu, který zůstal na podobných hodnotách jako základní model, vidíme, že přidané proměnné výsledek neovlivní. S největší pravděpodobností tyto proměnné v černé pojišťovně do výpočtu pojistného opravdu zahrnuty nejsou a odhad nám nijak nevylepší.

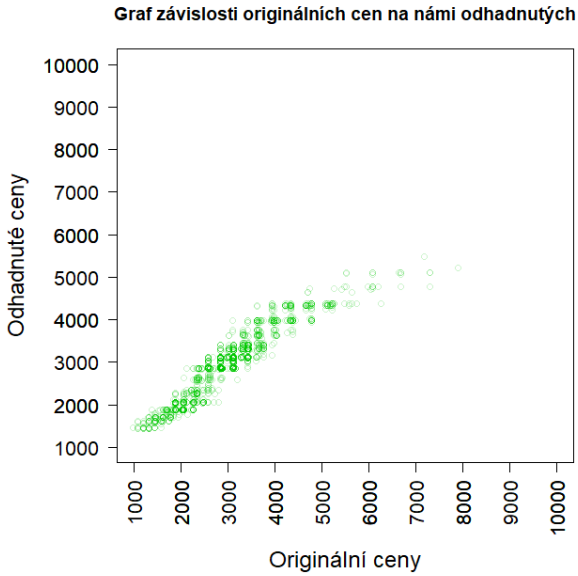
4.4.2.2 Větší množství tarifních skupin ve vstupních datech

Tím, že jsme sloučili některé kategorie značek a typů vozidel, je pravděpodobné, že nám unikla nějaká informace. Nebudeme ovšem zahrnovat kompletně všechny značky a typy aut, ale zahrneme námi prvotní rozdělení vozidel na 38 tarifních skupin (15 nejvyskytovanějších značek, 15 nejvyskytovanějších typů a 8 kategorií zbylých aut podle karoserie). Výsledky v tabulce 4.6 a vykreslení výsledků na obrázku 4.5. Adjustovaný koeficient determinace je $\overline{R^2} = 0,9372$.

Odchylka	Vzorky	% vzorků
do 5%	825	41,3%
do 10%	748	37,4%
do 15%	206	10,3%
do 20%	133	6,7%
do 25%	53	2,7%
do 30%	19	1,0%
do 35%	9	0,5%
do 40%	2	0,1%
do 45%	3	0,2%
do 50%	0	0,0%
nad 50%	2	0,1%

Tabulka 4.6: Celková chybovost modelu neuronové sítě černé pojišťovny s 38 kategoriemi aut

Odhad cen se nám mírně vylepšil, adjustovaný koeficient determinace se nepatrně zvýšil, ale stále nedosahujeme výraznějšího zlepšení.



Obrázek 4.5: Odhady cen a jejich skutečná hodnota pro model černé pojišťovny s 38 kategoriemi aut

4.5 Výsledky

Model s využitím neuronových sítí bohužel nepřinesl žádné lepší výsledky než klasická metoda zobecněných lineárních modelů. Můžeme říci, že v případě, kdy máme dostatečné množství dat, dosahuje metoda neuronových sítí stejné úspěšnosti odhadu jako zobecněné lineární modely. A tedy by se k odhadu cen konkurence dala využít, především pro svoji rychlost. V případě, že dat není dostatečné množství je lepší využívat zobecněné lineární modely či nějaký jednodušší model.

Závěr

Tato práce byla zaměřena na odhad cen konkurence v pojištění odpovědnosti z provozu vozidla. K odhadu byly použity dvě metody. První klasická metoda zobecněného lineárního modelu a druhá více experimentální metoda neuronových sítí. K dispozici jsme měli data od dvou pojišťoven, přičemž od bílé pojišťovny jsme dostali rozsáhlé vzorky cen, černá pojišťovna pak disponovala pouze jednou sadou o malém počtu vzorků.

První část práce byla zaměřena na odhad cen konkurence pomocí zobecněných lineárních modelů. V případě dostatečného počtu vzorků (bílá pojišťovna) jsme dosáhli 77% úspěšnosti odhadu do odchylky 5% a 90% odhadu do odchylky 10%. Model byl zvalidován pomocí validačních množin a označen za spolehlivý. V případě nedostatečného počtu vzorků (černá pojišťovna) jsme takové úspěšnosti nedosáhli, do odchylky 5% jsme byli schopni odhadnout pouze 44%, do 10% odchylky pak 75%. Model černé pojišťovny jsme se pokusili vylepšit znalostí cen z modelu bílé pojišťovny, avšak tyto vylepšení se na daných datech neprokázaly jako lepší než základní model.

V druhé části práce byla na tato data použita alternativní metoda, a to metoda neuronových sítí. Byl sestaven model jak pro bílou, tak pro černou pojišťovnu. Parametry modelu byly voleny na základě vstupních dat. Úspěšnost odhadu cen bílé pojišťovny pomocí této alternativní metody je srovnatelná se zobecněným lineárním modelem (79% úspěšnosti odhadu do odchylky 5%). Model černé pojišťovny nedosahoval takové úspěšnosti, avšak můžeme říci, že v případě malého počtu dat není potřeba kategorizovat data tak důkladně a hledět na významnost a nevýznamnost vstupních proměnných.

Lze tedy říci, že zobecněný lineární model dobře sedí na strukturu tarifů, které i dnes u převážné většiny pojistitelů mají formu součinů multiplikátorů závisejících na jednotlivých parametrech. Výhodou modelu je určitě snadná interpretace faktorů, avšak nevýhodou je nutnost důkladné předchozí analýzy proměnných. Ačkoli model s neuronovou sítí nevykazoval tak dobré výsledky jako zobecněný lineární model, jeho nespornou výhodou by mohla být větší volnost v prvotní analýze a rychlost predikce.

Pro malý počet vzorků a pro ceny pocházející z jednoduššího modelu výpočtu ceny se osvědčily spíše jednodušší metody. U rozsáhlejšího vzorku cen pak záleží na struktuře cen konkurence, zda volit jednodušší či složitější metody.

V této diplomové práci bohužel nebyla použita reálná data se zahrnutými bonusy a malusy, proto otázka úplné znalosti cen konkurence zůstává stále otevřena.

Literatura

[1] CIPRA, Tomáš. Pojistná matematika - teorie a praxe. EKOPRESS, s.r.o., Vydání I., 1999. ISBN 80-86119-17-3.

[2] Česká kancelář pojistitelů [online]. Dostupné z: https://www.ckp.cz/tiskove-centrum/tiskove-zpravy/137-vyvoj-povinneho-ruceni-2019-prumerne-motoriste-zaplati-2969-kc-za-povinne-ruceni-o-4-vice-nez-loni-cena-za-opravu-vozidel-ale-vzrostla-temer-o-7

[3] Slovník pojmů pojišťovnictví | Klik.cz. Online srovnávač pojištění | Klik.cz [online]. Dostupné z: https://www.klikpojisteni.cz/slovnik-pojmu/

[4] Česká kancelář pojistitelů [online]. Dostupné z: https://www.ckp.cz/o-povinnem-ruceni/slovnik-pojmu

[5] 168/1999 Sb. Zákon o pojištění odpovědnosti za škodu způsobenou provozem vozidla. Zákony pro lidi - Sbírka zákonů ČR v aktuálním konsolidovaném znění [online]. Dostupné z: https://www.zakonyprolidi.cz/cs/1999-168

[6] DUNN, Peter K., SMYTH, Gordon K. Generalized Linear Models With Examples in R. Springer, 2018. ISBN 9781441901187.

[7] de JONG, Piet, HELLER, Gillian Z. Generalized Linear Models for Insurance Data. Cambridge University Press, 2008. ISBN 9780511380983.

[8] McCULLAGH P., NELDER, J. A. Generalized Linear Models. 2nd Ed. Chapman and Hall, London, 1989. ISBN 978-0412317606

[9] SFU Mathematics and Statistics Web Server . SFU Mathematics and Statistics Web Server [online]. Dostupné z: http://people.stat.sfu.ca/ raltman/stat402/402L26.pdf

[10] Zápisky z přednášky Zobecněné lineární modely (ZLIM), FJFI ČVUT, doc. Ing. Tomáš Hobza, Ph.D.

[11] Coefficient of Determination (R Squared): Definition, Calculation - Statistics How To. Statistics How To: Elementary Statistics for the rest of us! [online]. Dostupné z: https://www.statisticshowto.com/probability-and-statistics/coefficient-of-determination-r-squared/

[12] BISHOP, Christopher M. Pattern recognition and machine learning. New York: Springer, c2006. Information science and statistics. ISBN 0-387-31073-8.

[13] How neural networks work - A simple introduction. Explain that Stuff [online]. Dostupné z: https://www.explainthatstuff.com/introduction-to-neural-networks.html

[14] FAUSETT, Laurene. Fundamentals of neural network. Architecture, algorithms and aplication, 1993. ISBN 978-0133341867

[15] A Gentle Introduction to the Rectified Linear Unit (ReLU). Machine Learning Mastery [online]. Dostupné z: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/

[16] Towards Data Science [online]. Dostupné z: https://towardsdatascience.com/neural-network-optimization-7ca72d4db3e0

[17] FAUSETT, L. Fundamentals of Neural Networks: Architectures, Algorithms And Applications. Prentice-Hall, 1994.

[18] 5 Step Life-Cycle for Neural Network Models in Keras. Machine Learning Mastery [online]. Dostupné z: https://machinelearningmastery.com/5-step-life-cycle-neural-network-models-keras/

[19] ROGALEWICZ, Vladimír. Pravděpodobnost a statistika pro inženýry. Vyd. 2., přeprac. V Praze: Nakladatelství ČVUT, 2007, c1998. ISBN 978-80-01-03785-0.

[20] NIELSEN, Michael. Neural networks and deep learning. Neural networks and deep learning [online]. Dostupné z: http://neuralnetworksanddeeplearning.com/