

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



APLIKÁCIA DATA MININGU V ZDRAVOTNÍCTVE

DIPLOMOVÁ PRÁCA

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

APLIKÁCIA DATA MININGU V ZDRAVOTNÍCTVE

DIPLOMOVÁ PRÁCA

Študijný program: Ekonomicko-finančná matematika a modelovanie
Študijný odbor: Aplikovaná matematika
Školiace pracovisko: Katedra aplikovanej matematiky a štatistiky
Vedúci práce: Mgr. Henrieta Tulejová, MSc.

Bratislava 2018

Bc. Annamária BODÓOVÁ



Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Annamária Bodóová
Študijný program: ekonomicko-finančná matematika a modelovanie
(Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Aplikácia data miningu v zdravotníctve.
Data mining in health care system.

Cieľ: Vybrať a popísať metódy data miningu nad anonymizovanými dátami zdravotnej poisťovne a použiť ich na riešenie vybraného problému v zdravotníctve.

Vedúci: Mgr. Henrieta Tulejová, MSc.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.
Dátum zadania: 25.01.2017

Dátum schválenia: 27.01.2017
prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Abstrakt v štátnom jazyku

BODÓOVÁ Annamária: Aplikácia data miningu v zdravotníctve [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Henrieta Tulejová, MSc., Bratislava, 2018, 70 strán

Cieľom tejto práce bolo vybrať a popísať metódy data miningu a aplikovať ich na vybraný problém v zdravotníctve - kompenzácie rizikovej štruktúry. Najskôr sme zaviedli pojem zdravotného rizika, vysvetlili sme význam kompenzácie rizikovej štruktúry a následne sme popísali metódy používané v práci. Zdôvodnili sme, prečo bolo potrebné zaviesť kompenzáciu rizikovej štruktúry na Slovensku a zadefinovali sme model farmaceuticko-nákladových skupín, ktorý sme potom aplikovali na dáta zdravotnej poisťovne. Zistili sme, že model má dobrú predikčnú silu, ale táto predikčná sila je významne limitovaná pre nové pozorovania. V štvrtej kapitole sme zisťovali, aké lieky by mali vstupovať do modelu farmaceuticko-nákladových skupín. Našli sme zoznam ATC skupín, ktoré považujeme za dobrých kandidátov na vstup do modelu. V poslednej kapitole sme nakoniec identifikovali skupinu takých poistencov, ktorí sú podkompenzovaní a navrhli sme riešenie ako zlepšiť predikciu ich nákladov.

Kľúčové slová: kompenzácia rizikovej štruktúry, farmaceuticko-nákladové skupiny, lineárna regresia

Abstract

BODÓOVÁ Annamária: Data mining in health care system [Master thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Henrieta Tulejová, MSc., Bratislava, 2018, 70 pages

The aim of this thesis was to select and describe data mining methods and to apply them to the selected problem from the field of healthcare, which was risk equalisation. We first introduced the concept of health risk, explained the importance of risk equalisation, and then described the methods used in our thesis. We justified the necessity of risk equalisation in Slovakia and we defined a model of pharmacy-based cost groups that which then we applied to the data of the health insurance company. We have found that the model has a good predictive power, but this predictive power is limited, the model performs much worse for new observations. In the fourth chapter, we investigated which drugs should enter the model of pharmacy-based cost groups. We have found a list of ATC groups that we consider to be good candidates for entering the model as predictors. Finally, in the last chapter, we identified a group of people who are undercompensated and we proposed a solution to improve their cost prediction.

Keywords: risk equalisation, pharmacy-based cost groups, linear regression

Obsah

Úvod	7
1 Zdravotné riziko	9
1.1 Individuálne riziko	9
1.2 Akumulácia rizika a populačné riziko	10
1.3 Selekcia rizika	10
1.4 Oceňovanie rizika a kompenzácia rizikovej štruktúry	12
2 Analytické nástroje používané v práci	14
2.1 Lineárna regresia	14
2.1.1 Odhad metódou najmenších štvorcov	15
2.1.2 Odhad pre σ^2	16
2.1.3 Koeficient determinácie	17
2.1.4 Gauss-Markovova veta	18
2.1.5 Testovanie lineárnych hypotéz	19
2.1.6 Diagnostika modelu	20
2.1.7 Box-Coxova transformácia	21
2.1.8 Lineárny model s obmedzeniami na parametre	21
2.2 Zhluková analýza	23
2.2.1 Funkcie podobnosti a vzdialenosti	23
2.2.2 Metóda DBSCAN	25
3 Model prerozdelenia poistného na Slovensku	26
3.1 Verejné zdravotné poistenie na Slovensku	26
3.1.1 Trh zdravotného poistenia na Slovensku	27
3.2 Model farmaceuticko-nákladových skupín	30
3.2.1 Farmaceuticko-nákladové skupiny	31
3.2.2 Model farmaceuticko-nákladových skupín	32
3.2.3 Prerozdelenie poistného	33
3.3 Odhad indexov rizika	34
3.3.1 Vlastnosti modelu farmaceuticko-nákladových skupín	34

3.3.2	Tranformácia modelu farmaceuticko-nákladových skupín	39
3.3.3	Verifikácia modelu farmaceuticko-nákladových skupín	40
4	Výber liekov do prerozdeľovacieho modelu	42
4.1	Dáta	42
4.2	Postup analýzy	43
4.3	Výsledok analýzy	46
5	Kompenzácia vybranej skupiny poistencov	47
5.1	Kompenzácia poistencov s menšou spotrebou liekov	47
5.2	Rozdiel medzi predikovanými a skutočnými nákladmi vybraných skupín	48
5.3	Obmedzenie pre predikovanú hodnotu v lineárnom modeli	49
5.4	Aplikácia obmedzení na model farmaceuticko-nákladových skupín . . .	51
	Záver	53
	Zoznam použitej literatúry	57
	Príloha A	60
	Príloha B	61
	Príloha C	62
	Príloha D	65
	Príloha E	67

Úvod

Úlohou poisťovní je preberanie rizika od iných subjektov v ekonomike. Nie je to inak ani v prípade zdravotných poisťovní. Je však ale jeden výrazný rozdiel medzi komerčnými a zdravotnými poisťovňami. Komerčné poisťovne môžu ešte pred uzavretím poisťnej zmluvy zhodnotiť nákladové riziko, a na základe očakávaných nákladov určiť výšku poistného.

V prípade zdravotného poistenia to však funguje inak. Zdravotné poistenie je založené na princípe solidarity, čo znamená, že výška poistného pre jednotlivca závisí len od výšky jeho príjmu, a nie od jeho rizikovosti, čiže od očakávaných nákladov na jeho zdravotnú starostlivosť. Na základe tohto princípu práve ekonomicky aktívni, väčšinou zdraví poistenci s nízkymi nákladmi na zdravotnú starostlivosť platia vysoké poistné, a dlhodobo chorí, ekonomicky neaktívni poistenci s vysokým nákladovým rizikom prispievajú najmenšou sumou. Takáto anomália na trhu by mohla viesť k selekcii rizika, čo znamená, že zdravotné poisťovne by mohli uprednostniť zdravých, ekonomicky aktívnych poistencov.

Úlohou zdravotného poistenia a zdravotných poisťovní by však malo byť to, aby zabezpečili kvalitnú zdravotnú starostlivosť každému poistencovi, a nediskriminovali nikoho kvôli jeho zdravotnému stavu. Preto je potrebné dodatočne upraviť príjmy poisťovní tak, aby tieto príjmy odzrkadlili rizikovosť ich kmeňov.

Riešením na tento problém je prerozdeľovanie poistného medzi zdravotnými poisťovňami na základe nákladovej rizikovosti ich kmeňov. Prvé modely prerozdeľovania sa objavili už v deväťdesiatich rokoch, dôležitým krokom však bolo zavedenie farmaceuticko-nákladových skupín v Holandsku v roku 2003. Odvtedy rôzne modely prerozdeľovania zaviedla už väčšina vyspelého sveta s viacerými zdravotnými poisťovňami na trhu.

Na Slovensku už dlhšie fungovalo rizikové váženie poistencov na základe demografických premenných ako vek a pohlavie, ktorých schopnosť vysvetliť variabilitu nákladov na zdravotnú starostlivosť je výrazne limitovaná. Preto sa v roku 2012 aj u nás zaviedli farmaceuticko-nákladové skupiny. Výhodou farmaceuticko-nákladových skupín je to, že sú schopné zachytiť, či daný poistenec má nejakú chorobu v chronickom štádiu na základe spotrebovaných liekov. V procese prerozdeľovania potom poisťovne majú väčšie príjmy za týchto chronicky chorých poistencov, ktorí znamenajú pre nich vyššie riziko,

vyššie očakávané náklady na zdravotnú starostlivosť.

Model prerozdelenia poistného funguje na Slovensku už vyše 5 rokov. Máme už k dispozícii množstvo dát, na základe ktorých je možné vyhodnotiť jeho výhody a nevýhody, a ako zmenil podmienky na trhu zdravotného poistenia. Cieľom našej práce je pozrieť sa bližšie na jeho výhody, nevýhody, navrhnúť riešenie jeho nedostatkov a určiť dopad jeho modifikácií na zdravotnú poisťovňu.

V prvej kapitole sa venujeme zdravotnému riziku, objasníme, čo predstavuje riziko pre zdravotné poisťovne, pričom vychádzame z [3] a [12]. Popíšeme rizikové faktory jednotlivca a celej populácie. Nakoniec vysvetlíme prečo selekcia rizika je nežiaduca na trhu zdravotného poistenia a prečo je dôležitý proces kompenzácie rizikovej štruktúry.

V druhej kapitole bližšie popíšeme metódy používané v práci, lineárnu regresiu a zhlukú analýzu vychádzajúc z [14], [5], [13] a [15].

Ďalej vysvetlíme ako vyzerá trh verejného zdravotného poistenia na Slovensku, prečo je opodstatnená kompenzácia rizikovej štruktúry a ako prebieha prerozdelenie poistného na základe [4], [16], [17], [18], [19] a [20] Aplikujeme model farmaceuticko-nákladových skupín na dáta zdravotnej poisťovne, odhadneme a vysvetlíme jednotlivé parametre. Bližšie sa pozrieme na vlastnosti modelu a validujeme ho.

Keďže anatomicko-terapeuticko-chemické skupiny, ktoré vstupujú do modelu, sú dané zákonom [20], v štvrtej kapitole sa pozrieme bližšie aké anatomicko-terapeuticko-chemické skupiny by mali vstupovať do modelu na základe našich dát a navrhnúť ich zmenu. Najprv zadefinujeme pravidlá, ktoré musia tieto skupiny spĺňať. Hlavnou podmienkou však je, aby lieky zoradené do týchto anatomicko-terapeuticko-chemických skupín vedeli jednoznačne spojiť s nejakou chronickou chorobou.

V piatej, poslednej kapitole sa pozrieme na podkompenzáciu zdravotných poisťovní za určitej podskupiny poistencov - chronicky chorých, ktorí ale nie sú zaradení do žiadnej farmaceuticko-nákladovej skupiny. Nakoniec upravíme model tak, aby aj táto skupina poistencov bola dostatočne kompenzovaná.

1 Zdravotné riziko

Na úvod našej práce by sme radi ujasnili pojem zdravotné riziko, keďže tento pojem je kľúčový pre našu prácu. Zaujíma nás, čo predstavuje riziko pre zdravotné poisťovne. Najskôr vysvetlíme, aké faktory vplývajú na rizikovosť jednotlivca. Ďalej popíšeme proces akumulácie rizika (ang. risk pooling), a ukážeme rozdiely medzi rizikovosťou jednotlivca a rizikovosťou populácie, čiže na základe akých faktorov sa posudzuje rizikovosť celého kmeňa poisťovne. Nakoniec vysvetlíme proces kompenzácie rizikovej štruktúry, a objasníme, prečo je dôležitý v prípade verejného zdravotného poistenia.

Čo je vôbec riziko? Jednoducho povedané, riziko nie je nič iné, ako potencinálna strata určitej hodnoty v budúcnosti. Je kombináciou dvoch faktorov: výškou straty a pravdepodobnosťou nastatia straty. V prípade zdravotného rizika strata vôbec nemusí znamenať len finančnú stratu. Literatúra hovorí, že udalosť ktorá nastala je stratou v prípade, ak stav jednotlivca po nastatí udalosti je menej priaznivý ako jeho pôvodný stav [3]. V rámci našej práce sa však zameriavame výlučne na finančnú stratu.

1.1 Individuálne riziko

Z pohľadu poisťovne je dôležité identifikovať vysoko rizikových jednotlivcov. Ako sme už vyššie spomenuli, v rámci našej práce vnímame zdravotné riziko z finančného hľadiska. Preto nás zaujíma, aké faktory vplývajú na očakávané náklady na zdravotnú starostlivosť jednotlivcov. Tieto rizikové faktory sme zaradili podľa [3] do piatich kategórií.

1. Vrodené rizikové faktory ako vek, pohlavie alebo rasa. Tieto rizikové faktory sú dané pre každého jednotlivca, sú fixné. Viacero štúdií [8], [12] ukázalo, že variabilita nákladov na zdravotnú starostlivosť naozaj závisí do určitej miery od vrodených rizikových faktorov.
2. Rizikové faktory súvisiace so zdravotným stavom (napr. diabetes alebo rakovina): jednotlivci s takýmito zdravotnými stavmi budú mať jednoznačne vyššie náklady na zdravotnú starostlivosť ako jednotlivci, ktorí takýto zdravotný stav nemajú.
3. Rodinná predispozícia: niektoré rizikové faktory súvisiace so zdravotným stavom sú dedičné. Takýto zdravotný stav je napr. hemofília, alebo určité typy rakoviny.

4. Rizikové faktory súvisiace so životným štýlom ako napr. fajčenie, stres, nedostatok pohybu alebo nedostatočná výživa prispievajú k vyšším nákladom na zdravotnú starostlivosť.
5. Exogénne rizikové faktory: pracovné podmienky jednotlivca, podmienky jeho bývania, jeho najvyššie dosiahnuté vzdelanie atď.

Zdravotné poisťovne však väčšinou disponujú výrazne limitovanými informáciami o svojich poistencoch - pre nich sú známe len niektoré vrodené rizikové faktory, čiže poznajú pohlavie a vek poistencov. Môžu mať informácie aj o zdravotnom stave svojich poistencov na základe čerpania zdravotnej starostlivosti. Ostatné rizikové faktory však nevedia vyhodnotiť.

1.2 Akumulácia rizika a populačné riziko

Populačné riziko je riziko určitej skupiny, v našom prípade riziko celého kmeňa danej poisťovne. Populačné riziko sa síce odvíja od rizikovosti jednotlivcov v danej skupine, ale kvôli rozloženiu rizika má iné charakteristiky. V tomto prípade sa totiž hovoríme o systéme akumulácie rizika (ang. risk pooling). To znamená, že riziká celej skupiny sa akumulujú do jedného celku, pričom každý člen skupiny prispieva určitou sumou do spoločného balíka. Z tohto balíka sú potom vykryté náklady na zdravotnú starostlivosť celej skupiny [12].

Dôležitými faktormi v prípade populačného rizika sú veľkosť a štruktúra danej skupiny. Ak je skupina jednotlivcov príliš malá, už jedna katastrofická udalosť môže spôsobovať vážne problémy. Čo sa týka štruktúry skupiny, je nevyhnutné, aby jej súčasťou boli nízko rizikovní jednotlivci, čiže jednotlivci s nízkymi očakávanými nákladmi. Práve oni sú totiž kľúčoví pri rozložení rizika. Aj keď sa v skupine nachádzajú takí jednotlivci, u ktorých sú prítomné vyššie spomenuté rizikové faktory, vďaka nízko rizikovým jednotlivcom rizikovosť celej skupiny nemusí byť významná [3].

1.3 Selekcia rizika

Ako sme už vyššie písali, každý poistenec prispieva určitou sumou na zdravotné poistenie, čiže do "balíka", z ktorého sa potom uhradzujú náklady na ich zdravotnú sta-

roastlivoť.

Výhodou takýchto systémov akumulácie rizika je ochrana jednotlivca pred nepredvídateľnými výdavkami na zdravotnú starostlivoť. Avšak, na druhej strane, ak jednotlivci majú právo výberu z viac než jednej zdravotnej poisťovne, takéto systémy môžu viesť k nerovnomernému rozloženiu rizikovej štruktúry medzi týmito poisťovňami [12].

Každá racionálne správajúca sa poisťovňa sa totiž snaží o to, aby docielila čím lepší hospodársky výsledok, čiže o maximalizáciu zisku. Tento cieľ sa dá dosiahnuť vytvorením zdravého, málo rizikového poistného kmeňa. V odbornej literatúre sa tento jav nazýva selekcia rizika.

K selekcii rizika dôjde vždy, keď majú jednotlivci právo výberu z viac než jednej zdravotnej poisťovne [12]. Dokonca aj v tom prípade, keď platí tzv. "open enrollment", t.j. keď zdravotné poisťovne zo zákona musia akceptovať každého poistenca, ktorý sa k nim prihlási.

Okrem toho, že vysoko rizikovní jednotlivci budú pre zdravotné poisťovne neatraktívni, má selekcia rizika ešte ďalšie nežiaduce dôsledky.

Po prvé, selekcia rizika znižuje kvalitu zdravotnej starostlivosti. Keďže zdravotné poisťovne nemajú motiváciu osloviť vysoko rizikových jednotlivcov, ani nebudú dbať na to, aby im poskytovali kvalitné služby. Môžu sa napríklad poisťovne rozhodnúť tak, že neuzavrú zmluvu so špičkovým špecialistom, čím odradia vysoko rizikových jednotlivcov. Potom tie zdravotné poisťovne, ktoré sa špecializujú na liečbu vysoko rizikových jednotlivcov, musia určiť oveľa vyššie poistné ako ostatné poisťovne, aby pokryli náklady na liečenie svojich poistencov. V dôsledku selekcii rizika majú teda vysoko rizikovní jednotlivci dve možnosti: buď príjmu zlé služby a zlú zdravotnú starostlivoť, alebo platia vysoké poistné, aby im bola poskytovaná kvalitná zdravotná starostlivoť. V krajinách, kde výška poistného je regulovaná štátom (ako napr. aj Slovesnko), poskytovanie kvalitných služieb vysoko rizikovým jednotlivcov môže priam znamenať samovraždu pre poisťovňu [9]

Po druhé, selekcia rizika znižuje efektívnosť zdravotnej starostlivosti. Na náklady na zdravotnú starostlivoť danej poisťovne totiž oveľa viac vplýva zdravotný stav poistencov, ako akékoľvek aktivity na zvýšenie efektivity poskytovanej zdravotnej starostlivosti. Až 35% poistencov nečerpá počas roka nijakú zdravotnú starostlivoť, kým 5%

najrizikovejších jednotlivcov tvorí až 50% nákladov [12].

1.4 Oceňovanie rizika a kompenzácia rizikovej štruktúry

Vidíme, že selekcia rizika je nežiadúci jav na trhu zdravotného poistenia a je potrebné dosiahnuť, aby aj vysoko rizikovní jednotlivci boli pre zdravotné poisťovne atraktívni. Tento problém rieši kompenzácia rizikovej štruktúry (ang. risk equalization). Jej cieľom je financovanie poisťovní v závislosti od rizikovosti ich kmeňa [12].

Predpokladá sa, že v systéme je dostatok zdrojov na pokrytie nákladov na zdravotnú starostlivosť každého jednotlivca, len sú nerovnomerne rozmiestnené. Preto je potrebné ich prerozdelenie medzi poisťovňami na základe rizikovosti ich kmeňov. Dostatočná kompenzácia poisťovní potom podporuje ich konkurenciu založenú na kvalite a efektívite poskytovaných služieb [12].

A ako funguje mechanizmus kompenzácie rizika? Najskôr sa ocení rizikovosť každého jednotlivca podľa rizikových faktorov z podkapitoly 1.1. Pripomíname, že ide o finančné riziko, takže sa vlastne odhadujú náklady na ich zdravotnú starostlivosť. Následne sa vyhodnocuje rizikovosť celého kmeňa zdravotnej poisťovne, na základe čoho sú potom medzi nimi prerozdelené dostupné zdroje.

V tomto celom mechanizme je kľúčové oceňovanie rizika jednotlivcov. Ako vybrať vhodné rizikové faktory? Na základe [9] a [10] zdefinujeme niekoľko podmienok, ktoré musia rizikové faktory spĺňať, aby mohli vstupovať do mechanizmu kompenzácie rizikovej štruktúry.

- 1. Prediktívna hodnota** - Triviálnou podmienkou pre rizikové faktory je, aby aspoň do určitej miery vysvetlili budúce náklady na zdravotnú starostlivosť jednotlivca.
- 2. Validita** - Daný rizikový faktor musí silne súvisieť s nejakým chronickým stavom, ako napr. diabetes, rakovina, srdcovo-cievne ochorenia atď. Tento predpoklad je dôležitý preto, lebo poisťovne by mali byť kompenzované len za chronický stav svojich poistencov, čiže len za predvídateľné vysoké náklady.
- 3. Vhodnosť motivácií poisťovní** - Tento predpoklad hovorí, že zaradenie nejakého rizikového faktora do mechanizmu kompenzácie rizikovej štruktúry by nemalo znížiť motiváciu zdravotných poisťovní efektívne hospodáriť. Rizikové faktory, ktoré

sú založené napr. na predchádzajúcich nákladoch nie sú vhodné, lebo trestajú efektívne zdravotné poisťovne.

4. Realizovateľnosť - Ako sme spomínali v podkapitole 1.1, poisťovne majú limitované informácie o svojich poistencoch. Rizikové faktory, ktoré vstupujú do mechanizmu kompenzácie rizikovej štruktúry však musia byť ľahko dostupné - musia to byť také informácie, ktoré sú k dispozícii každej zdravotnej poisťovni o svojich poistencoch. Cieľom tejto podmienky je to, aby zbieranie informácií o poistencoch nebolo drahé.

2 Analytické nástroje používané v práci

Na začiatku našej práce zhrnieme metódy používané v práci - lineárnu regresiu a zhlukovú analýzu. V prípade lineárnej regresie sme vychádzali z kníh o lineárnej regresii [13], [14] a prednášok z Ekonometrie [15]. Snažili sme sa zameriavať hlavne na vlastnosti lineárneho modelu, aby sme v ďalšej časti našej práce mohli charakterizovať model farmaceuticko-nákladových skupín.

Druhú podkapitolu sme venovali zhlukovej analýze, pričom základnú teóriu sme prebrali z Bakalárskej práce autorky [2] a metódu DBSCAN sme popísali na základe knihy o Data Miningu [5]. Zhlukovanie sme využili v Kapitole 4.

2.1 Lineárna regresia

Lineárny model definujeme ako

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (1)$$

kde

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \text{ je vysvetľovaná premenná, } \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \text{ je náhodná odchýlka,}$$
$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \text{ sú vysvetľujúce premenné,}$$
$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \text{ je neznámy parameter.}$$

2.1.1 Odhad metódou najmenších štvorcov

Metóda najmenších štvorcov je založená na predpoklade, že čím menšie sú rezíduá, tým lepší je model. Zo všetkých možných hodnôt β odhad metódou najmenších štvorcov (označme $\hat{\beta}$) minimalizuje súčet štvorcov rezíduí.

$$\begin{aligned}\hat{\beta} &= \underset{\beta}{\operatorname{argmin}} S \\ &= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik})^2 \\ &= \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \\ &= \underset{\beta}{\operatorname{argmin}} \mathbf{y}^T \mathbf{y} - \beta^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \beta + \beta^T \mathbf{X}^T \mathbf{X} \beta \\ &= \underset{\beta}{\operatorname{argmin}} \mathbf{y}^T \mathbf{y} - 2\beta^T \mathbf{X}^T \mathbf{y} + \beta^T (\mathbf{X}^T \mathbf{X}) \beta\end{aligned}\tag{2}$$

Keďže úloha (2) je konvexná, jej riešenie môžeme nájsť pomocou gradientu

$$\frac{\partial S}{\partial \beta} = -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \beta.\tag{3}$$

Úlohu 2 minimalizuje ten parameter $\hat{\beta}$, pre ktorý je gradient (3) nulový.

Tvrdenie 2.1. *Nech matica $\mathbb{X}$ je regulárna. Potom odhad metódou najmenších štvorcov je jednoznačne daný formulou*

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.\tag{4}$$

Otázkou zostáva, či odhad metódou najmenších štvorcov je najlepší možný odhad. Pri odpovedi nám pomôžu Gauss-Markovove podmienky.

Definícia 2.2. *Gauss-Markovove podmienky definujeme ako*

- $E(\epsilon) = 0$,
- $E(\epsilon\epsilon^T) = \sigma^2 I$.

Všimnime si, že za platnosti Gauss-Markovových podmienok platí, že

$$E(\mathbf{y}) = \mathbf{X}\beta\tag{5}$$

a

$$\operatorname{Var}(\mathbf{y}) = E[(\mathbf{y} - \mathbf{X}\beta)(\mathbf{y} - \mathbf{X}\beta)^T] = E(\epsilon\epsilon^T) = \sigma^2 I.\tag{6}$$

Tvrdenie 2.3. *Nech $\hat{\beta}$ je odhad metódou najmenších štvorcov a nech platia Gauss-Markovove podmienky 2.2. Potom pre strednú hodnotu a varianciu $\hat{\beta}$ platí*

- $E(\hat{\beta}) = \beta$,
- $Var(\hat{\beta}) = \sigma^2(X^T X)^{-1}$.

Dôkaz. Ukážeme platnosť Tvrdenia 2.3. Pri prvej časti tvrdenia o strednej hodnote odhadu metódou najmenších štvorcov využijeme (5).

$$E(\hat{\beta}) = E[(X^T X)^{-1} X^T y] = (X^T X)^{-1} X^T X \beta = \beta.$$

Nech $A = (X^T X)^{-1} X^T$. Potom podľa (4) $\hat{\beta} = Ay$ a zo vzťahu (6) dostaneme

$$\begin{aligned} Var(\hat{\beta}) &= A Var(y) A^T = \sigma^2 A I A^T = \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1}. \end{aligned}$$

□

2.1.2 Odhad pre σ^2

Snažíme sa nájsť nevychýlený odhad pre σ^2 , aby sme mohli odhadnúť varianciu odhadu $\hat{\beta}$.

Nech $\hat{y} = X\hat{\beta}$ je odhad pre y a nech $\hat{\epsilon} = y - \hat{y}$ sú rezíduá. Zadefinujme $M = I - X(X^T X)^{-1} X^T$. Táto matica je symetrická, idempotentná a pozitívne semidefinitná. Pre rezíduá potom platí

$$\begin{aligned} \hat{\epsilon} &= y - \hat{y} = y - X(X^T X)^{-1} X^T y = My \\ &= MX\beta + M\epsilon = M\epsilon, \end{aligned} \tag{7}$$

keďže $MX = (I - X(X^T X)^{-1} X^T)X = X - X = 0$.

Definícia 2.4. *Reziduálny súčet štvorcov (ang. residual sum of squares) definujeme ako*

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \hat{\epsilon}_i^2 = \hat{\epsilon}^T \hat{\epsilon}.$$

Z výrazu (7), z Definície 2.4 a z vlastnosti matice $\mathbf{M}$ máme potom

$$\sum_{i=1}^n \hat{\epsilon}_i^2 = \hat{\epsilon}^T \hat{\epsilon} = \epsilon^T \mathbf{M}^T \mathbf{M} \epsilon = \epsilon^T \mathbf{M} \epsilon = \sum_{i=1}^n m_{ii} \epsilon_i^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^n m_{ij} \epsilon_i \epsilon_j. \quad (8)$$

Pre strednú hodnotu $\mathbf{RSS}$ potom platí

$$\begin{aligned} E(\hat{\epsilon}^T \hat{\epsilon}) &= \sum_{i=1}^n m_{ii} E(\epsilon_i^2) + \sum_{\substack{i,j=1 \\ i \neq j}}^n m_{ij} E(\epsilon_i \epsilon_j) \\ &= \sigma^2 \sum_{i=1}^n m_{ii} = \sigma^2 \text{tr}(\mathbf{M}) = (n - k - 1) \sigma^2, \end{aligned} \quad (9)$$

keďže $\text{tr}[\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] = \text{tr}[\mathbf{I}_n] - \text{tr}[\mathbf{X}^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1}] = \text{tr}[\mathbf{I}_n] - \text{tr}(\mathbf{I}_{k+1}) = n - k - 1$ pre model s interceptom.

Preto

$$s^2 = \sum_{i=1}^n \frac{\hat{\epsilon}_i^2}{n - k - 1} \quad (10)$$

je nevychýlený odhad pre σ^2 pre model s interceptom.

Podobne

$$s^2 = \sum_{i=1}^n \frac{\hat{\epsilon}_i^2}{n - k} \quad (11)$$

je nevychýlený odhad pre σ^2 pre model bez interceptu.

Keďže sme našli nevychýlený odhad pre σ^2 , máme aj odhad pre varianciu odhadu $\hat{\beta}$.

Tvrdenie 2.5. *Nech platia Gauss-Markovove podmienky (Definícia 2.2). Potom s^2 je nevychýlený odhad pre σ^2 a $s^2(\mathbf{X}^T \mathbf{X})^{-1}$ je nevychýlený odhad pre $\mathbf{Var}(\hat{\beta})$.*

2.1.3 Koefficient determinácie

Nech $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ je odhad pre $\mathbf{y}$ a nech $\hat{\epsilon} = \mathbf{y} - \hat{\mathbf{y}}$ sú rezíduá.

Definícia 2.6. *Celkový súčet štvorcov (ang. total sum of squares) definujeme ako*

$$CSS = \sum_{i=1}^n (y_i - \bar{y})^2,$$

kde $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ je priemer vysvetľovanej premennej.

Definícia 2.7. *Vysvetľovaný súčet štvorcov (ang. explained sum of squares) definujeme ako*

$$VSS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

Celkovú variabilitu vysvetľovanej premennej $\mathbf{y}$ vyjadruje celkový súčet štvorcov. Táto variabilita vysvetľovanej premennej môže mať dve príčiny. Prvá príčina je variabilita vysvetľujúcej premennej $\mathbf{X}$, ktorú popisuje vysvetľovaný súčet štvorcov, táto variabilita sa totiž dá odhadnúť pomocou modelu. Druhým dôvodom je variabilita náhodnej chyby ϵ modelu. Túto variabilitu lineárny model nie je schopný vysvetliť, vyjadrujeme ho pomocou rezíduí, teda pomocou reziduálneho súčtu štvorcov RSS .

Definícia 2.8. *Koeficient determinácie definujeme ako*

$$R^2 = \frac{VSS}{CSS}.$$

Tvrdenie 2.9. *Nech R^2 je koeficient determinácie z Definície 2.8. Potom platí*

- $R^2 \in \langle 0, 1 \rangle$,
- $R^2 = 1 - \frac{RSS}{CSS}$.

Toto tvrdenie uvedieme bez dôkazu (dôkaz nájdeme v [15]).

Vidíme, že čím väčší je koeficient determinácie R^2 , tým väčšiu variabilitu vysvetľovanej premennej $\mathbf{y}$ vyjadruje model. Preto koeficient determinácie určuje kvalitu modelu.

2.1.4 Gauss-Markovova veta

Vo väčšine aplikácií lineárnej regresie nás zaujímajú odhady určitých lineárnych funkcií $\mathbf{l}^T \beta$, kde $\mathbf{l}$ je vektor. Odhady takéhoto typu sú aj napríklad $\hat{y}_i$, odhad nejakého budúceho pozorovania $\hat{y}_0$, alebo aj samotná $\hat{\beta}$.

Síce môžu existovať rôzne odhady, my sa obmedzíme na lineárne odhady - t.j. na odhady, ktoré sú lineárnou funkciou $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$, v tvare $\mathbf{c}^T \mathbf{y}$. Tiež predpokladáme, že tieto lineárne funkcie sú nevychýlené odhady pre $\mathbf{l}^T \beta$.

Odhad metódou najmenších štvorcov $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ je tiež lineárnou funkciou $\mathbf{y}$ a vyššie sme už ukázali, že je nevychýlený (5). Nasledujúca veta nám hovorí, že zo všetkých lineárnych nevychýlených odhadov práve odhad metódou najmenších štvorcov má najmenšiu varianciu.

Tvrdenie 2.10. *Nech $\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$ a $\mathbf{y} = X\beta + \epsilon$. Potom za platnosti Gauss-Markovových podmienok $l^T \hat{\beta}$ je najlepší lineárny nevychýlený odhad pre $l^T \beta$.*

Dôkaz. Nech $c^T \mathbf{y}$ je ďalší lineárny nevychýlený odhad pre $l^T \beta$. Potom $l^T \beta = E(c^T \mathbf{y}) = c^T X\beta$ pre všetky β , z čoho dostaneme

$$c^T X = l^T. \quad (12)$$

Z (6) a (12) vieme vyjadriť variancie odhadov

$$\text{Var}(c^T \mathbf{y}) = c^T \text{Var}(\mathbf{y}) c = c^T (\sigma^2 I) c = \sigma^2 c^T c, \quad (13)$$

a

$$\text{Var}(l^T \hat{\beta}) = l^T \text{Var}(\hat{\beta}) l = \sigma^2 l^T (X^T X)^{-1} l = \sigma^2 c^T X (X^T X)^{-1} X^T c. \quad (14)$$

Teraz pre rozdiel (13) a (14) platí

$$\begin{aligned} \text{Var}(c^T \mathbf{y}) - \text{Var}(l^T \hat{\beta}) &= \sigma^2 [c^T c - c^T X (X^T X)^{-1} X^T c] \\ &= \sigma^2 c^T [I - X (X^T X)^{-1} X^T] c \geq 0. \end{aligned}$$

Keďže matica $M = I - X (X^T X)^{-1} X^T$ je pozitívne semidefinitná, Tvrdenie 2.10 sme dokázali.

□

2.1.5 Testovanie lineárnych hypotéz

Po ohadnutí parametrov β často nás zaujímajú ich rôzne vlastnosti, napr. či vôbec vysvetľujú varianciu $\mathbf{y}$. Tieto vlastnosti vieme sformulovať formou hypotéz v tvare

$$H : C\beta - \gamma = 0, \quad (15)$$

kde C je $m \times (k+1)$ matica s hodnotou m , pričom $m < (k+1)$.

Tvrdenie 2.11. *Predpokladajme, že náhodné odchýlky ϵ_i sú nezávislé a pochádzajú z rovnakého normálneho rozdelenia*

$$\epsilon_i \sim N(0, \sigma^2).$$

Potom hypotéza (15) je zamietnutá na hladine významnosti α , ak

$$\frac{m^{-1}(C\hat{\beta} - \gamma)^T [C(X^T X)^{-1} C^T]^{-1} (C\hat{\beta} - \gamma)}{s^2} \geq F_{m, n-k-1, \alpha} \quad (16)$$

kde $F_{m, n-k-1, \alpha}$ je $100 - \alpha$ kvantil Fisherovho-Snedecorovo rozdelenia so stupňami voľnosti m a $n - k - 1$,

$$\hat{\beta} = (X^T X)^{-1} X^T y \text{ a } s^2 = (n - k - 1)^{-1} y^T [I - X(X^T X)^{-1} X^T] y.$$

Dôkaz Tvrdenia 2.11 nájdeme v knihe [14].

2.1.6 Diagnostika modelu

Vyššie sme dokázali, že odhad metódou najmenších štvorcov je najlepší lineárny nevychýlený odhad, ak sú splnené Gauss-Markovove podmienky. Ukázali sme tiež, že ak sa rezíduá dokonca riadia normálnym rozdelením, veľmi efektívne vieme testovať lineárny hypotézy o odhadoch. Tieto predpoklady však nemusia byť vždy splnené, dokonca vo väčšine prípadov aspoň jeden z predpokladov je porušený.

Heteroskedasticita Gauss-Markovove podmienky hovoria, že náhodné odchýlky ϵ_i musia mať rovnakú varianciu σ^2 (Definícia 2.2). Porušenie tohto predpokladu sa volá heteroskedasticita.

Z Tvrdenia 2.3 vyplýva, že prítomnosť heteroskedasticity v modeli ešte neznamená, že odhad metódou najmenších štvorcov je vychýlený, no $s^2(X^T X)^{-1}$ už naďalej nebude nevychýlený odhad pre jeho varianciu, čo môže ovplyvniť koeficient determinácie R^2 , aj testy lineárnych hypotéz.

Korelované náhodné odchýlky Podľa Gauss-Markovových podmienok platí, že

$E(\epsilon_i \epsilon_j) = 0$ pre $\forall i \neq j$. Keď totiž táto vlastnosť neplatí pre náhodné odchýlky, tak

$$E(\epsilon \epsilon^T) = \sigma^2 \Omega, \quad (17)$$

čo je porušením Gauss-Markovových podmienok. Podľa Tvrdenia 2.3 ani kvôli korelovaným náhodným odchýlkam nebude odhad $\hat{\beta}$ vychýlený. Aj v tomto prípade je problém s varianciou, keďže už malé nediagonálne prvky v matici Ω môžu spôsobovať výrazné zvýšenie variance odhadu $\hat{\beta}$.

Normalita Vyššie sme videli, že normalita náhodných odchýlok je dôležitým predpokladom v prípade lineárnych hypotéz. Preto aj keď tento predpoklad nie je splnený, odhad $\hat{\beta}$ je stále nevychýlený a ani jeho variancia nebude ovplyvnená. Výrazne však sťažuje testovanie lineárnych hypotéz, ako napr. skúmanie signifikantnosti odhadnutých parametrov, alebo signifikantnosť modelu.

2.1.7 Box-Coxova transformácia

Keď Gauss-Markovove podmienky neplatia pre lineárny model, nelineárna transformácia vysvetľovanej premennej $\mathbf{y}$ môže viesť k lepšiemu splneniu podmienok. To je aj cieľom Box-Coxovej transformácie - zabezpečiť, aby $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I})$.

Box-Coxova transformácia je definovaná pre $\mathbf{y}_i > 0$ ako trieda transformácií v tvare

$$\mathbf{y}_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \text{ak } \lambda \neq 0 \\ \log(y_i) & \text{ak } \lambda = 0, \end{cases} \quad (18)$$

kde $i = 1, 2, \dots, n$.

Vhodný parameter λ nájdeme metódou maximálnej vierohodnosti.

2.1.8 Lineárny model s obmedzeniami na parametre

V tejto podkapitole predpokladáme, že máme lineárny model (1 s obmedzením pre parameter β v tvare

$$\mathbf{C}\beta - \mathbf{d} = \mathbf{0}, \quad (19)$$

kde $\mathbf{C}$ je $\mathbf{m} \times (\mathbf{k} + 1)$ matica s hodnotou $\mathbf{m}$ a $\mathbf{m} < (\mathbf{k} + 1)$.

Odhad pre parameter β chceme nájsť aj v tomto prípade metódou najmenších štvorcov. Nech $\lambda = (\lambda_1, \dots, \lambda_m)$ je Lagrangeov multiplikátor. Potom máme

$$\frac{\partial}{\partial \lambda} [S + \lambda^T (\mathbf{d} - \mathbf{C}\beta)] = \mathbf{d} - \mathbf{C}\beta \quad (20)$$

a

$$\frac{\partial}{\partial \beta} [S + \lambda^T (\mathbf{d} - \mathbf{C}\beta)] = 2(\mathbf{X}^T \mathbf{X})\beta - 2\mathbf{X}^T \mathbf{y} - \mathbf{C}^T \lambda \quad (21)$$

Odhad metódou najmenších štvorcov nájdeme položením rovníc (20) a (21) rovné 0. Za predpokladu, že matica $\mathbf{X}^T \mathbf{X}$ je regulárna, odhad metódou najmenších štvorcov je

$$\hat{\beta}_{OBM} = (\mathbf{X}^T \mathbf{X})^{-1} [\mathbf{X}^T \mathbf{y} + \frac{1}{2} \mathbf{C}^T \hat{\lambda}] = \hat{\beta} + \frac{1}{2} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \hat{\lambda} \quad (22)$$

Z rovnice (20) máme potom

$$\mathbf{d} = \mathbf{C}\hat{\boldsymbol{\beta}}_{OBM} = \mathbf{C}\hat{\boldsymbol{\beta}} + \frac{1}{2}\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\hat{\boldsymbol{\lambda}} \quad (23)$$

Po vyjadrení $\hat{\boldsymbol{\lambda}}$ z (23) a dosadení do (22) dostaneme odhad metódou najmenších štvorcov s ombedzeniami pre $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}_{OBM} = \hat{\boldsymbol{\beta}} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}(\mathbf{d} - \mathbf{C}\hat{\boldsymbol{\beta}}). \quad (24)$$

2.2 Zhluková analýza

Cieľom zhlukovej analýzy je vytvorenie skupiny z daných objektov na základe ich spoločných črtov. Z n objektov skonštruujú k zhlukov, čiže ich rozdeľujú do k skupín. Musia však byť splnené dve podmienky: každá skupina musí obsahovať aspoň jeden objekt, a každý objekt môže byť presne v jednej skupine. Vo všeobecnosti sa metódy zhlukovej analýzy snažia o to, aby podobné objekty boli v jednej skupine a objekty v rozličných skupinách boli čo najviac odlišné.

Existujú viaceré metódy zhlukovej analýzy, vo všeobecnosti ich však môžeme rozdeliť na dve skupiny:

1. Nehierarchické metódy

Nehierarchické metódy sa používajú, keď dané objekty chceme roztriediť do k skupín, kde k je fixné. Najznámejšie nehierarchické metódy sú metódy k -priemerov a k -medoidov.

2. Hierarchické metódy

Hierarchické metódy nevytvoria jediné rozdelenie n objektov do k zhlukov, ale určujú rozdelenie objektov pre každú možnú hodnotu k , čiže pre $k = 1, \dots, n$.

2.2.1 Funkcie podobnosti a vzdialenosti

Ako sme už vyššie spomenuli metódy zhlukovej analýzy sa snažia o to, aby najpodobnejšie objekty boli v jednej skupine. Čo presne znamená, že objekty sú podobné, vysvetlíme pomocou funkcie podobnosti a funkcie vzdialenosti.

Definícia 2.12. *Nech je daných n objektov $O_1, \dots, O_n$ a $\forall i$ platí $O_i \in \mathbb{R}^m$. Funkcia $s : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ sa nazýva funkcia podobnosti, ak pre $\forall i, j \in \{1, 2, \dots, n\}$ splňa nasledujúce predpoklady:*

1. $0 \leq s(i, j) \leq 1,$

2. $s(i, i) = 1,$

3. $s(i, j) = s(j, i).$

Na určenie podobností sa používa väčšinou Pearsonov alebo Spearmanov korelačný koeficient. Čím väčšia je teda korelácia medzi objektmi, tým väčšiu hodnotu má funkcia podobnosti. Keďže korelačné koeficienty môžu nadobudnúť aj záporné hodnoty, nie je možné ich priamo použiť ako funkciu podobnosti. Často sa preto používa nasledujúca transformácia.

$$s(i, j) = \frac{1 + R(i, j)}{2},$$

kde $R(i, j)$ môže byť Pearsonov alebo Spearmanov korelačný koeficient. Vidíme, že keď korelačný koeficient má hodnotu -1, ich funkcia podobnosti má hodnotu 0. Občas potrebujeme, aby aj negatívne korelované objekty boli v jednom zhľuku. Vtedy môžeme použiť transformáciu:

$$s(i, j) = |R(i, j)|.$$

Veľa metód zhľukovej analýzy však na určenie jednotlivých zhľukov nepoužíva funkciu podobnosti, ale funkciu vzdialenosti, ktorú definujeme nasledovne:

Definícia 2.13. *Nech je daných n objektov $O_1, \dots, O_n$ a $\forall i$ platí $O_i \in \mathbb{R}^m$. Funkcia $d : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ sa nazýva funkcia vzdialenosti, ak pre $\forall i, j, h \in \{1, 2, \dots, n\}$ spĺňa nasledujúce predpoklady:*

1. $d(i, j) \geq 0$,
2. $d(i, i) = 0$,
3. $d(i, j) = d(j, i)$,
4. $d(i, j) \leq d(i, h) + d(h, j)$.

Lahko vidíme, že na základe tejto definície je funkciou vzdialenosti napríklad Minkowského norma

$$M(x, y) = \left(\sum_{k=1}^m (x_k - y_k)^q \right)^{1/q}. \quad (25)$$

Špeciálnym prípadom Minkowskeho normy pre $q = 2$ je Euklidova norma. Práve Euklidova norma je najpopulárnejšou funkciou vzdialenosti, keďže reprezentuje skutočnú geometrickú vzdialenosť.

$$E(x, y) = \sqrt{\sum_{k=1}^m (x_k - y_k)^2}. \quad (26)$$

2.2.2 Metóda DBSCAN

Názov metódy DBSCAN vznikol ako skratka z anglického výrazu *Density-based spatial clustering of applications with noise*, čiže je to zhukovacia metóda založená na hustote pre aplikácie so šumom.

Výhodou metód založených na hustote je to, že pomocou nich vieme nájsť zhluky ľubovoľného tvaru - tieto tvary nemusia byť nutne konvexné. Sú založené na predpoklade, že v priestore existujú husté oblasti - kde objekty sú blízko k sebe, a tieto husté oblasti sú oddelené riedkymi oblasťami.

Hustotu objektu O pričom môžeme merať ako počet objektov v blízkosti O . Metóda DBSCAN nájde hlavné objekty, teda tie objekty, ktoré majú husté okolie - v ich okolí sa nachádza veľa iných objektov. Zhluky potom vytvára tak, že hlavné objekty spojí s ich okolím.

Ako kvantifikuje ale metóda DBSCAN okolie objektu? Užívateľom zadaný parameter ϵ určuje ϵ -okolie daného objektu. Hustotu objektu O potom definujeme ako počet objektov v jeho ϵ -okolí.

Existuje ešte ďalší parameter, ktoré potrebujeme zadávať pri aplikovaní tejto metódy. Tento parameter sa volá *MinPts*, ktorý vlastne určí hranicu hustoty. Objekt O je potom hlavný objekt, ak jeho ϵ -okolie obsahuje aspoň *MinPts* objektov.

Metóda DBSCAN potom vytvára také zhluky, kde hustota je aspoň *MinPts*.

3 Model prerozdelenia poistného na Slovensku

V rámci tejto kapitoly popíšeme model kompenzácie rizikovej štruktúry na Slovensku. Najskôr charakterizujeme trh verejného poistenia na základe článku [4], v ktorom sa autori zaoberajú s verejným zdravotným poistením na Slovensku a Správy o zdravotnom poistení [16], ktorú vydáva každoročne Úrad pre dohľad nad zdravotnou starostlivosťou. Cieľom tejto charakteristiky je, aby sme pochopili, prečo bolo potrebné zaviesť model kompenzácie rizikovej štruktúry na Slovensku. V ďalšej podkapitole popíšeme model, na základe ktorého sa kompenzujú zdravotné poisťovne, pričom vychádzame zo Zákona č. 580/2004 Z. z. o zdravotnom poistení [20]. V poslednej časti aplikujeme tento model na dáta zdravotnej poisťovne, analyzujeme jeho vlastnosti a validujeme ho.

3.1 Verejné zdravotné poistenie na Slovensku

Úlohu verejného zdravotného poistenia na Slovensku definuje Zákon č. 580/2004 Z. z. o zdravotnom poistení [20] ako poistenie, na základe ktorého sa poskytuje zdravotná starostlivosť. Ak poistenec spĺňa predpoklady stanovené v zákone, má právo na úhradu zdravotnej starostlivosti. Verejné zdravotné poistenie teda funguje ako systém akumulácie rizika z Kapitoly 1, a jeho cieľom je ochrana jednotlivcov pred nepredvídateľnými nákladmi na ich zdravotnú starostlivosť.

Povinné verejné zdravotné poistenie na Slovensku je založené na princípe solidarity. Výška poistného vôbec nezávisí od rizikovosti daného poistenca, rozhodujúci je jeho príjem. Čím vyšší príjem má, tým väčšou sumou prispieva na verejné zdravotné poistenie. Jednoducho povedané, každý prispieva takou sumou, akú si môže dovoliť, aby boli pokryté náklady na zdravotnú starostlivosť každého jednotlivca.

Najväčšou sumou na zdravotné poistenie prispievajú ekonomicky aktívni poistenci, čiže poistenci, ktorí vykonávajú zárobkovú činnosť. Do tejto skupiny patria zamestnanci, samostatne zárobkovo činné osoby, alebo poistenci, ktorí majú príjem z dividend. V prípade zamestnanca je to napríklad 14% z jeho hrubej mzdy, pričom 10% platí zamestnávateľ a 4% on sám [20].

V prípade ekonomicky neaktívny poistencov, ktorí nemajú vlastný príjem zo zárob-

kovej činnosti, hradí poistné na verejné zdravotné poistenie štát. Sem patria napríklad nezaopatrené deti, osoby preberajúce starobný alebo invalidný dôchodok, alebo osoby, ktoré preberajú rodičovský príspevok. (Tento zoznam je však oveľa dlhší, presné podmienky nájdeme v Zákone č. 580/2004 Z. z. [20].) Suma, ktorou prispieva štát, je vo všeobecnosti oveľa nižšia. Za rok 2018 sadzba poistného pre štát je 3,71% z vymeriavacieho základu [17].

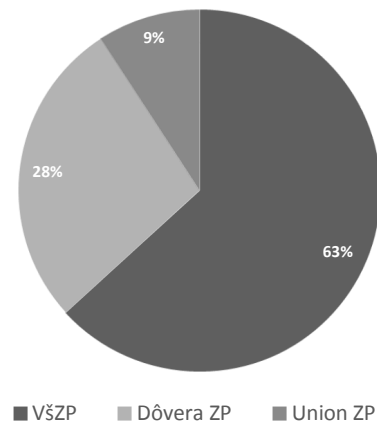
Aby sme lepšie pochopili rozdiel vo výške poistného pre ekonomicky aktívnych a neaktívnych poistencov, uvedieme presný údaj z najnovšej Správy Úradu pre dohľad nad zdravotnou starostlivosťou [16]. V roku 2016 ročné predpísané poistné prepočítané na jedného ekonomicky aktívneho poistenca bolo 1 401,73 eur, kým na jedného poistenca štátu 461,83 eur. Čiže ekonomicky neaktívni poistenci prispievali na verejné zdravotné poistenie o 67,1% menšou sumou.

V skutočnosti však práve vysoko rizikovní poistenci nemajú možnosť vykonávať zárobkovú činnosť kvôli svojej chorobe. To znamená, že sa stanú poistencami štátu, ktorý za nich prispieva na verejné zdravotné poistenie len s nízkou sumou. Solidarita systému sa spočíva v tom, že na náklady na ich zdravotnú starostlivosť sa poskladajú ostatní poistenci. Teda tí, ktorí prispievajú väčšou sumou ako sú ich náklady na zdravotnú starostlivosť.

3.1.1 Trh zdravotného poistenia na Slovensku

Napriek tomu, že Slovensko je pomerne malá krajina, jeho trh verejného zdravotného poistenia je pluralitný, čiže na Slovensku pôsobí viacero konkurenčných verejných zdravotných poisťovní: štátna Všeobecná zdravotná poisťovňa a.s., a súkromné Dôvera zdravotná poisťovňa a.s. a Union zdravotná poisťovňa a.s.

Podľa poslednej Správy o stave vykonávania verejného zdravotného poistenia [16], ktorú vydáva každý rok Úrad pre dohľad nad zdravotnou starostlivosťou, najväčší trhovú podiel (63,19%) má Všeobecná zdravotná poisťovňa s počtom poistencov 3 252 794. Druhý v rebríčku je Dôvera zdravotná poisťovňa (trhový podiel 27,66%, počet poistencov 1 423 956) a najmenšia zdravotná poisťovňa na Slovensku je Union (trhový podiel 9,14%, počet poistencov 470 658). Trhový podiel jednotlivých zdravotných poisťovní ukážeme na Obrázku 1.



Obr. 1: Trhový podiel jednotlivých zdravotných poisťovní [16]

Pluralita trhu verejného zdravotného poistenia má svoje výhody aj nevýhody. Veľkou výhodou je slobodný výber zdravotnej poisťovne. Každý jednotlivec si môže na základe svojich preferencií vybrať tú zdravotnú poisťovňu, ktorá mu najviac vyhovuje. Čiže sa uplatňujú princípy voľného trhu, a poisťovne by mali byť prinútené k poskytovaniu kvalitnejších služieb a k zvyšovaniu efektivity.

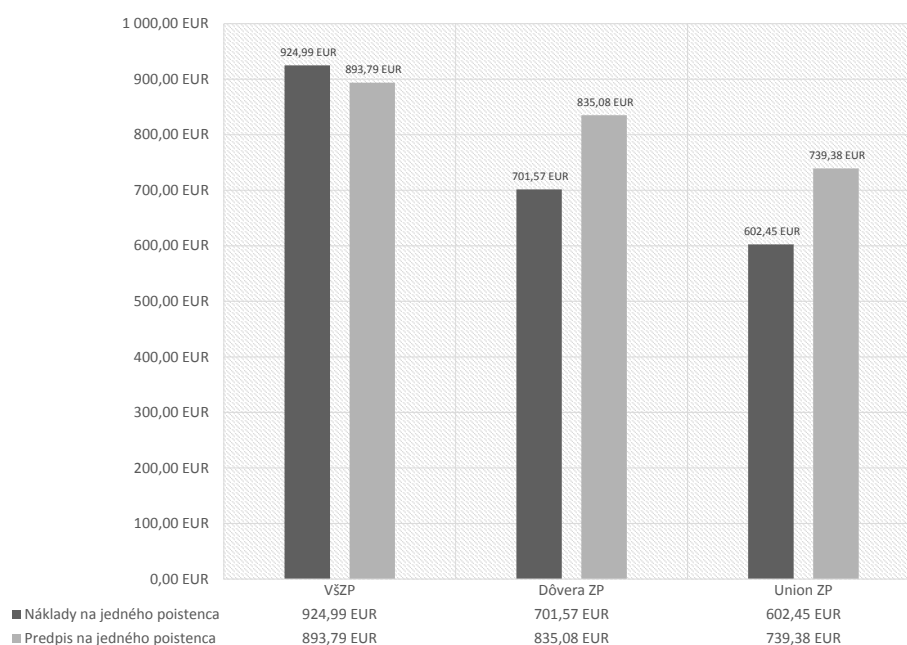
Ako sme však už v Kapitole 1 spomenuli, k selekcii rizika dôjde vždy, keď majú jednotlivci právo výberu z viac než jednej zdravotnej poisťovne. Selekcia rizika pričom spôsobí práve opak - v jej dôsledku sa znižuje kvalita a efektívnosť poskytovaných služieb.

Prítomnosť selekcie rizika naznačuje aj rozdiel v štruktúre poistného kmeňa jednotlivých poisťovní. Na Obrázku 2 sme ilustrovali ročné náklady a ročný predpis poistného na jedného poistenca pre každú zdravotnú poisťovňu. Dáta máme zo Správy Úradu pre dohľad na zdravotnou starostlivosťou za rok 2016 [16].

Vidíme, že Všeobecná zdravotná poisťovňa mala v roku 2016 výrazne vyššie priemerné náklady, čiže má viac rizikový poistný kmeň ako ostatné dve súkromné zdravotné poisťovne. Mala aj vyšší priemerný predpis na jedného poistenca, no aj tak nestačil na pokrytie nákladov - chýbali jej zdroje vo výške 31,2 eur na poistenca. Keď zohľadňujeme

počet jej poistencov, bez kompenzácie rizikovej štruktúry by mala poisťovňa stratu vo výške 101 miliónov eur. Situáciu nezľahčuje ani fakt, že mala podpriemernú úspešnosť výberu poistného - 96,97%, čiže strata by bola ešte vyššia.

V prípade dvoch súkromných zdravotných poisťovní boli priemerné náklady výrazne nižšie, teda majú menej rizikový poistný kmeň. Na druhej strane však mali aj výrazne vyšší priemerný predpis ako priemerné náklady. Dôvera zdravotná poisťovňa by mala bez kompenzácie rizikovej štruktúry zisk 133,51 eur na jedného poistenca, celkovo 190 miliónov eur. Pre zdravotnú poisťovňu Union by bol tento zisk 136,84 eur na jedného poistenca, čo predstavuje celkový zisk vo výške 64 milióna eur. Tieto dve súkromné poisťovne mali aj vyšší úspech výberu poistného: 98,1% a 98,07%.



Obr. 2: Ročný predpis a ročné náklady na jedného poistenca podľa zdravotných poisťovní [16]

Vidíme teda, že zdravotné poisťovne na Slovensku majú rozdielnú štruktúru kmeňa, čiže je prítomná selekcia rizika. Aby sa neznížila kvalita a efektivita zdravotnej starostlivosti, bolo potrebné aj na Slovensku zaviesť kompenzáciu rizikovej štruktúry poistencov, čiže kompenzáciu poisťovní za vysoko rizikových poistencov.

Prvý model kompenzácie rizika bol zavedený v roku 1995 a jediné rizikové faktory ktoré vstupovali do modelu boli demografické premenné - vek a pohlavie jednotlivcov [4]. Tento model bol založený na predpoklade, že výška nákladov na zdravotnú starostlivosť rastie vekom, alebo je ovplyvnená pohlavím jednotlivca. Veľkou výhodou takéhoto modelu je to, že spĺňa všetky podmienky pre rizikové faktory z Podkapitoly 1.4. Je veľmi jednoduchý, demografické údaje sú ľahko dostupné pre poisťovne a neznižuje ich motiváciu efektívne hospodáriť. Na druhej strane však má nízku prediktívnu silu - vysvetľuje len 2% variability nákladov [12].

Keď model kompenzácie rizikovej štruktúry vysvetľuje len takýto nízky podiel celkovej variability nákladov, stále necháva veľký priestor zdravotným poisťovniam pre selekciu rizika. Preto bolo nutné zavedenie ďalších rizikových faktorov do modelu kompenzácie rizikovej štruktúry. Od roku 2010 sa tak pridala faktor ekonomická aktivita/neaktivita do modelu [4]. Takýto model má ale ešte stále nízku prediktívnu silu, preto bolo veľmi dôležitým krokom, že súčasťou modelu kompenzácie rizikovej štruktúry sa stali aj rizikové faktory súvisiace so zdravotným stavom.

V roku 2012 sa zaviedli farmaceuticko-nákladové skupiny na základe Holandského modelu. Model farmaceuticko-nákladových skupín je založený na preskripcii liekov, pričom jeho predpokladom je, že existuje silný vzťah medzi zdravotným rizikom a užívaním liekov. Prediktívna sila takéhoto modelu je už výrazne vyššia, môže dosiahnuť až 20-25%. Ako presne vyzerá model farmaceuticko-nákladových skupín používaný na kompenzáciu rizikovej štruktúry, vysvetlíme v ďalšej podkapitole.

3.2 Model farmaceuticko-nákladových skupín

Model farmaceuticko-nákladových skupín a celý proces prerozdelenia sú zadefinované v Zákone č. 580/2004 o zdravotnom poistení o zmene a doplnení niektorých zákonov v znení neskorších predpisov [20].

Cieľom modelu je určiť očakávané náklady zdravotných poisťovní za určité fixné obdobie. Vstupom je informácia na úrovni poistenca, a tiež očakávané náklady sa určia pre každého poistenca.

3.2.1 Farmaceuticko-nákladové skupiny

Aby sme pochopili model farmaceuticko-nákladových skupín, potrebujeme najskôr ujasniť, čo sú farmaceuticko-nákladové skupiny. Každý kategorizovaný liek, čiže liek ktorý je plne alebo čiastočne hrađený zo zdravotného poistenia, sa zaradí do práve jednej anatomicko-terapeuticko-chemickej skupiny (ďalej len ATC skupina). To znamená, že lieky sú klasifikované podľa toho, na aký orgán je zameraný ich terapeutický účinok. Farmaceuticko-nákladová skupina je potom definovaná jedným alebo viacerými ATC skupinami. Ministerstvo zdravotníctva určí formou vyhlášky každý rok aktuálny zoznam farmaceuticko-nákladových skupín. Na rok 2018 ich zadefinovala 26, ktoré nájdeme v Prílohe C.

Tieto farmaceuticko-nákladové skupiny sú spôsobilé odhadnúť budúce náklady na poskytnutú zdravotnú starostlivosť poistenca s chorobou v chronickom štádiu na základe spotreby kategorizovaných liekov [20].

Sú však určité podmienky, ktoré musia farmaceuticko-nákladové skupiny spĺňať, aby mohli vstupovať do modelu. Po odhadnutí parametrov nasleduje verifikácia, a ak niektorá farmaceuticko-nákladová skupina nespĺňa všetky nižšie popísané podmienky, nebude zaradená do aktuálneho zoznamu farmaceuticko-nákladových skupín na daný rok [20]. Tieto podmienky sú nasledujúce:

- Jej zaradenie zvýši predpovedaciu silu modelu (koeficient determinácie) aspoň o 0,01.
- Koeficient farmaceuticko-nákladovej skupiny je štatisticky významný aspoň na hladine významnosti 0,01%.
- Podiel dodatočných nákladov farmaceuticko-nákladovej skupiny na celkových ročných nákladoch je aspoň 0,01%.
- Dodatočné náklady farmaceuticko-nákladovej skupiny tvoria aspoň 15% priemerých nákladov na poistenca. Dodatočné náklady sú určené ako rozdiel priemerých nákladov na poistenca zoradeného v danej farmaceuticko-nákladovej skupine a priemerných nákladov na poistenca.
- Viac, než polovica štandardných dávok liečiva boli vydané na chorobu priradenú

danej farmaceuticko-nákladovej skupiny (po vyradení ATC skupín, kde tento podiel je menší ako 30%).

3.2.2 Model farmaceuticko-nákladových skupín

Ako sme už vyššie spomenuli, vstupom do modelu farmaceuticko-nákladových skupín je informácia na úrovni poistenca. Každý jeden poistenec sa zaradi do práve jednej demografickej skupiny, a do jednej, alebo žiadnej farmaceuticko-nákladovej skupiny.

Dokopy na rok 2018 je 62 demografických skupín, do ktorých sa zaradia poistenci na základe pohlavia, 17 vekových skupín a platiteľa poistného. Takouto demografickou skupinou je napríklad ženy vo veku 40 až 44 rokov, za ktorých platí poistné štát. Zoznam všetkých demografických skupín nájdeme vo Vyhláske Ministerstva Zdravotníctva [18].

Ako sme už na začiatku tejto kapitoly spomenuli, na rok 2018 je zadefinovaných 26 farmaceuticko-nákladových skupín [19]. Pre každého poistenca sa určí počet dní, keď mu bol podávaný liek definujúci farmaceuticko-nákladovú skupinu (súčin počtu spotrebovaných balení a štandardnej dávky liečiva daného lieku). Poistenca zaradíme do danej farmaceuticko-nákladovej skupiny, ak počet dní presiahol alebo sa rovnal hodnote 181 dní. Ak možno poistenca zaradiť do viacerých farmaceuticko-nákladových skupín, vždy sa zaradi do farmaceuticko-nákladovej skupiny s najvyššími očakávanými nákladmi.

V modeli farmaceuticko-nákladových skupín vysvetľovanou premennou sú celkové ročné náklady na zdravotnú starostlivosť daného poistenca a vysvetľujúcimi premennými sú binárne premenné vyjadrujúce, či daný poistenec patrí do danej demografickej, či farmaceuticko-nákladovej skupiny. Nech n je počet poistencov. Potom

$$\mathbf{Y} = \bar{\mathbf{y}} + \mathbf{X}^d \boldsymbol{\beta} + \mathbf{X}^{PCG} \boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad (27)$$

kde $\mathbf{Y}$ je n -rozmerný vektor celkových ročných nákladov, $\bar{\mathbf{y}}$ je konštanta - populačný priemer celkových ročných nákladov, $\mathbf{X}_{n \times 62}^d$ je matica vyjadrujúca či daný poistenec patrí danej demografickej skupiny, a $\mathbf{X}_{n \times 26}^{PCG}$ je matica vyjadrujúca, či daný poistenec patrí do danej farmaceuticko-nákladovej skupiny.

Odhady pre $\boldsymbol{\beta}$ a $\boldsymbol{\gamma}$ sa určia metódou najmenších štvorcov. Potom odhadnuté parametre môžeme interpretovať takto:

- Odhady pre koeficientov $\beta_1, \beta_2, \dots, \beta_{62}$ vyjadrujú odchýlku od priemeru priemerných nákladov poistencov z jednotlivých demografických skupín podľa vekových skupín, pohlavia a typu platiteľa.
- Koeficienty $\gamma_1, \gamma_2, \dots, \gamma_{26}$ potom vyjadrujú odchýlky priemerných nákladov poistencov rovnakej vekovej skupiny, pohlavia a typu platiteľa zaradených do konkrétnej farmaceuticko-nákladovej skupiny od populačného priemeru.

Pripomeňme si, že cieľom tohto modelu je určiť, aké sú očakávané náklady pre poistencov v daných demografických skupinách, a aké sú dodatočné náklady na poistencov s chorobou v chronickom štádiu. Odhadnuté parametre zákon nazýva indexmi rizika. Pre každého poistenca sa určí jeho osobitné index rizika ako súčet jeho demografického a farmaceuticko-nákladového indexu. Ak jeho index rizika sa rovná 1, jeho očakávané náklady sa rovnajú populačnému priemeru.

3.2.3 Prerozdeľovanie poistného

Indexy rizika za rok t určí Ministerstvo zdravotníctva Slovenskej republiky formou vyhlášky na základe nákladov na zdravotnú starostlivosť jednotlivcov v roku $t - 2$ a ich spotrebu liekov v roku $t - 3$. Pri zadefinovaní indexov rizika za rok 2018 sa zohľadnili náklady v roku 2016 a spotreba liekov v roku 2015. Odhady sa robia na základe dát od zdravotných poisťovní - čiže na základe celej populácie.

Prerozdeľovanie poistného sa vykonáva mesačne, pričom jeho základom je 95% z celkovej sumy zaplatených preddavkov na poistné. Každý poistenec sa zaradí do jednej demografickej skupiny a do jednej alebo žiadnej farmaceuticko-nákladovej skupiny na základe spotrebovaných liekov za predchádzajúcich 12 mesiacov. Jeho index rizika sa potom určí ako súčet odhadnutého indexu rizika jeho demografickej skupiny a odhadnutého indexu jeho farmaceuticko-nákladovej skupiny z Vyhlášky Ministerstva vnútra [19]. Inak povedané, jeho očakávané náklady za rok 2018 sa určia na základe odhadnutých indexov rizika, kde boli zohľadnené náklady za rok 2016 a spotreba liekov za rok 2015.

Index rizika celej poisťovne sa potom určí ako súčet indexov rizika jej poistencov. Dostupné zdroje sa potom rozdelia medzi poisťovňami v pomere ich indexov rizika.

3.3 Odhad indexov rizika

Model (27) sme aplikovali na dáta zdravotnej poisťovne s cieľom odhadnúť indexy rizika nákladov na zdravotnú starostlivosť za rok 2018. Pri výpočte sme sa riadili Zákonom [20].

Ako vysvetľované premenné vstúpili do modelu náklady na zdravotnú starostlivosť za rok na úrovni poistencov. Každého poistenca sme zaradili do práve jednej demografickej skupiny na základe jeho pohlavia, veku ku dátumu 31.12.2016 a platiteľa poistného. Na základe spotreby liekov sme ich zaradili do jednej alebo žiadnej farmaceuticko-nákladovej skupiny.

Jednotlivé parametre sme odhadli metódou najmenších štvorcov. Odhady pre jednotlivé demografické, resp. farmaceuticko-nákladové skupiny sme zhrnuli v Tabuľke 1 a v Tabuľke 2.

Odhadnuté parametre sme následne testovali, či spĺňajú všetky zákonom definované podmienky z Podkapitoly 3.2.1.

Tieto odhadnuté parametre môžeme interpretovať nasledovne. Vidíme, že demografický index rizika žien, za ktorých platiteľom nie je štát vo veku 40-44 rokov je 0,50367. To znamená, že očakávané náklady poistencov v tejto demografickej skupine predstavujú zhruba polovicu populačného priemeru. Ak poistenec z tejto demografickej skupiny je zaradený aj do farmaceuticko-nákladovej skupiny 'DM1', čiže užíva lieky na ochorenie diabetes typu I, tak jeho index rizika bude súčet demografického indexu a indexu farmaceuticko-nákladovej skupiny: 3,96. Jeho očakávané náklady sú teda skoro štvornásobok populačného priemeru.

Predikčná sila modelu - koeficient determinácie R^2 - je 23%, čiže model farmaceuticko-nákladových skupín je schopný vysvetliť 23% variability nákladov na zdravotnú starostlivosť.

3.3.1 Vlastnosti modelu farmaceuticko-nákladových skupín

Veľkou výhodou modelu farmaceuticko-nákladových skupín je jeho jednoduchosť a jasná interpretovateľnosť. V podkapitole 2.1 sme ukázali, že odhad metódou najmenších štvorcov je nevychýlený, no je najlepší lineárny nevychýlený odhad len v tom prípade, ak platia Gauss-Markovove podmienky. Teraz sa pozrieme bližšie na náš model, či tieto

Veková kategória	Poistenci, za ktorých platiteľom nie je štát		Poistenci, za ktorých platiteľom je štát	
	Muži	Ženy	Muži	Ženy
00-04			0,89904	0,81828
05-09			0,48785	0,42578
10-14			0,49152	0,44390
15-19	0,30797	0,54877	0,43251	0,62786
20-24	0,29212	0,42530	0,37838	0,64507
25-29	0,27736	0,51524	0,45655	0,89099
30-34	0,30139	0,58435	0,59230	1,01684
35-39	0,32634	0,53061	0,69753	0,90343
40-44	0,35187	0,50367	0,79521	0,91142
45-49	0,40514	0,54144	1,1129	1,14697
50-54	0,47824	0,59674	1,53511	1,39322
55-59	0,58116	0,65031	1,90040	1,50015
60-64	0,68899	0,69325	1,48017	1,14237
65-69	1,06383	0,85467	1,61581	1,26998
70-74	1,58997	0,85670	2,04512	1,4884
75-79	1,91371	2,07372	2,11558	1,67566
80-xx	0,87291	1,37059	1,95847	1,53597

Tabuľka 1: Odhadnuté indexy rizika nákladov na zdravotnú starostlivosť na rok 2018 podľa veku, pohlavia a platiteľa poistného

Kód	Názov farmaceuticko-nákladovej skupiny	Index rizika
PAH	Primárna pľúcna hypertenzia	91,19049
REN	Renálne zlyhanie	37,56222
CFP	Cystická fibróza alebo ochorenie exokrinnej funkcie pankreasu	18,22985
HIV	HIV/AIDS	17,68894
TNF	Reumatické choroby liečené inhibítormi TNF	15,42571
ONK	Malignity	17,2980
CNS	Ochorenie mozgu a miechy	10,9312
RAS	Liečba rastovým hormónom	9,75724
TRA	Transplantácie	7,81131
DM1	Diabetes typu I	3,42398
HOR	Hormonálna onkoliečba	3,18747
NPP	Neuropatická bolesť	2,82857
COP	Chronická obštrukčná choroba a ťažká astma	2,55652
PAR	Parkinsonova choroba	2,4497
CRO	Crohnova choroba, ulcerózna kolitída	1,86097
KVS	Srdcové choroby	1,90323
REU	Reumatické choroby liečené inak ako inhibítormi TNF	1,60909
PSY	Antipsychotiká, Alzheimerova choroba, liečba závislostí	1,57925
EPI	Epilepsia	1,45706
AST	Astma	1,06889
DMH	Diabetes s hypertenziou	1,03143
DM2	Diabetes typu II	0,78647

DEP	Liečba antidepresívami	0,72651
THY	Poruchy štítnej žľazy	0,52002
GLA	Glaukóm	0,45950
CHO	Hypercholesterolémia	0,42680

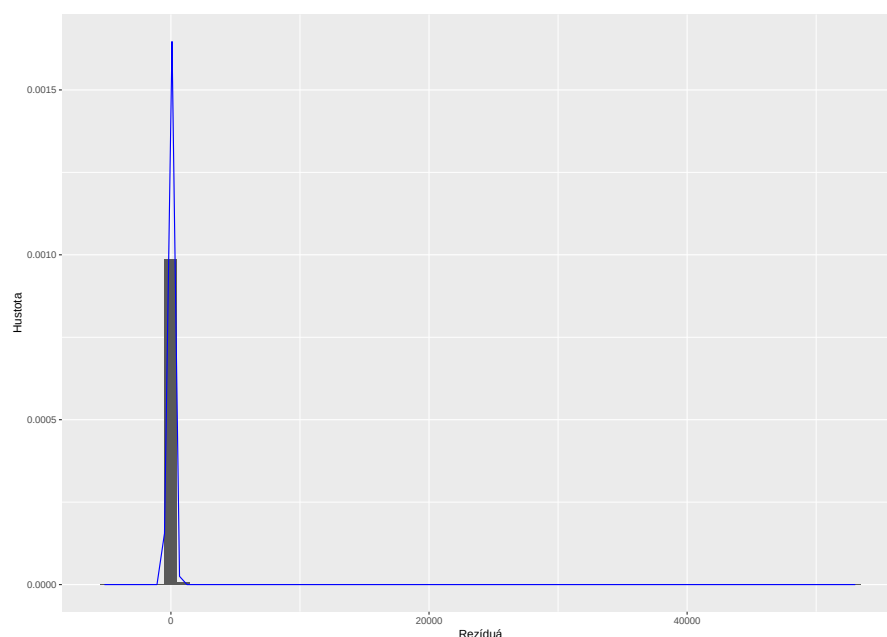
Tabuľka 2: Odhadnuté indexy rizika nákladov na zdravotnú starostlivosť na rok 2018 pre jednotlivé farmaceuticko-nákladové skupiny

podmienky nie sú porušené.

Normalita

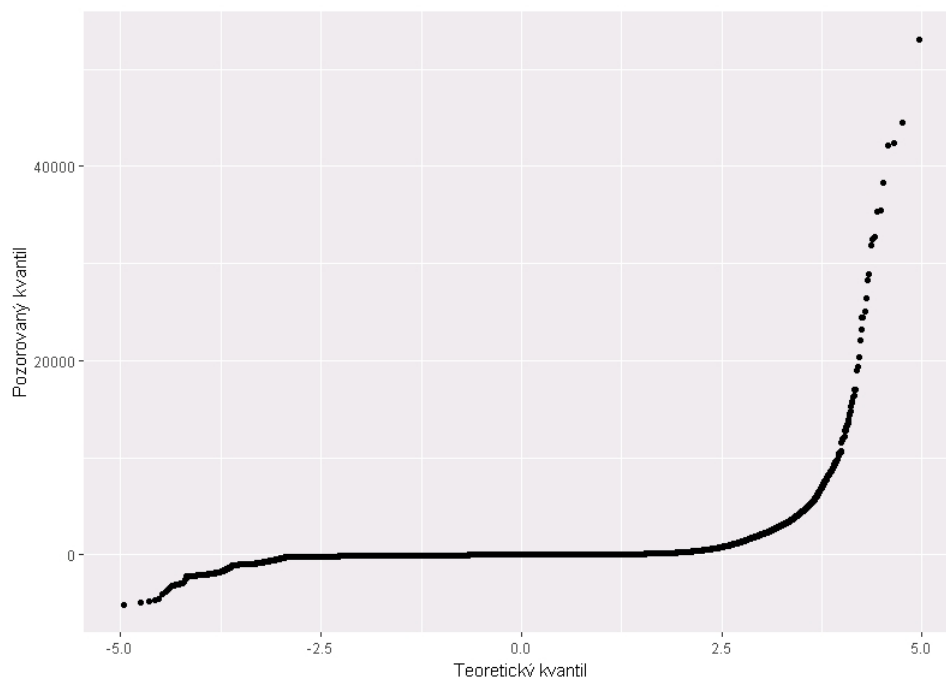
Normalita rezíduí je dôležitá pri testovaní lineárnych hypotéz. Ak by sme zistili, že rezíduá neriadia normálnym rozdelením, odhady metódou najmenších štvorcov $\hat{\beta}$ a $\hat{\gamma}$ môžu byť ešte stále najlepšie lineárne nevychýlené odhady, no ostatné inferencie o modeli (ako napríklad signifikantnosť odhadnutých parametrov, alebo konfidenčné intervaly) nemôžeme považovať za správne.

Pozrime sa najskôr na rozdelenie rezíduí. Na Obrázku 3 vidíme histogram pravdepodobností rezíduí, nad ktorým sme vykreslili hustotu normálneho rozdelenia, ktorá má strednú hodnotu rovnú priemeru rezíduí a tiež varianciu rezíduí.



Obr. 3: Histogram pravdepodobností rezíduí s krivkou hustoty normálneho rozdelenia

Na potvrdenie či vyvrátenie normality najskôr využijeme grafickú metódu - vykreslíme kvantil-kvantilový graf rezíduí. Na x-ovej osi sa nachádzajú kvantily normálneho rozdelenia, na y-ovej osi sú kvantily rezíduí. Ak by sa rezíduá riadili normálnym rozdelením, závislosť by bola medzi kvantilmi normálneho rozdelenia a kvantilmi rezíduí lineárna. Na Obrázku 4 však vidíme, že táto závislosť lineárna rozhodne nie je. Ne-normalitu rezíduí potvrdil aj Kolgomorov-Smirnov test. Keďže sme dostali menšiu p-hodnotu ako 2×10^{-16} , zamietame nulovú hypotézu o normalite náhodných odchýlok.



Obr. 4: Kvantil-kvantilový graf rezíduí

Homoskedasticita a nekorelované náhodné odchýlky

Heteroskedasticita aj korelované náhodné odchýlky sú porušením druhej Gauss-Markovovej podmienky $E(\epsilon\epsilon^T) = \sigma^2 I$. Keď táto podmienka nie je splnená, nemôžeme si byť istí, že odhady metódou najmenších štvorcov sú najlepšie lineárne nevychýlené odhady, čiže odhady s najmenšou možnou varianciou.

Heteroskedasticitu sme testovali pomocou Breusch-Paganovho testu. Hypotézu o homoskedasticite sme zamietli, keďže sme dostali p-hodnotu menšiu ako 2×10^{-16} .

Na otestovanie prítomnosti korelovaných náhodných chýb sme použili Breusch-Godfreyho test. S p-hodnotou **0.013** sme nulovú hypotézu o nekorelovaných náhodných odchýl-

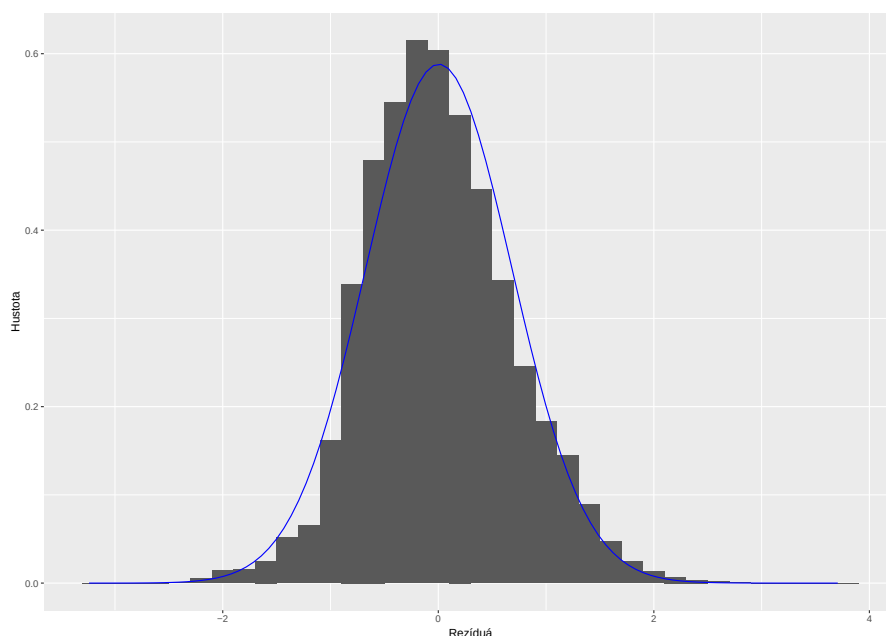
kach zamietli.

3.3.2 Tranformácia modelu farmaceuticko-nákladových skupín

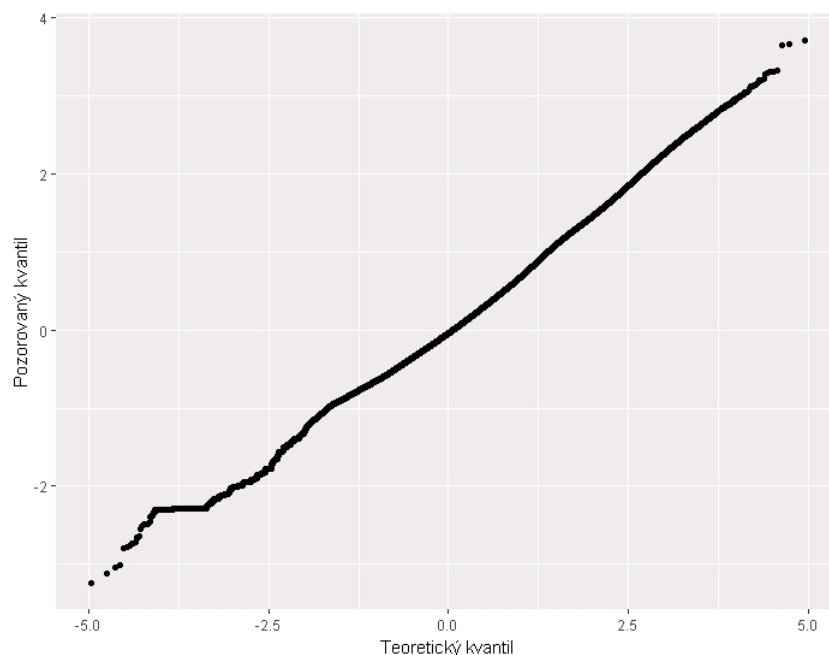
Keďže náš model nespĺňa Gauss-Markovove podmienky, rozhodli sme aplikovať Box-Coxovu transformáciu na vysvetľovanú premennú. Metódou maximálnej vierohodnosti sme našli vhodný parameter $\lambda = -0,1414$. Po pretransformovaní vysvetľovanej premennej sme fitovali model (27). Predikčná sila modelu sa výrazne zvýšila, $R^2 = 92\%$ v tomto prípade, čiže model vysvetľuje až 92% variability pretransformovanej vysvetľovanej premennej.

Normalita

Teraz otestujeme, či takto pretransformovaný model spĺňa Gauss-Markovove podmienky. Začíname normalitou, ktorá je potrebná len pre testovanie lineárnych hypotéz. Na Obrázku 5 sme znázornili histogram pravdepodobností rezíduí. V tomto prípade sa histogram podobá viac na hustotu normálneho rozdelenia, čo potvrdí aj kvantil-kvantilový graf na Obrázku 6. Hypotézu o normalite náhodných odchýlok však Kolmogorov-Smirnov test odmietol na hladine významnosti 5%.



Obr. 5: Histogram pravdepodobností rezíduí po Box-Coxovej transformácie s krivkou normálneho rozdelenia



Obr. 6: Kvantil-kvantilový graf rezíduí po Box-Coxovej transformácie

Homoskedasticita a nekorelované náhodné odchýlky

Heteroskedasticitu sme aj v tomto prípade testovali pomocou Breusch-Paganovho testu. Hypotézu o homoskedasticite sme zamietli, s p-hodnotou menšou ako 2×10^{-16} . To, či náhodné odchýlky sú korelované sme testovali Breusch-Godfreyho testom. Hypotézu o nekorelovaných náhodných chybách sme žiaľ aj v tomto prípade museli odmietnuť.

Box-Coxova transformácia síce zlepšila vlastnosti modelu farmaceuticko-nákladových skupín, no heteroskedasticita je ešte stále prítomná v modeli, preto si nemôžeme byť istí, že odhad metódou najmenších štvorcov je naozaj najlepší možný odhad.

Na druhej strane vďaka tejto transformácii sa výrazne zvýšila predikčná sila modelu. V tomto prípade $R^2 = 92\%$. Interpretácia odhadnutých koeficientov však už nie je triviálna, čo výrazne sťažuje vyhodnotenie rizikovosti kmeňa poisťovne. Tento pretransformovaný model preto ani neodporúčame používať na takéto účely.

3.3.3 Verifikácia modelu farmaceuticko-nákladových skupín

Kedže model farmaceuticko-nákladových skupín aj po Box-Coxovej transformácii porušuje Gauss-Markovove podmienky, testy o signifikantnosti parametrov a o signifikantnosti modelu nie sú správne. Potrebujeme preto model validovať iným spôsobom.

Na verifikáciu modelu sme použili k-násobnú krížovú validáciu. Náš dátový súbor sme najskôr rozdelili na trénovaciu sadu (80% dát) a na testovaciu sadu (20% dát). K-násobnú krížovú validáciu sme potom aplikovali na trénovaciu sadu.

Trénovaciu sadu sme rozdelili na $k = 50$ menších sád s približne rovnakým počtom pozorovaní. Vykonalí sme potom 50 odhadov takým spôsobom, že vždy sme vynechali jednu sadu, a odhady sme spravili na základe ostatných 49 sád. Na konci sme mali 50 odhadov pre jednotlivé parametre. Finálny odhad sme dostali spriemerovaním týchto 50 odhadov. Výhodou takéhoto prístupu je to, že finálny odhad je ovplyvnený každým jedným pozorovaním (čo napríklad pri Bootstrap agregácie nemusí platiť).

Po tom, ako sme získali odhad z trénovacej sady, pozreli sme sa na to, ako presne tento odhad predikuje náklady pre testovaciu sadu. Predikčná sila takéhoto odhadu bola len okolo 13% na testovacej sade aj po viacerých opakovaníach k-násobnej krížovej validácie, čo znamená, že model predikuje výrazne horšie pre nové dáta (pripomíname, že koeficient determinácie R^2 bol 23% pre model (27)).

4 Výber liekov do prerozdeľovacieho modelu

Modely založené na preskripcii liekov môžu mať viacero nevýhod. Niektoré ochorenia napríklad môžu byť liečené celou paletou rôznych liekov, a preto nemusí byť vždy jasné spojenie medzi zdravotným stavom jednotlivca a preskripciou lieku [12]. Táto vlastnosť liekov je v rozpore s podmienkou validity pre prediktory v modeli kompenzácie rizikovej štruktúry, ktorú sme zadefinovali v Kapitole 1. V rámci tejto kapitoly sa venujeme práve tomuto nedostatku modelu farmaceuticko-nákladových skupín. Naším cieľom je nájsť také lieky, na základe ktorých vieme jednoznačne identifikovať zdravotný stav jednotlivca.

Podmienka validity taktiež hovorí, že prediktor v modeli musí úzko súvisieť s nejakým chronickým stavom. Tento predpoklad je dôležitý preto, lebo poisťovne by mali byť kompenzované len za náklady, ktoré sú predvídateľné. Preto v rámci našej práce hľadáme také lieky, ktorých užívanie jednoznačne implikuje chronické ochorenie jednotlivca.

Našu analýzu môžeme rozdeliť na dve časti. V prvej časti hľadáme také lieky, ktoré vieme jednoznačne spojiť s nejakým zdravotným stavom. V Kapitole 3 sme už vysvetlili, že každý liek je zaradený do anatomickeo-terapeuticko-chemickej skupiny na základe jeho terapeutického účinku. Preto skúmame závislosti medzi užívaním lieku z nejakej ATC skupiny a zdravotným stavom jednotlivca.

V druhej časti sme zisťovali, či daný zdravotný stav je chronický. Chronické ochorenia sú ochorenia, ktoré trvajú dlhý čas, a tým pádom ich liečenie generuje dlhodobo vysoké náklady. Hľadali sme preto také ATC skupiny, ktoré sú užívané jednotlivcami dlhšiu dobu a ich užívanie významne zvyšuje nákladovosť daného jednotlivca. Sústredili sme sa taktiež na také ATC skupiny, ktoré sú užívané vysokým počtom jednotlivcov.

4.1 Dáta

K dispozícii sme mali dáta zdravotnej poisťovne - vykázanú zdravotnú starostlivosť. Vykázaná zdravotná starostlivosť obsahuje všetok zdravotnú starostlivosť, ktorá bola poskytnutá poistencom zdravotnej poisťovne. Z týchto dát sme vybrali spotrebu liekov. Každá preskripcia lieku predstavuje jeden riadok vo vykázanej zdravotnej starostlivosti,

pričom obsahuje

- kód lieku,
- ATC skupinu, do ktorej je daný liek zaradený,
- kód diagnózy na základe Medzinárodnej klasifikácie chorôb (10. revízia) [7].

Kód diagnózy je trojmiestna skratka - skladá sa z jedného písmena a dvoch čísel a určuje špecifickú chorobu. Základné skupiny diagnóz sme uviedli v Príloha A.

4.2 Postup analýzy

Pre každú preskripciu sme teda mali k dispozícii presnú diagnózu, na ktorú bol daný liek predpísaný a ATC skupinu, ktorá bola predpísaná. Hľadali sme také ATC skupiny, ktoré sme vedeli jednoznačne spojiť s nejakou diagnózou.

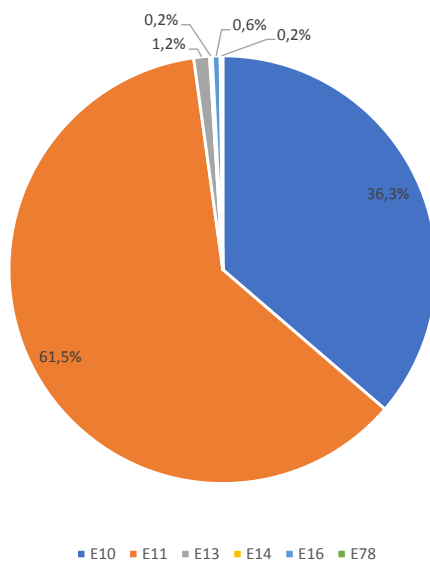
Hľadať súvislosť medzi liekmi a jednotlivými diagnózami nie je však príliš efektívne. Niektoré chronické stavy sú totiž pokryté viacerými diagnózami. Takéto chronické ochorenie je napríklad aj diabetes mellitus, ktoré môže byť vykázané pod rôznymi kódmi diagnóz (Tabuľka 3), pričom rôzne typy ochorenia môžu byť liečené rovnakými liekmi. Preto sme sa rozhodli hľadať súvislosti medzi jednotlivými ATC skupinami a skupinami diagnóz.

Kód	Názov ochorenia
E10	Diabetes mellitus závislý do inzulínu
E11	Diabetes mellitus nezávislý do inzulínu
E12	Diabetes mellitus súvisiaci s podvýživou
E13	Iný bližšie určený diabetes mellitus
E14	Nešpecifikovaný diabetes mellitus

Tabuľka 3: Rôzne typy ochorenia diabetes mellitus podľa [7]

Najskôr sme však potrebovali vytvoriť skupiny diagnóz. Poistencom zdravotnej poisťovne bolo predpísaných 814 ATC skupín na 1222 diagnóz za skúmané obdobie. Pre každú ATC skupinu sme určili podiel preskripcií na jednotlivé diagnózy. Pre lepšie

pochopenie podiel diagnóz pre ATC skupinu A10AB01 (ľudský inzulín) sme znázornili na Obrázku 7. Vidíme, že túto ATC skupinu by sme vedeli spojiť s diagnózami E10 a E11, a v takomto prípade by bolo napríklad efektívne spojiť tieto dve diagnózy. Potom by sme mohli povedať, že na tieto diagnózy bolo predpísaných 94,8% liekov z ATC skupiny A10AB01.



Obr. 7: Podiel jednotlivých diagnóz pri preskripcii ATC skupiny A10AB01

Spojiť diagnózy je teda užitočné na základe toho, či na sa ich liečenie používa tá istá ATC skupina. Každéj diagnóze sme preto priradili jeden 814-rozmerný vektor, kde každá zložka prislúchala podielu preskripcii jednej ATC skupiny na danú diagnózu. Diagnózy sme potom zhlučkovali metódou DBSCAN. Táto metóda je špecifická tým, že určuje aj tzv. body šumu, čiže také body, v ktorých blízkosti žiaden iný bod nie je. V našom prípade ale táto metóda môže mať nevýhody. Ak napríklad 100% určitej ATC skupiny bola predpísaná na danú diagnózu, tak bod tej diagnózy bude v našom 814-rozmernom priestore výrazne posunutý v smere tej danej ATC skupiny, čo môže znamenať, že žiadne iné body nebudú v jeho blízkosti, a táto diagnóza bude označená ako bod šumu. Toto my ale rozhodne nechceme, keď totiž 100% určitej ATC skupiny

bola predpísaná na danú diagnózu, tak tá ATC skupina jednoznačne určuje zdravotný stav pacienta.

Z tohto dôvodu sme z nášho súboru najskôr vybrali tie ATC skupiny, ktorých viac ako 80% bolo predpísaných na jednu diagnózu. Takýchto ATC skupín bolo 251, takže zhlukovanie sme nakoniec vykonali v 563-rozmernom priestore. Dostali sme nakoniec 34 zhlukov diagnóz na základe toho, či na tieto diagnózy boli predpísané podobné lieky. Tieto zhluky nájdeme v Prílohe D.

Teraz sme už len potrebovali zistiť, ktoré ATC skupiny implikujú jednoznačný zdravotný problém. Vypočítali sme podiel jednotlivých ATC skupín na jednotlivé diagnózy, príp. zhluky diagnóz. Hranicu sme určili na 50%, čiže ATC skupinu sme spojili s danou diagnózou, resp. zhlukom diagnóz, ak aspoň polovica danej ATC skupiny bola predpísaná na danú diagnózu, resp. skupinu diagnóz. Takto sme dostali 560 takých ATC skupín, na základe ktorých máme celkom presné informácie o zdravotnom stave poistencov.

Týmto sme sa dostali k druhej časti našej analýzy. Z týchto 560 ATC skupín ešte potrebujeme vybrať tie, ktoré indikujú chronický stav.

Prvou podmienkou, aby sme mohli tvrdiť, že užívanie danej ATC skupiny implikuje chronické ochorenie je, aby poistenci užívajúce lieky z ATC skupiny mali vyššie priemerné náklady ako populačný priemer. Tie ATC skupiny, ktoré túto podmienku nesplnili, sme vyradili z nášho zoznamu.

Dôležitým faktorom bol aj počet poistencov užívajúcich lieky z danej ATC skupiny. Do modelu kompenzácie rizikovej štruktúry by totiž mali vstupovať hlavne také ochorenia, ktoré sa týkajú väčšiu skupinu poistencov. Pre zdravotné poisťovne predstavuje síce veľké riziko, ak pár poistencov má extrémne vysoké náklady, no oveľa nebezpečnejšie sú mierne vyššie náklady veľkej skupiny poistencov. Preto sme skúmali aj počet poistencov užívajúcich lieky v jednotlivých ATC skupinách. Hranicu počtu poistencov sme určili ako 25%-ný kvantil počtov poistencov užívajúcich lieky v jednotlivých ATC skupinách. Tie ATC skupiny, ktoré boli užívané menším počtom poistencov, ako táto hranica, sme z finálneho zoznamu vyradili.

Keďže chronické ochorenia sú dlhodobé ochorenia, skúmali sme aj dĺžku užívania liekov v jednotlivých ATC skupinách. Zohľadnili sme dva faktory: priemerný počet

spotrebovaných balení a priemerný počet spotrebovaných štandardných dávok. V oboch prípadoch sme vyradili tie ATC skupiny, ktoré neprekročili hranicu 25%-ného kvantilu.

4.3 Výsledok analýzy

Výsledok analýzy, čiže zoznam ATC skupín, ktoré vyhoveli vyššie popísaným podmienkam nájdeme v Prílohe E. Nami identifikované ATC skupiny môžeme teda spojiť s nejakou diagnózou, či skupinou diagnóz, teda sú dobrými kandidátmi pre vstup do modelu farmaceuticko-nákladových skupín.

Aby podmienka validity nebola porušená, je potrebné aby diagnózy, s ktorým sme spájali jednotlivé ATC skupiny naznačovali chronické ochorenie. Túto ich vlastnosť sme posudzovali na základe rôznych faktorov. Plnenie nami nastavených podmienok však ešte nezaručuje, že ide o chronické ochorenie, preto odporúčame preverenie zoznamu odborníkom.

5 Kompenzácia vybranej skupiny poistencov

Napriek tomu, že vo svete sa objavajú čoraz sofistikovanejšie modely kompenzácie rizikovej štruktúry, ešte aj tieto modely podstatne podkompenzujú vybrané skupiny poistencov. Pod pojmom podkompenzácia rozumieme jav, že modelom predikované náklady sú systematicky nižšie ako skutočné náklady týchto vybraných skupín.

Keď zdravotné poisťovne sú podkompenzované za určité skupiny poistencov, tieto skupiny pre nich nie sú atraktívne, a dôjde k selekcii rizika. Dôsledky selekcie rizika sme podrobne popísali v Kapitole 1, no pripomínáme, že znižuje kvalitu a efektivitu zdravotnej starostlivosti, preto nie žiadajúca na trhu verejného poistenia.

Najpodstatnejším dôvodom podkompenzácie je to, že rizikové faktory vstupujúce do modelov kompenzácie rizikovej štruktúry sú do určitej miery neúplné. Nie každá informácia s prediktívnou hodnotou pre náklady na zdravotnú starostlivosť spĺňa požiadavky z Kapitoly 1, a tým pádom nemôže vstúpiť do modelu. Takýto rizikový faktor je napríklad aj výška nákladov na zdravotnú starostlivosť jednotlivca v nejakom predchádzajúcom období. Tento rizikový faktor má jednoznačne prediktívnu hodnotu, no zvyšuje motiváciu poisťovní pre neefektívnu liečbu.

V modeli farmaceuticko-nákladových skupín sú jednotlivci klasifikovaní do farmaceuticko-nákladovej skupiny len ak za posledný rok spotrebovali aspoň 181 štandardných dávok relevantného lieku. Táto hranica má svoje opodstatnenie. Ak by napríklad pre zaradenie do farmaceuticko-nákladovej skupiny stačila jedna štandardná dávka relevantného liečiva, zdravotné poisťovne by boli motivované podporovať zbytočný predpis liekov, aby dostávali vyššiu kompenzáciu v budúcnosti.

Na jednej strane teda takáto vysoká hranica eliminuje motiváciu poisťovní podporovať zbytočný predpis liekov. Na druhej strane však spôsobuje nedostatočnú kompenzáciu poistencov, ktorí sú síce chronicky chorí, no lieky užívajú len v menšom množstve.

5.1 Kompenzácia poistencov s menšou spotrebou liekov

Naším predpokladom je, že v modeli farmaceuticko-nákladových skupín poistenci klasifikovaní v nejakej farmaceuticko-nákladovej skupine (skupina X) sú dostatočne kompenzovaní. Na druhej strane však tá skupina poistencov, ktorá relevantné lieky tiež

užívala, no menej než 181 štandardných dávok (skupina Y), bude pravdepodobne podkompenzovaná. Ako výsledok, celá skupina užívajúca relevantný liek (skupina XY) je podkompenzovaná. Keďže parametre modelu sa odhadujú metódou najmenších štvorcov, platí že súčet štvorcov rezíduí musí rovnať nule. Z tohto dôvodu skupina neužívajúca žiadne lieky (skupina Z) bude nadkompenzovaná. Autori článku [10] potom navrhujú, že podkompenzácia skupiny XY a nadkompenzácia skupiny Z sa dá odstrániť s dodatočnou kompenzáciou skupiny X.

Vychádzajúc z článku [10] autori článku [11] navrhli analytický nástroj na určenie výšky dodatočnej kompenzácie - lineárnu regresiu s obmedzeniami na predikované hodnoty.

V rámci tejto kapitoly sa pozrieme na kompenzáciu tých poistencov, ktorí síce užívali relevantné lieky pre farmaceuticko-nákladové skupiny, no nespĺňali hranicu 181 štandardných dávok. Ak sa nám potvrdí hypotéza o ich podkompenzácie, aplikujeme návrh z článku [10] na náš model farmaceuticko-nákladových skupín a zanalyzujeme, či sa celková kompenzácia zlepšila.

5.2 Rozdiel medzi predikovanými a skutočnými nákladmi vybraných skupín

Poistencov sme zaradili do vyššie spomenutých skupín X, Y a Z. Poznamenáme, že aj v prípade zaradení poistencov do skupiny Y sme zachovali princíp, ktorý platí pri zaradení poistencov do farmaceuticko-nákladových skupín. Poistenca sme zaradili vždy do najviac nákladovej skupiny.

Priemerný rozdiel medzi predikovanými a skutočnými nákladmi poistencov v jednotlivých skupinách sme zhrnuli v Tabuľke 4. Vidíme, že priemerná kompenzácia jednotlivých skupín sa výrazne líši, čo nám potvrdil aj Studentov t-test. Vykonali sme 3 testy (testovali sme rovnosť priemerov po pároch), každý na hladine významnosti 5%. Hypotézu o rovnosti strednej hodnoty v každom prípade sme zamietli s p-hodnotou menšou ako 2×10^{-16} .

Skupina poistencov	Počet poistencov v skupine	Priemerný rozdiel
Skupina X	202 965	-117,1692
Skupina Y	139 452	-428,1456
Skupina Z	1 114 291	32,3316

Tabuľka 4: Rozdiel medzi predikovanými a skutočnými nákladmi jednotlivých skupín poistencov

Keďže sme sa nám potvrdila hypotéza, že skupina poistencov užívajúcich relevantné lieky pre jednotlivé farmaceuticko-nákladové skupiny je výrazne podkompenzovaná - ich predikované náklady sú približne o 428 eur menšie na jedného poistenca ročne, ako ich skutočné náklady - , chceme sa pozrieť aj na to, či výška podkompenzácie závisí od farmaceuticko-nákladovej skupiny. Túto informáciu sme zhrnuli v Tabuľke 5.

Najvyšší rozdiel medzi predikovanými a skutočnými nákladmi majú poistenci užívajúce lieky relevantné pre farmaceuticko-nákladovú skupinu renálne zlyhanie, 25 000 eur. Keď sa však zohľadňujeme súčet odchýlok pre celú skupinu poistencov, model predikuje najhoršie pre jednotlivcov, ktorí užívajú lieky pre srdcové ochorenia. Celkový rozdiel pre túto skupinu je vyše 8.5 milióna eur. Z Tabuľky 5 je nám zrejmé, že pre zdravotnú poisťovňu znamená síce vysoké riziko ak určitá malá skupina poistencov je výrazne podkompenzovaná (renálne zlyhanie, malignity), no ešte nebezpečnejšie je len mierna podkompenzácia veľkej skupiny poistencov, ako srdcové choroby, astma, alebo neuropatická bolesť. Preto v ďalšej časti sa sústredíme práve na tieto ochorenia.

5.3 Obmedzenie pre predikovanú hodnotu v lineárnom modeli

Na základe článku [11] zavedieme obmedzenia pre parametre lineárneho modelu s účelom zníženie podkompenzácie s fixným percentom pre skupinu Y.

Pre každého poistenca modelom predikovaná hodnota je $\hat{\mathbf{y}}_i = \sum_{j=1}^k \hat{\beta}_j \mathbf{x}_{ij}$. Majme skupinu $\mathbf{g}$, ktorá je podkompenzovaná a nech počet poistencov v skupine $\mathbf{g}$ je n_g . Potom priemerná predikovaná hodnota v skupine $\mathbf{g}$ je

$$\bar{\mathbf{y}}_g = \frac{1}{n_g} \sum_{i \in g} \sum_{j=1}^k \hat{\beta}_j \mathbf{x}_{ij}. \quad (28)$$

Skupina	Priemerný rozdiel v skupine	Celkový rozdiel
KVS	-921,29	-8 537 600,47
ONK	-10 134,02	-7 823 462,09
AST	-195,51	-7 609 374,77
NPP	-708,86	-5 534 773,77
PSY	-600,33	-4 486 900,50
COP	-388,22	-3 798 733,38
DEP	-290,69	-3 442 627,20
THY	-158,45	-3 092 469,05
REN	-25 378,21	-2 461 686,38
EPI	-466,73	-2 440 057,56
CNSn	-905,96	-1 616 226,61
HOR	-1 493,61	-1 317 367,56
REU	-656,22	-1 278 322,28
DM1	-1 472,47	-986 551,62
TRA	-2 096,51	-865 857,69
CHO	-64,73	-853 101,59
DMH	-250,95	-844 948,90
CRO	-646,85	-607 389,04
PAR	-757,84	-570 653,65
GLA	-141,19	-499 398,48
TNFn	-8 040,13	-458 287,42
DM2	-176,14	-200 275,90
RAS	-2 853,99	-125 575,66
HIV	-5 964,89	-101 403,05
PAH	-14 294,36	-100 060,54
CFP	-7 529,77	-52 708,41

Tabuľka 5: Rozdiel medzi predikovaným a skutočnými nákladmi pre skupinu Y

Rovnicu (28) môžeme prepísať ako

$$\bar{\hat{\mathbf{y}}}_g = \sum_{i \in g} \sum_{j=1}^k \hat{\beta}_j \bar{x}_{ij}, \quad (29)$$

kde $\bar{x}_{ij}$ sú priemerné vysvetľujúce premenné pre skupinu g . Po určení tohto priemeru pre skupinu g môžeme jednoducho nastaviť obmedzenie pre $\bar{\hat{\mathbf{y}}}_g$ tak, aby sme dostali požadovanú hodnotu. Požadovanú hodnotu $\bar{\hat{\mathbf{y}}}_g$ môžeme určiť tak, aby sme celkovú podkompenzáciu skupiny g znížili s fixným podielom podkompenzácie základného modelu, čo môžeme vyjadriť ako

$$\bar{\hat{\mathbf{y}}}_g = \bar{c}_g - \alpha(\bar{c}_g - \sum_{j=1}^k \hat{\beta}_j \bar{x}_{gj}), \quad (30)$$

kde $\bar{c}_g$ sú priemerné náklady poistencov v skupine g .

5.4 Aplikácia obmedzení na model farmaceuticko-nákladových skupín

Ako sme už vyššie písali, sústredili sme sa na tie skupiny poistencov, ktoré znamenajú vysoké riziko pre zdravotnú poisťovňu.

Vidíme však, že predikované náklady každej skupiny z Tabuľky 5 sú výrazne nižšie ako ich skutočné náklady, ale obmedzenie na každý parameter nie je príliš efektívne - výrazne sa môže totiž zhoršiť prediktívna sila modelu. Preto sme sa rozhodli obmedziť parametre piatich skupín s najväčším celkovým rozdielom medzi predikovanými a skutočnými nákladmi: KVS, ONK, AST, NPP, PSY. Týchto 5 skupín je zodpovedný totiž za 56% celkového rozdielu v rámci skupiny Y.

Obmedzenia na predikované náklady sme určili na základe (30). Metódou najmenších štvorcov sme potom odhadli $\hat{\beta}_{OBM}$ a určili výšku podkompenzácie pre relevantné skupiny. Výsledok sme zhrnuli v Tabuľke 6.

Čo z Tabuľky 6 hneď vidíme, že kompenzácia každej podskupiny poistencov sa zlepšila v dôsledku zavedení obmedzení na predikované hodnoty modelu farmaceuticko-nákladových skupín. Celková kompenzácia skupín XY sa výrazne zvýšila, čo môže znížiť motiváciu poisťovní na selekciu rizika.

Takýto prístup však má aj nedostatky. Keď obmedzíme predikované hodnoty, pre-

dikčná sila lineárneho modelu sa zníži. V našom prípade koeficient determinácia sa zhoršila extrémne - z 23% na 13%. Pripomíname, že pri validácii modelu farmaceuticko-nákladových skupín sme zistili, že na nových dátach bola jeho predikčná sila len 13%, s obmedzeniami by bola ešte o dosť nižšia, čo koniec koncom môže práve podporovať poisťovne pri selekcii rizika.

Skupina	Priemerný rozdiel v pôvodnom modeli	Priemerný rozdiel v novom modeli
KVS (Skupina X)	-225,72	257,16
KVS (Skupina XY)	-380,40	28,18
ONK (Skupina X)	-868,56	-210,24
ONK (Skupina XY)	-3 272,88	-2 747,28
AST (Skupina X)	-36,10	625,54
AST (Skupina XY)	-231,6	87,72
NPP (Skupina X)	-96,77	1 256,70
NPP (Skupina XY)	-601.20	-241,92
PSY (Skupina X)	-83,134	928,51
PSY (Skupina XY)	-347,88	215,28

Tabuľka 6: Rozdiel predikovaných a skutočných nákladov jednotlivých skupín poistencov v prípade pôvodného modelu farmaceuticko-nákladových skupín a modelu s obmedzeniami na parametre

Záver

Cielom našej práce bolo vybrať a popísať metódy data miningu a použiť ich na riešenie vybraného problému v zdravotníctve. Ako vybraný problém sme si zvolili kompenzáciu rizikovej štruktúry zdravotných poisťovní.

Prvú kapitolu sme venovali vysvetľovaniu, prečo je potrebná kompenzácia rizikovej štruktúry. V Kapitole 2 sme sa podrobne zaoberali s lineárnym modelom a vysvetlili sme, kedy môžeme považovať odhad metódou najmenších štvorcov za najlepší možný odhad a ujasnili sme aj základy zhlukovej analýzy. Ďalej sme objasnili, prečo bolo potrebné zaviesť kompenzáciu rizikovej štruktúry na Slovensku. Poukázali sme na rozdielnú rizikovosť kmeňa jednotlivých zdravotných poisťovní, čo naznačuje prítomnosť selekcie rizika na trhu verejného poistenia. Podrobne sme popísali aj model kompenzácie rizikovej štruktúry na Slovensku - model farmaceuticko-nákladových skupín (27).

Zákonom definovaný model farmaceuticko-nákladových skupín sme následne aplikovali na dáta zdravotnej poisťovne. Získali sme indexy rizika nákladov na zdravotnú starostlivosť pre jednotlivé demografické a farmaceuticko-nákladové skupiny. Zistili sme, že model je schopný vysvetliť 23% variability nákladov, čo je podľa literatúry [8], [12] v prípade nákladov na zdravotnú starostlivosť dobrý výsledok.

Pri testovaní predpokladov modelu sme zistili, že je prítomná heteroskedasticita aj korelované náhodné odchýlky, ktoré sa ani neriadia normálnym rozdelením. Preto sme si nemohli byť istí, že odhad metódou najmenších štvorcov je najlepší možný a ani sme nemohli testovať signifikantnosť odhadnutých parametrov. Model sme preto pretransformovali Box-Coxovou metódou. Nový model mal síce oveľa vyššiu predikčnú silu (až 92%) a aj lepšie vlastnosti (Obrázky 5 a 6), no interpretácia odhadnutých koeficientov ďalej už nebola jednoznačná. Na základe tohto modelu sa už nedá jednoznačne určiť index rizika nákladov na zdravotnú starostlivosť jednotlivca a zdravotnej poisťovne, preto ho ani neodporúčame ďalej používať.

Po aplikovaní zákonom daného modelu farmaceuticko-nákladových skupín na dáta zdravotnej poisťovne nás zaujímalo, akú má skutočnú predikčnú silu, t.j. ako kvalitná je jeho predikcia pre nové dáta. Na odhadnutie parametrov sme použili k-násobnú krížovú validáciu na trénovacej sade. Zo získanými parametrami sme spravili odhad pre testovaciu sadu, a zistili sme, že predikčná sila modelu na testovacej sade je 13%,

čo je oproti pôvodnému modelu pokles o 10%.

Toto zistenie považujeme za jeden z prínosov našej práce. Ako sme v Podkapitole 3.2.3 vysvetlili, v procese prerozdelenia sa indexy rizika odhadujú na dátach spred dvoch rokov, a potom tieto odhady sa používajú na predikciu súčasných nákladov. Preto každého jednotlivca s súčasnými charakteristikami, ktoré vstupujú do procesu prerozdelenia, môžeme považovať za nové pozorovanie, pre ktorý chceme nájsť odhad nákladov. Je dôležité preto povedať, že síce model farmaceuticko-nákladových skupín vysvetľuje 23% variabilitu nákladov, no v procese prerozdelenia predikované náklady jednotlivcov nie sú také presné - v tomto prípade predikčná sila modelu je len 13%.

Štvrtú kapitolu sme venovali liekom vstupujúcim do modelu farmaceuticko-nákladových skupín. Cieľom tejto časti práce bolo, aby sme našli také anatomicko-terapeuticko-chemické skupiny liečiv, na základe ktorých je možné jednoznačne identifikovať chronické ochorenia jednotlivcov. Zistili sme, že je efektívnejšie hľadať súvislosti medzi skupinami diagnóz a skupinami ATC skupín, ako medzi jednotlivými diagnózami a ATC skupinami. Preto sme metódou DBSCAN vytvorili skupiny diagnóz na základe toho, či sa na ich liečenie používajú podobné ATC skupiny. Po vytvorení zhlukov sme určili, aký podiel jednotlivých ATC skupín bolo predpísaných na jednotlivých diagnóz, resp. zhluky diagnóz. Keď tento podiel bol väčší ako 50%, diagnózy, resp. zhluky diagnóz sme spojili s ATC skupinou.

V ďalšom kroku sme skúmali, či sa jednotlivé ATC skupiny užívajú na chronické ochorenie, pričom sme zohľadnili rôzne faktory. Nakoniec sme našli vyše 250 takých ATC skupín, ktoré vyhovel našim podmienkam.

Tieto ATC skupiny môžeme považovať za dobrých kandidátov na vstup do modelu farmaceuticko-nákladových skupín. Na základe našej analýzy však nemôžeme tvrdiť, že tieto ATC skupiny jednoznačne indikujú chronické ochorenie. Naše podmienky totiž neboli príliš prísne. Stačilo napríklad, ak priemerné náklady poistencov užívajúcich liek z niektorej ATC skupiny boli len trochu väčšie ako populačný priemer.

Podmienky sme nastavili preto tak mierne, lebo sme považovali za menší hriech, ak sa na finálny zoznam dostala nakoniec taká ATC skupina, ktorá neimplikuje chronické ochorenie, ako keby sa tam nedostala taká ATC skupina, ktorá tam určite musí byť. Finálny zoznam sa dá takto ešte zúžiť, preskúmaním jednotlivých kódov diagnóz. Tiež

odporúčame skontrolovanie zoznamu odborníkom - lekárom.

Cieľ štvrtrej kapitoly považujeme za splnený - podarilo sa nám nájsť také ATC skupiny, ktoré sú dobrými kandidátmi na vstup do farmaceuticko-nákladových skupín. Splnenie tohto cieľa považujeme za ďalší prínos našej diplomovej práce.

V Kapitole 5 bolo našim cieľom identifikovať skupinu poistencov, ktorí sú podkompenzovaní a navrhnúť riešenie na zlepšenie ich kompenzácie. Zistili sme, že modelom predikované hodnoty sú pre poistencov užívajúcich relevantné lieky pre farmaceuticko-nákladové skupiny v menšom množstve ako 181 štandardných dávok systematicky nižšie, ako ich skutočné náklady. Navrhli sme preto na základe článku [11] úpravu modelu - obmedzenie na predikované hodnoty. Nový model výrazne znížil rozdiel medzi predikovanými a skutočnými nákladmi pre vybranú skupinu poistencov. Predikčná sila tohto modelu však bola výrazne nižšia, ako predikčná sila pôvodného modelu farmaceuticko-nákladových skupín. To znamená, že pre ostatných poistencov predikoval nový model horšie hodnoty, ako model farmaceuticko-nákladových skupín. Preto takáto úprava modelu síce zníži motiváciu poistovní pri selekcii rizika vybraných skupín, na druhej strane však ostatné skupiny poistencov môžu byť pre nich menej atraktívne.

Aby sme tomuto problému predišli, odporúčame obmedziť predikované hodnoty len pre menšiu skupinu poistencov. V našej analýze sme nastavili obmedzenie pre skoro 5% pozorovaní. Ukázalo sa však, že toto je príliš silné obmedzenie. Predmetom ďalšieho výskumu môže byť skúmanie menšieho obmedzenia na predikčnú silu modelu a nájdenie rovnováhy medzi zhoršením predikčnej sily modelu a zlepšením podkompenzácie vybraných skupín.

Tematika podkompenzácie vybraných skupín poistencov je veľmi aktuálna a necháva ešte veľký priestor na ďalší výskum. My sme v našej práci identifikovali jednu skupinu poistencov ktorá je podkompenzovaná, no určite existujú aj ďalšie. Autori článku [11] napríklad zistili, že v Holandsku sú podkompenzovaní tí poistenci, ktorí v predchádzajúcom roku užívali fyzioterapiu. Podobná identifikácia podkompenzovaných skupín by bola možná aj na Slovensku.

Cieľ piatej kapitoly sme splnili, navrhli sme spôsob na zlepšenie podkompenzácie vybranej skupiny poistencov. Cieľ práce považujeme tiež za splnený. Vybrali a popísali sme metódy data miningu - lineárnu regresiu, zhlukovú analýzu alebo krížovú validá-

ciu a aplikovali sme ich v kontexte kompenzácie rizikovej štruktúry na dáta zdravotnej poisťovne. Dostali sme pri tom zaujímavé výsledky - predikčná sila modelu na nových dátach, zoznam ATC skupín, ktoré sú vhodnými kandidátmi pre vstup do modelu farmaceuticko-nákladových skupín a zlepšenie kompenzácie vybranej skupiny poistencov.

Zoznam použitej literatúry

- [1] ATC klasifikácia liečiv, dostupné na internete (1.5.2018):
<https://www.adc.sk/databazy/klasifikacie/atc-skupiny.html>
- [2] Bodóová, A., *Vplyv kvantitatívneho uvoľňovania na štruktúru finančných trhov*, Bakalárska práca, FMFI UK, 2016
- [3] Duncan, I., *Healthcare Risk Adjustment and Predictive Modeling*, ACTEX Publications, Inc., Winsted, 2011
- [4] Glova J., Gavurová B., *Vybrané aspekty systémov zdravotnej starostlivosti*, GRANT journal (2013), 16-24, dostupné na internete (1. 4. 2018):
<http://www.grantjournal.com/issue/0202/PDF/0202glova.pdf>
- [5] Han, J., Kamber, M. Pei, J.: *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, Watham, 2012
- [6] van Kleef, R. C., McGuire, T. G., van Vliet, R. C. J. A., van de Ven, W. P. P. M.: *Improving risk equalization with constrained regression*, The European Journal of Health Economics (2017), 1137-1156, dostupné na internete (8.1.2017):
<https://link.springer.com/article/10.1007>
- [7] NCZI, *Medzinárodná klasifikácia chorôb, 10. revízia*, dostupné na internete (1.5.2018): <http://www.nczisk.sk/Standardy-v-zdravotnictve/Pages/MKCH-10-Revizia.aspx>
- [8] Lamers, L. M., van Vliet, R. C. J. A.: *The Pharmacy-based Cost Group model: validating and adjusting the classification of medications for chronic conditions to the Dutch situation*, Health Policy (2004), 113-121, dostupné na internete (8.1.2018):
<http://www.sciencedirect.com/science/article/pii/S0168851003001623>
- [9] Van de Ven, W.P.M.M., Ellis, R.P.: *Risk adjustment in competitive health insurance markets* In: Culyer, A.J., Newhouse, J.P. (eds.) *Handbook of health economics*, pp. 755–845. North- Holland, Amsterdam, 2000

- [10] van Kleef, R. C., van Vliet, R. C. J. A., van de Ven, W. P. P. M.: *Overpaying morbidity adjusters in risk equalization models*, The European Journal of Health Economics (2016), 885-895, dostupné na internete (8.1.2018): <https://link.springer.com/article/10.1007/s10198-015-0729-2>
- [11] van Kleef, R. C., McGuire, T. G., van Vliet, R. C. J. A., van de Ven, W. P. P. M.: *Improving risk equalization with constrained regression*, The European Journal of Health Economics (2017), 1137-1156, dostupné na internete (18.4.2018): <https://link.springer.com/article/10.1007/s10198-016-0859-1>
- [12] Pažitný, P. a kol., *Dobrý poistenec je nad zlato. Vieme ho vyvážiť?*, Stredoeurópsky inštitút pre zdravotnú politiku - Health Policy Institute, 2008
- [13] Rawlins, J. O., Pantula, S. G., Dickey D. A., *Applied Regression Analysis: A Research Tool, Second edition*, Springer-Verlag New York Inc., 1998
- [14] Sen, A., Srivastava M., *Regression Analysis Theory, Methods and Applications*, Springer-Verlag New York Inc., 1990
- [15] Somorčík, J., *Ekonometria, prednášky*, FMFI UK, 2016
- [16] Správa o stave vykonávania verejného zdravotného poistenia za rok 2016, Úrad pre dohľad nad zdravotnou starostlivosťou, dostupné na internete (1.4.2018): <http://www.udzs-sk.sk/documents/77197/0/sprava+o+stave+VZP.pdf/b60496ba-a30d-47b1-9247-8038a61bd251>
- [17] Uznesenie Výboru Národnej rady Slovenskej republiky pre zdravotníctvo z 28. novembra 2017 k vládnemu návrhu zákona, ktorým sa mení a dopĺňa zákon č. 576/2004 Z. z. o zdravotnej starostlivosti, službách súvisiacich s poskytovaním zdravotnej starostlivosti a o zmene a doplnení niektorých zákonov v znení neskorších predpisov a ktorým sa menia a dopĺňajú niektoré zákony
- [18] Vyhláška č. 346/2017 Z. z. Ministerstva zdravotníctva Slovenskej republiky, ktorou sa ustanovuje index rizika nákladov na zdravotnú starostlivosť na rok 2018

- [19] Vyhláška Ministerstva zdravotníctva Slovenskej republiky, ktorou sa ustanovuje zoznam farmaceuticko-nákladových skupín na rok 2018, LP/2017/855
- [20] ZÁKON z 20. júna 2012, ktorým sa mení a dopĺňa zákon č. 580/2004 Z. z. o zdravotnom poistení a o zmene a doplnení zákona č. 95/2002 Z. z. o poisťovníctve a o zmene a doplnení niektorých zákonov v znení neskorších predpisov a o zmene a doplnení niektorých zákonov

Príloha A

Názov	Kód
Infekčné a parazitárne choroby	(A00-B99)
Nádory	(C00-D48)
Choroby krvi a krvotvorných orgánov a daktoré poruchy imunitných mechanizmov	(D50-D89)
Choroby žliaz s vnútorným vylučovaním, výživy a premeny látok	(E00-E90)
Duševné poruchy a poruchy správania	(F00-F99)
Choroby nervového systému	(G00-G99)
Choroby oka a jeho adnexov	(H00-H59)
Choroby ucha a hlávkového výbežku	(H60-H95)
Choroby obehovej sústavy	(I00-I99)
Choroby dýchacej sústavy	(J00-J99)
Choroby tráviacej sústavy	(K00-K93)
Choroby kože a podkožného tkaniva	(L00-L99)
Choroby svalovej a kostrovej sústavy a spojivového tkaniva	(M00-M99)
Choroby močovej a pohlavnej sústavy	(N00-N99)
Ťarchavosť, pôrod a popôrodie	(O00-O99)
Niektoré choroby vznikajúce v perinatálnej perióde	(P00-P96)
Vrodené chyby, deformácie a chromozómové anomálie	(Q00-Q99)
Subjektívne a objektívne príznaky a abnormálne klinické a laboratórne nálezy nezatriedené inde	(R00-R99)
Poranenia, otravy a daktoré iné následky vonkajších príčin	(S00-T98)
Vonkajšie príčiny chorobnosti a úmrtnosti	(V01-Y98)
Faktory ovplyvňujúce zdravotný stav a styk so zdravotníckymi službami)	(Z00-Z99)

Tabuľka 7: Medzinárodná klasifikácia chorôb - 10. revízia [7]

Príloha B

Hlavná skupina	Názov
A	Tráviaci trakt a metabolizmus
B	Krv a krvotvorné orgány
C	Kardiovaskulárny systém
D	Dermatologiká
G	Urogenitálny trakt a pohlavné hormóny
H	Systémové hormonálne liečivá s výnimkou pohlavných hormónov a inzulínov
J	Antiinfektíva na systémové použitie
L	Cytostatiká a imunomodulátory
M	Muskuloskeletálny systém
N	Centrálna nervová sústava
P	Antiparazitiká
R	Respiračný systém
S	Zmyslové orgány
V	Rôzne (vária)

Tabuľka 8: Anatomicko-terapeutické skupiny

Príloha C

Pora- die	Kód	Názov	ATC skupina	Kód diagnózy
1.	AST	Astma	R03 okrem (R03AC18, R03AK03, R03BB, R03BC01, R03CC13, R03CC02)	J40 - J47
2.	CFP	Cystická fibróza a/alebo ochorenie exokrinnej funkcie pankreasu	J01GB01, J01XB01, R05CB13	E84
3.	CNS	Ochorenie mozgu/miechy	L03AB07, L03AB08, L03AB13, L03AX13, L04AA23, L04AA27, L04AA31, L04AA34, M03BX01, M03BX02, N07XX09	G35 - G37, G80 -G83, I60 -I69
4.	COP	CHOCHP/ťažká astma	R03AC18, R03AK03, R03BB	J40 - J47
5.	CRO	Crohnova choroba, ulcerózna kolitída	A07EA06, A07EC02	K50 - K52
6.	DEP	Liečba antidepresívami	N06A okrem (N06AA09, N06AX21)	
7.	DM1	Diabetes typu I	A10A	E10-E14
8.	DM2	Diabetestypu II	A10B	E10-E14

9.	DMH	Diabetes s hypertenziou	(A10A a zároveň [C02 okrem (C02KX), C03 okrem (C03CA01), C07, C08 okrem (C08CA06), C09]) alebo (A10B a zároveň [C02 okrem (C02KX), C03 okrem (C03CA01), C07, C08 okrem (C08CA06), C09])	I10-I25
10.	EPI	Epilepsia	N03 okrem (N03AX12, N03AX16, N03AE01)	G40, G41
11.	HIV	HIV/AIDS	J05AF okrem (J05AF07, J05AF08, J05AF10, J05AF11), J05AG, J05AR, J05AX okrem (J05AX15, J05AX16, J05AX65, J05AX67)	B20 - B24, Z21
12.	CHO	Hypercholesterolémia	C10 okrem C10AC01	E78
13.	KVS	Srdcové choroby	C01A, C01B, C01D, C01EB15, C01EB17, C03CA01	I20 -I52
14.	ONK	Malignity	L01 okrem (L01BA01), L03AA, L04AX04	C00 - D48
15.	PAR	Parkinsonova choroba	N04B	G20, G21.G22
16.	PSY	Antipsychotiká, Alzheimerova choroba, liečba závislostí	N05A okrem (N05AL03, N05AN01), N06DA, N06DX01, N07BB, N07BC51	F00 - F99, G30 - G32
17.	RAS	Liečba rastovým hormónom	H01AC01	E23, Q96

18.	REN	Renálne zlyhanie	B03X, V03AE	N17-N19, D46
19.	REU	Reumatické choroby liečené inak ako inhibítormi TNF	A07EC01, L01BA01, L04AA13, L04AX03, M01CB01.P01BA02	M05 - M09, M30 - M36
20.	TRA	Transplantácie	L04AA06, L04AA10, L04AA18, L04AD01, L04AD02	
21.	HOR	Hormonálna onkoliečba	L02	COO - D48
22.	NPP	Neuropatická bolesť	N01BX04, N03AX12, N03AX16	
23.	HEM	Hemofília	B02BD okrem B02BD01	D66 - D68
24.	AUT	Autoimunitné ochorenia liečené biologickou liečbou	L04AA11, L04AA24, L04AA26, L04AA32, L04AA33, L04AB, L04AC	K50 - K52, L40, M05 - M09, M30 - M36, M45
25.	PAH	Primárna pľúcna hypertenzia	B01AC11, B01AC21, C02KX, G04BE03	I27.0
26.	THY	Poruchy štítnej žľazy	H03A, H03B	E00 - E07

Príloha D

Zhluk	Diagnóza
1	B20, B23
2	E10, E11
3	N79, Z31
4	L20, L30
5	F32, F33
6	K50, K51
7	C19, C20
8	J04, J06
9	I70, I83
10	E55, Z00
11	B00, B02
12	N31, N39, R32
13	C17, C76
14	A14, A16
15	C44, L57
16	C43, C96, D45, D47
17	M07, M45
18	S01, S61, S81
19	L63, L64
20	I63, I69
21	O28, Z34
22	L02, L08
23	D80, D81, D83, D89
24	E29, N48, N50
25	N41, N42, N46
26	H59, Z96
27	P27, Q21
28	K22, R13

29	L94, M34
30	C67, R31
31	K71, K73
32	H82, H90, H93
33	B82, B83
34	H04, K02, K14

Tabuľka 9: Zhluky diagnóz na základe predpisu podobných liekov

Príloha E

Diagnóza	ATC skupiny
A14, A16	J04AB02
B18	J05AB04, J05AF07, L03AB11
C18	L01XA03, L01XX19
C34	L01CB01, L01XE03
C43, C96, D45, D47	L01XX35, L03AB04
C50	L01CD01, L01DB03, L01XC03, L02AE03, L02BA01, L02BG03, L02BG04, L02BG06
C61	L02AE02, L02BB03, L02BX02
C64	L01XE04
C90	L01XX32, L04AX04
D25	G03XB02
D50	B03AD, B03AE10
D66	G03XB02
E05	H03BA02, H03BB02
E10, E11	A10AB01, A10AB04, A10AB05, A10AB06, A10AC01, A10AD01, A10AD04, A10AD05, A10AE04, A10AE05, A10AE06, A10BA02, A10BB01, A10BB07, A10BB08, A10BB09, A10BB12, A10BD02, A10BD05, A10BD07, A10BD08, A10BD11, A10BD13, A10BG03, A10BH01, A10BH02, A10BH04, A10BH05, A10BX04, A10BX07, A10BX09, A10BX10, A10BX11, A10BX12
E22	G02CB03, G02CB04
E23	H01AC01

E78	C10AA01, C10AA02, C10AA04, C10AA05, C10AA07, C10AB05, C10AX09, C10BA02, C10BA04
F00	N06DA02, N06DA04, N06DX01
F10	N07BB03, N07BB04
F20	N05AB02, N05AE03, N05AE04, N05AH02, N05AX12, N05AX13
F32, F33	N06AA08, N06AX11, N06AX12, N06AX16, N06AX21, N06AX22, N06AX26
F90	N06BA04, N06BA09
G20	N04BA02, N04BA03, N04BC09, N04BD02
G35	L03AB07, L03AB08, L03AX13, L04AA23, N07XX09
G40	N03AB02, N03AF02, N03AG01, N03AG04, N03AX03, N03AX09, N03AX11, N03AX14, N03AX18
G70	N07AA02
H40	S01EA05, S01EB01, S01EC03, S01EC04, S01EC54, S01ED01, S01ED02, S01ED05, S01ED51, S01EE01, S01EE03, S01EE04, S01EE05
I10	C02AB01, C02AC05, C02AC06, C02CA06, C03AA03, C03BA, C03BA11, C03EA01, C03EA06, C07AA, C07AB03, C07AB05, C07AB07, C07AB08, C07AB12, C07BB03, C07BB07, C07FB07, C08CA01, C08CA02, C08CA03, C08CA08, C08CA09, C08CA13, C08DA01, C08DA51, C08DB01, C09AA01, C09AA02, C09AA03, C09AA04, C09AA05, C09AA06, C09AA09, C09AA10, C09AA11, C09AA16, C09BA02, C09BA03, C09BA04, C09BA05, C09BA06, C09BA09, C09BB03, C09BB04, C09BB07, C09BX01, C09CA01, C09CA02, C09CA03, C09CA04, C09CA06, C09CA07, C09DA01, C09DA02, C09DA03, C09DA04, C09DA06, C09DA07, C09DB01, C09DB04, C09DB06, C09XA02, C09XA52, C10BX03
I25	B01AC10, C01AA05, C01DA14, C01DX12, C01EB15, C01EB17, C03DA04
J30	R06AX26, R06AX28
J44	R03AC19, R03AL04, R03BB04, R03BB05, R03BB06, R03DX07

J45	R03AC04, R03AC12, R03AK03, R03AK06, R03AK07, R03AK10, R03BA02, R03BA05, R03BA07, R03BA08, R03DC03, R03DX05
K50, K51	A07EA06, A07EC02, A07FA, L04AB02, L04AX01
K59	A06AX05
K86	A09AA02
L20, L30	D02AE01
L40	D05BB02, D07XC01
M05	L01BA01, L04AA13, L04AC07
M10	M04AA03
M80	H05AA02
M81	A11CC03, A12AX, G03XC01, M05BA06, M05BA07, M05BX03
N18	A11GA01, B03XA01, B03XA02, B03XA03, H05BX02, V03AE02, V03AE03, V03AE04, V06DD
N20	G04BC
N31, N39, R32	G04BD04, G04BD06, G04BD08, G04BD09, G04BD10, G04BD11, G04BD12, N07AA03
N40	G04CA01, G04CA02, G04CA03, G04CA04, G04CB01
N79, Z31	G03GA02
N80	G03D
N95	G03CA03, G03CX01, G03FA01, G03FA14, G03FA17, G03FB08, G03FB09
Z94	J05AB14, L04AA06, L04AD02

Tabuľka 10: Výsledok analýzy