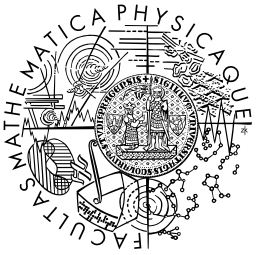

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Bc. Pavel Škopek

Predikce rozdělení počtu úmrtí s aplikacemi v ocenění životních smluv

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. RNDr. Branda Martin, Ph.D.
Studijní program: Finanční a pojistná matematika
Studijní obor: Finanční a pojistná matematika

Praha 2023

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

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

V ..... dne ....

Podpis autora

Chtěl bych tímto poděkovat doc. RNDr. Martinu Brandovi, Ph.D. za vedení mé diplomové práce, za cenné rady, vstřícný přístup a trpělivost. Rovněž bych chtěl poděkovat svým kamarádům PhDr. Josefu Försterovi Ph.D. a Bc. Pavlu Kortusovi za korekturu a v neposlední řadě svým rodičům. Těm děkuji za korekturu, ale hlavně za toleranci, psychickou podporu a prostor, který mi během psaní této práce dopřáli.

Název práce: Predikce rozdělení počtu úmrtí s aplikacemi v ocenění životních smluv

Autor: Bc. Pavel Škopek

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. RNDr. Branda Martin, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá modelováním úmrtností a oceňováním životních smluv. Nejprve jsou představeny základní pojmy z oblasti demografického modelu a úmrtnostních tabulek. Následuje popis Leeova–Carterova modelu včetně tří metod pro odhad parametrů a predikce budoucích hodnot. Dále se v textu rozebírá Renshawův–Habermanův model a metoda využívající analýzu kompozičních dat včetně neparametrické bootstrap metody pro intervalové odhady. Kromě teoretické části práce obsahuje i praktickou kapitolu, kde se modelují česká data úmrtnosti zvlášť pro muže a ženy. Na základě dat z období 1970-2021 nalezneme vhodný model, provedeme predikci budoucích hodnot na 30 let dopředu a oceníme životní smlouvy v roce 2021 i v budoucnosti.

Klíčová slova: úmrtnostní míra, počet zemřelých, Leeův–Carterův model, Renshawův–Habermanův model, analýza kompozičních dat

Title: Forecasting age distribution of death counts with applications in life insurance pricing

Author: Bc. Pavel Škopek

Department: Department of Probability and Mathematical Statistics

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

Abstract: This thesis deals with the topic of mortality modelling and life insurance pricing. First, basic concepts from the demographic model and life tables are introduced. Following is a description of the Lee–Carter model including three methods for estimating of parameters and predicting of future values. The thesis also analyses the Renshaw–Haberman model and the method, which uses compositional data analysis including the non-parametric bootstrap for interval estimations. Besides the theoretical part the thesis also contains a practical one, where Czech mortality data are modeled separately for men and women. Based on the data from 1970-2021 we select the best model, predict future values for 30 years ahead and price the life insurance in 2021 and in the following years.

Keywords: mortality rate, death counts, Lee–Carter model, Renshaw–Haberman model, compositional data analysis

Obsah

Seznam symbolů

x — *Diskrétní věk*
Nabývá hodnot $x_1, \ldots, x_m$.

t — *Diskrétní čas*
Nabývá hodnot $t_1, \ldots, t_n$.

T_x — *Zbývající délka života pro jedince ve věku x*
Jedná se o spojitou náhodnou veličinu na pravděpodobnostním prostoru (Ω, F, P).

${}_tq_x$ — *Pravděpodobnost, že jedinec ve věku x zemře v následujících t rocích*
Jedná se o pravděpodobnost $P(T_x \leq t)$, což také odpovídá distribuční funkci náhodné veličiny T_x. Můžeme značit též $F_x(t)$.

${}_tp_x$ — *Pravděpodobnost, že jedinec ve věku x nezemře v následujících t rocích*
Jedná se o pravděpodobnost $P(T_x > t)$, což také odpovídá funkci přežití $S_x(t)$.

η_x — *Intenzita úmrtnosti ve věku x*

$D_{x,t}$ — *Počet úmrtí jedinců mezi věky x a x+1 v čase t*

$d_{x,t}$ — *Tabulkový počet zemřelých jedinců z hypotetické generace mezi věky x a x+1 v čase t*

$\mathbf{D}$ — *Matice počtu úmrtí*
Prvky této matice jsou $D_{x,t}$ resp. $d_{x,t}$, což odpovídá počtu úmrtí mezi věky x a $x+1$ v čase t, resp. tabulkovému počtu zemřelých mezi věky x a $x+1$ v čase t. Matice má rozměr $(m \times n)$.

$P_{x,t}$ — *Střední stav populace ve věku x v čase t*
Jedná se o počet jedinců ve věku x k 1. červenci roku t.

$l_{x,t}$ — *Tabulkový počet dožívajících se věku x v čase t*

$L_{x,t}$ — *Tabulkový počet žijících*
Jedná se o součet dob všech jedinců prožitých mezi věky x a $x+1$.

$N_{x,t}$ — *Centrální expozice ve věkovém intervalu (x, x+1) v čase t*
Jedná se o součet všech dob strávených jednotlivými jedinci populace v daném kalendářním roce t ve věkovém intervalu $(x, x+1)$.

a_x — *Neznámý parametr závislý na věku x, v LC modelu odhadnut průměrem úmrtnostních měr*

$\mathbf{a}$ — *Vektor neznámých parametrů a_x závislých na věku x*
Tato proměnná je specifická pro každý věk, má rozměr $(m \times 1)$ a platí $\mathbf{a} = (a_{x_1}, \ldots, a_{x_m})^T$.

$\mathbf{A}$	*Matice neznámých parametrů a_x závislých na věku x* Jedná se o matici, která je složena z vektorů $(\mathbf{a},\ldots,\mathbf{a})$. Matice má rozměr $(m \times n)$.
$\dot{\mathbf{a}}_x$	*Vektor shodných a_x* Platí $\dot{\mathbf{a}}_x = (a_x,\ldots,a_x)^T$ a tento vektor má rozměr $(n \times 1)$.
b_x	*Neznámý parametr závislý na věku x vyjadřující citlivost na změnu v časovém indexu*
$\mathbf{b}$	*Vektor neznámých parametrů b_x závislých na věku x* Tato proměnná je specifická pro každý věk, má rozměr $(m \times 1)$ a platí $\mathbf{b} = (b_{x_1},\ldots,b_{x_m})^T$.
k_t	*Časová řada úrovně úmrtnosti v čase t*
$\mathbf{k}$	*Vektor časové řady úrovně úmrtnosti* Tato proměnná je specifická pro každý uvažovaný čas, má proto rozměr $(n \times 1)$ a platí $\mathbf{k} = (k_{t_1},\ldots,k_{t_n})^T$.
$m_{x,t}$	*Centrální míra úmrtnosti pro věk x v čase t*
$\mathbf{m}$	*Matice centrálních měr úmrtností* Prvky této matice jsou $m_{x,t}$, což odpovídá centrální míře úmrtnosti pro věk x v čase t. Matice má rozměr $(m \times n)$.
$\mathbf{m}_{\cdot,t}$	*Vektor centrálních měr úmrtnosti v čase t* Platí $\mathbf{m}_{\cdot,t} = (m_{x_1,t},\ldots,m_{x_m,t})^T$.
$\mathbf{m}_{x,\cdot}$	*Vektor centrálních měr úmrtnosti ve věku x* Platí $\mathbf{m}_{x,\cdot} = (m_{x,t_1},\ldots,m_{x,t_n})^T$.
$_{x,t}$	*Chybový člen modelu pro věk x v čase t*
	Matice chybových členů modelu
$\varphi_{x,l}$	*l-tá hlavní komponenta pro věk x*
$\beta_{l,t}$	*Faktorové skóre l-té hlavní komponenty pro čas t*
$z_{x,t}$	*Transponovaná kompoziční data*
$\hat{p}$	*Modelový odhad parametru p*
$\tilde{p}$	*Budoucí hodnota parametru p*
$\bar{p}$	*Aritmetický průměr parametrů p_i*
$\frac{\partial}{\partial x_i} f(x)$	*Parciální derivace funkce $f(x)$ podle x_i*
$\frac{d}{dx} f(x),\ f'(x)$	*Derivace funkce $f(x)$*
O_{LS}	*Součet čtvercových chyb modelu*
$L(\beta)$	*Věrohodnostní funkce*

$\mathrm{Dev}(X, \hat{X})$	*Odchylka odhadnutého parametru $\hat{X}$ od empirické hodnoty X*
AIC	*Akaikeho informační kritérium*
$x^{(b)}$	*Bootstrap hodnota parametru x*
R^2	*Koeficient determinace*
MAPE	*Střední absolutní procentuální chyba*
S_γ	*Intervalové skóre s pravděpodobností pokrytí* $100(1-\gamma)$
$a_{x:\overline{n}\vert}$	*Střední současná hodnota budoucích pojistných plnění (v závislosti na úrokové míře) na smlouvě dočasného okamžitého důchodu s výplatami na konci období*
a_x	*Střední současná hodnota budoucích pojistných plnění (v závislosti na úrokové míře) na smlouvě doživotního okamžitého důchodu s výplatami na konci období*
${}_{m\vert}a_{x:\overline{n}\vert}$	*Střední současná hodnota budoucích pojistných plnění (v závislosti na úrokové míře) na smlouvě odloženého dočasného důchodu s výplatami na konci období*
${}_{m\vert}a_x$	*Střední současná hodnota budoucích pojistných plnění (v závislosti na úrokové míře) na smlouvě odloženého doživotního důchodu s výplatami na konci období*

Úvod

To, že každý jednou zemřeme, je jisté, ale kdy to přijde? To samozřejmě nikdo neví, matematicky však lze určit mnoho údajů, které nám leccos napoví. Abychom mohli s úmrtími pracovat, potřebujeme data. Ty se shromažďují již mnoho let, konkrétně Český statistický úřad (Český statistický úřad (2023)) vede přehlednou evidenci od roku 1920.

Důvodů, proč se úmrtnostmi zabývat, je mnoho. Kromě zvědavosti každého z nás, jak dlouhý náš život bude, to může být například riziko pro pojišťovny či penzijní fondy, které nabízejí různé produkty, jež závisí právě na úmrtnosti populace. V posledních letech je taktéž velmi diskutovaným tématem stanovování hranice pro odchod do důchodu. I zde je zapotřebí znát úmrtnosti populace, aby se hranice zvolila vhodně a stát mohl zdravě fungovat.

Tato práce popisuje několik metod pro modelování úmrtnostních měr či počtu úmrtí. Tyto metody se následně aplikují na pozorovaná data, provede se predikce budoucích hodnot, vybere se nejlepší metoda a určí se cena životních produktů.

Práce je členěná do sedmi kapitol, kterým předchází seznam nejčastěji používaného značení. První kapitola obsahuje představení základních pojmů demografického modelu a úmrtnostních tabulek. Odvozeny budou rovněž důležité vztahy a vlastnosti mezi základními veličinami.

V druhé kapitole se seznámíme s Leeovým–Carterovým modelem (Lee a Carter (1992)), jedním z nejznámějších demografických modelů. Odhad parametrů modelu popíšeme pomocí tří metod (metoda singulárního rozkladu, Poissonova metoda, alternativní metoda) a provedeme předpověď budoucích hodnot. Třetí kapitola bude pojednávat o rozšířené verzi Leeova–Carterova modelu o druhou sadu singulárních vektorů - Renshawovým–Habermanovým modelu (Renshaw a Haberman (2003)). Model bude popsán, parametry modelu odhadnuty pomocí dvou metod (metoda singulárního rozkladu, Poissonova metoda) a budou predikovány budoucí hodnoty. Čtvrtá kapitola představí nejnovější model, který pro modelování využívá analýzu kompozičních dat (Shang a Haberman (2019)). Model bude rovněž popsán, parametry modelu odhadnuty a detailněji se zaměříme na možnosti predikování budoucích hodnot.

V páté kapitole zavedeme tři kritéria, podle kterých budeme schopni vybrat nejlepší model. Šestá kapitola se bude zabývat oceňováním životních smluv. Rozebereme si rozdíly mezi dočasnými a doživotními smlouvami a také mezi okamžitými a odloženými smlouvami.

Poslední kapitola obsahuje rozsáhlý praktický příklad, ve kterém se budeme snažit nalézt nejvhodnější model pro česká data úmrtí mužů a žen. Pokusíme se předpovědět, jak bude vypadat rozdělení počtu úmrtí v budoucnu, oceníme životní smlouvy popsané v šesté kapitole a ukážeme si, proč je pro pojišťovny či penzijní fondy důležité se zabývat predikcí počtu úmrtí.

Práce primárně vychází z článku Lee a Carter (1992) ve druhé kapitole, z díla Renshaw a Haberman (2003) ve třetí kapitole a z Shang a Haberman (2019) ve čtvrté kapitole.

1. Úmrtnost a základní pojmy

Tato kapitola má za úkol představení základních pojmů úmrtnosti, vztahů mezi nimi a značení, které se při analýze úmrtnosti používá. V této kapitole vycházíme ze zdrojů Cipra (1990) a Cipra (1999).

Analýza úmrtnosti spadá do oblasti demografie, což je věda, jež zkoumá strukturu, velikost a vývoj lidské populace. Důvodů, proč se zabývat analýzou úmrtnosti, je několik. Analýza úmrtnosti je důležitá například pro zdravotnictví, které na jejím základě může identifikovat problém a následně ho vyřešit. Další aktuální problém, ve kterém se využívá analýzy úmrtnosti, je hledání vhodné hranice pro odchod do důchodu. Protože v úmrtnosti je pozorován trend, že populace umírá ve vyšších věcích, je namístě se zamyslet nad vhodným věkem pro odchod do důchodu. V neposlední řadě se analýza úmrtnosti hojně využívá v penzijních fondech, pojišťovnictví či v bankovnictví, kde se pojí s cenotvorbou i rezervováním.

Pokud se míra úmrtnosti zvýší pro nižší věky, tedy lidé budou dříve umírat, pak to může signalizovat hrozbu pro pojišťovny, neboť budou muset více vyplácet v případě rizikového životního pojištění. Naopak, pokud by se míra úmrtnosti snížila, hrozbou pro pojišťovny mohou být produkty pojištění dožití či životní důchody, kde se vyplácí pojistné plnění, pokud jedinec přežije stanovenou dobu, nebo v případě důchodů probíhají výplaty pojistných plnění, pokud je jedinec naživu.

1.1 Demografický model

Uvažujme jedince ve věku x. Označme T_x spojitou náhodnou veličinu na pravděpodobnostním prostoru (Ω, F, P), která značí **zbývající délku života**. Pro tuto náhodnou veličinu T_x zaveďme distribuční funkci

$$F_x(t) = \mathrm{P}(T_x \leq t), \qquad t \geq 0. \tag{1.1}$$

Pokud má náhodná veličina T_x hustotu, pak ji budeme značit jako $f_x(t)$:

$$f_x(t) = F_x(t) \qquad t \geq 0, \tag{1.2}$$

kde $F_x(t)$ je derivace funkce $F_x(t)$, za předpokladu, že derivace existuje.

Funkci $F_x(t)$ můžeme popsat jako pravděpodobnost, že jedinec ve věku x zemře do věku $x+t$. Definujme proto pravděpodobnost, že jedinec ve věku x se dožije věku $x + t$ jako **funkci přežití** $S_x(t)$:

$$S_x(t) = 1 - F_x(t) = \mathrm{P}(T_x > t), \qquad t \geq 0. \tag{1.3}$$

Nyní zavedeme aktuárské značení[1] pro již výše zmíněné pravděpodobnosti $F_x(t)$ a $S_x(t)$. Označme

$$\begin{aligned} {}_tq_x &= \mathrm{P}(T_x \leq t) = F_x(t), \qquad t \geq 0, \\ {}_tp_x &= \mathrm{P}(T_x > t) = S_x(t), \qquad t \geq 0. \end{aligned} \tag{1.4}$$

[1] Pokud vynecháme přední spodní index u pravděpodobností ${}_tp_x$ či ${}_tq_x$, vždy je uvažován horizont $t = 1$, neboli $q_x = {}_1q_x$ a $p_x = {}_1p_x$.

V následujících výpočtech budeme dále využívat **fundamentální předpoklad** (FA) neboli že rozdělení zbývající délky života T_{x+t} je stejné jako podmíněné rozdělení náhodné veličiny $T_x - t$ za podmínky, že zbývající délka života T_x je větší než t.

Díky tomuto předpokladu platí, že pravděpodobnost, že jedinec ve věku x přežije dalších $t+s$ roků, je rovna součinu pravděpodobností, že jedinec ve věku x přežije dalších t roků, a pravděpodobnosti, že jedinec ve věku $x+t$ nezemře během následujících s roků:

$$\begin{aligned} {}_{t+s}p_x &= \mathrm{P}(T_x > t+s) = \frac{\mathrm{P}(T_x > s)}{\mathrm{P}(T_x > s)}\mathrm{P}(T_x > t+s) = {}_sp_x \frac{\mathrm{P}(T_x > s, T_x > t+s)}{\mathrm{P}(T_x > s)} \\ &= {}_sp_x \mathrm{P}(T_x - s > t | T_x > s) = {}_sp_x \mathrm{P}(T_{x+s} > t) = {}_sp_x \cdot {}_tp_{x+s}, \end{aligned}$$

kde (FA) jsme využili v předposlední rovnosti.

Dále definujme střední hodnotu náhodné veličiny T_x, kterou budeme značit e_x^0 a budeme ji nazývat **střední zbývající délka života** ve věku x :

$$\begin{aligned} e_x^0 = \mathsf{E}(T_x) &= \int_0^\infty t f_x(t)\mathrm{d}t = \Big[tF_x(t) \Big]_{t=0}^\infty - \int_0^\infty F_x(t)\mathrm{d}t = \int_0^\infty 1\mathrm{d}t - \int_0^\infty F_x(t)\mathrm{d}t \\ &= \int_0^\infty (1 - F_x(t))\mathrm{d}t = \int_0^\infty S_x(t)\mathrm{d}t = \int_0^\infty {}_tp_x \mathrm{d}t, \end{aligned} \tag{1.5}$$

kde ve třetí rovnosti jsme využili metodu per-partes.

Dalším důležitým ukazatelem demografického modelu je **intenzita úmrtnosti** ve věku x, μ_x :

$$\mu_x = \lim_{h \to 0_+} \frac{\mathrm{P}(T_x \le h)}{h} = \lim_{h \to 0_+} \frac{{}_hq_x}{h}. \tag{1.6}$$

Z intenzity úmrtnosti ve věku $x+t$ za pomoci fundamentálního předpokladu můžeme odvodit užitečné vztahy mezi μ_{x+t} a ${}_tp_x$:

$$f_x(t) = {}_tp_x \mu_{x+t}, \tag{1.7}$$

$${}_tp_x = \exp\left(- \int_0^t \mu_{x+s}\mathrm{d}s \right). \tag{1.8}$$

Platí:

$$\begin{aligned} \mu_{x+t} &= \lim_{h \to 0_+} \frac{\mathrm{P}(T_{x+t} \le h)}{h} = \lim_{h \to 0_+} \frac{\mathrm{P}(T_x - t \le h | T_x > t)}{h} \\ &= \lim_{h \to 0_+} \frac{\mathrm{P}(t < T_x \le t+h)}{h\mathrm{P}(T_x > t)} = \lim_{h \to 0_+} \frac{F_x(t+h) - F_x(t)}{h} \frac{1}{{}_tp_x} \\ &= \frac{F_x(t)}{{}_tp_x} = \frac{f_x(t)}{{}_tp_x}, \end{aligned}$$

kde v druhé rovnici jsme využili předpokladu (FA), ve třetí rovnici definici podmíněné pravděpodobnosti a v předposlední rovnici definici derivace funkce $F(x)$. Dále platí:

$$\mu_{x+t} = \frac{F_x(t)}{{}_tp_x} = \frac{1}{{}_tp_x}\frac{\mathrm{d}}{\mathrm{d}t}{}_tq_x = \frac{1}{{}_tp_x}\frac{\mathrm{d}}{\mathrm{d}t}(1 - {}_tp_x) = -\frac{1}{{}_tp_x}\frac{\mathrm{d}}{\mathrm{d}t}{}_tp_x = -\frac{\mathrm{d}}{\mathrm{d}t}\ln({}_tp_x), \tag{1.9}$$

neboť

$$\frac{\mathrm{d}}{\mathrm{d}t}\ln({}_tp_x) = \frac{1}{{}_tp_x}\frac{\mathrm{d}}{\mathrm{d}t}{}_tp_x.$$

Nyní již z 1.9 snadno získáme vztah 1.8.

Na závěr definujeme **centrální míru úmrtnosti** m_x^C :

$$m_x^C = \frac{\int_0^1 S_0(x+u)\mu_{x+u}\mathrm{d}u}{\int_0^1 S_0(x+u)\mathrm{d}u} = \frac{S_0(x) - S_0(x+1)}{\int_0^1 S_0(x+u)\mathrm{d}u}. \tag{1.10}$$

Protože pravděpodobnosti ${}_tp_x$ a ${}_tq_x$ jsou v praxi dostupné pouze pro celočíselné hodnoty x a t, znamená to pro nás, že pokud chceme v praxi popsat rozdělení náhodné veličiny T_x, musíme přijmout některý z aproximačních předpokladů. Předpokládejme, že

$$T_x = K_x + S_x,$$

kde

$$K_x = \lfloor T_x \rfloor$$

je největší celé číslo, které je menší nebo rovno zbývající délce života, a

$$0 \leq S_x < 1.$$

Aproximační předpoklady pak jsou:

- Předpoklad linearity funkce ${}_uq_x$: ${}_uq_x = uq_x \quad 0 \leq u \leq 1$,
- Předpoklad konstantní intenzity úmrtnosti μ_{x+t}: $\mu_{x+t} = \mu_x \quad 0 \leq u \leq 1$,
- Aproximace S_x funkcí $S_x^{(m)} = \lfloor mS_x + 1 \rfloor / m$.

Pokud uvažujeme předpoklad konstantní intenzity úmrtnosti, pak je intenzita úmrtnosti totožná s centrální mírou úmrtnosti, neboť:

$$m_x^C = \frac{\int_0^1 S_0(x+u)\mu_{x+u}\mathrm{d}u}{\int_0^1 S_0(x+u)\mathrm{d}u} = \frac{\int_0^1 S_0(x+u)\mu_x\mathrm{d}u}{\int_0^1 S_0(x+u)\mathrm{d}u} = \frac{\mu_x \int_0^1 S_0(x+u)\mathrm{d}u}{\int_0^1 S_0(x+u)\mathrm{d}u} = \mu_x \tag{1.11}$$

a pro vztah mezi intenzitou úmrtnosti a pravděpodobností q_x platí:

$$q_x = 1 - p_x = 1 - \exp\left(-\int_0^1 \mu_{x+s}\mathrm{d}s\right) = 1 - \exp\left(-\int_0^1 \mu_x\mathrm{d}s\right) = 1 - \exp(-\mu_x). \tag{1.12}$$

1.2 Úmrtnostní tabulky

Úmrtnostní tabulky hrají důležitou roli v naší práci, neboť prostřednictvím těchto tabulek budeme získávat v praktické části vstupní data. Svou důležitost ale mají i jinak, protože je to přehledný způsob zaznamenání informací o úmrtnostech. Tabulky může vytvářet například statistický úřad, od něhož mohou tabulky převzít pojišťovny či penzijní fondy pro výpočet cen jejich produktů. Nyní si strukturu úmrtnostních tabulek popíšeme. Samozřejmě existuje mnoho druhů úmrtnostních tabulek, zde budeme popisovat úmrtnostní tabulky, které využívá Český statistický úřad (2023) pro muže a ženy v ČR.

Zmiňme ještě, že součástí tvorby tabulek je také vyhlazování či vyrovnávání hodnot (úmrtnostních měr nebo pravděpodobností úmrtí) pomocí matematicko-statistických metod. Rovněž probíhá úprava výpočtů hodnot pro vysoké věky, neboť často není k dispozici dostatek dat, a tím by mohly vznikat nepřesnosti v odhadech a v predikcích. Jiným způsobem se taktéž modeluje úmrtnost ve věku 0 (kojenecká úmrtnost), protože se jedná o velmi specifickou věkovou skupinu. Více informací o vyhlazovacích a vyrovnávacích metodách se můžeme dočíst na stránkách Český statistický úřad (2023).

Jednotlivé řádky v úmrtnostních tabulkách odpovídají jednotlivým věkům $x = 0, 1, 2, \ldots, 105+$, kde řádek pro $x = 105+$ zahrnuje všechny věky větší nebo rovno 105. Věk x je zároveň prvním sloupcem úmrtnostních tabulek. Hodnoty na daném řádku pro věk x odpovídají jedincům ve věkovém intervalu x až $x + 1$. Ve druhém sloupci zpravidla najdeme D_x, neboli **počet úmrtí** jedinců ve věkovém intervalu x až $x + 1$ v populaci.

Třetí sloupec odpovídá veličině P_x, což je **střední stav populace** ve věku x (někdy zaměnitelné též s centrální expozicí[2] N_x), jinými slovy počet jedinců ve věku x k 1. červenci daného roku.

Následuje sloupec s **(tabulkovou, centrální) mírou úmrtnosti** m_x^T, kterou můžeme vypočítat jako podíl počtu úmrtí v daném věku a středního stavu populace v daném věku:

$$m_x^T = \frac{D_x}{P_x}. \tag{1.13}$$

V dalším sloupci je zastoupena pravděpodobnost q_x. Jedná se o pravděpodobnost, že jedinec, který se dožil věku x se nedožije věku $x + 1$. Český statistický úřad (2023) využívá pro výpočet pravděpodobnosti q_x míru úmrtnosti m_x a dále parametr a_x, jenž odpovídá **průměrnému počtu člověkoroků prožitých v daném věkovém intervalu zemřelými jedinci**, což je s výjimkou první a poslední věkové skupiny nastaveno na $a_x = 0.5$, neboť se předpokládá, že úmrtnost během daného roku je rovnoměrně rozdělená.

Ve skupině $x = 0$ se, vzhledem ke specifičnosti skupiny a úmrtí v ní, počítá a_x jako pozorovaný průměr prožitých dob zemřelých jedinců v této věkové skupině. Výsledná hodnota se pohybuje kolem 0.1, což je zároveň i hodnota, která se použila v případě, že tato data nejsou dostupná. Ve skupině 105+ se parametr a_x spočítá jako inverzní hodnota m_x^T, neboť tím máme zaručeno, že $q_x = 1$, což požadujeme:

$$q_x = \frac{m_x^T}{1 + (1 - a_x)m_x^T}. \tag{1.14}$$

Druhá část úmrtnostních tabulek se váže k (hypotetické) tabulkové generaci, což je skupina jedinců, kteří se narodili ve stejný čas, a kteří umírají podle stejné pravděpodobnosti. Všechny následující veličiny jsou získány na základě pravděpodobností q_x.

První z nich je **tabulkový počet dožívajících** l_x, tedy počet jedinců z hypotetické tabulkové generace, kteří se dožili věku x. Hodnota l_0 značí počá-

[2] Centrální expozice odpovídá součtu všech dob strávených jednotlivými jedinci populace v daném kalendářním roce ve věkovém intervalu $(x, x + 1)$. Pokud předpokládáme, že data narození jedinců jsou rovnoměrně rozložená během kalendářního roku, okamžiky přistěhování, odstěhování a úmrtí jedinců jsou taktéž rovnoměrně rozloženy během kalendářního roku, pak můžeme střední stav populace P_x zaměnit s centrální expozicí N_x.

teční velikost hypotetické generace, což Český statistický úřad (2023) nastavil na $l_0 = 100000$. Platí:

$$l_{x+1} = l_x(1 - q_x). \tag{1.15}$$

Další proměnnou v úmrtnostní tabulce je **tabulkový počet zemřelých d_x.** Jedná se o počet zemřelých jedinců z hypotetické generace mezi věky x a $x + 1$. Platí:

$$d_x = l_x - l_{x+1} = l_x - l_x(1 - q_x) = l_x - l_x + l_x q_x = l_x q_x. \tag{1.16}$$

V následujícím sloupci se nachází **tabulkový počet žijících L_x,** který vyjadřuje hypotetický počet člověkoroků prožitých mezi věky x a $x + 1$. Jinými slovy se jedná o součet dob všech jedinců prožitých mezi věky x a $x+1$. Každý jedinec, který se dožije věku $x + 1$, přispěje do L_x celým rokem, ti jedinci, kteří zemřou mezi věkem x a $x+1$, přispějou poměrem prožité doby v daném intervalu. Platí:

$$L_x = l_x - (1-a_x)d_x = l_{x+1} + a_x d_x = l_{x+1} + a_x(l_+ - l_{x+1}) = a_x l_x + (1-a_x)l_{x+1}. \tag{1.17}$$

Vidíme, že L_x, lze také zapsat jako lineární kombinace jedinců dožitých věku x a jedinců dožitých věku $x + 1$. Protože vyjma první a poslední věkové skupiny je nastaveno $a_x = 0.5$, můžeme psát

$$L_x = \frac{1}{2}(l_x + l_{x+1}) \qquad x = 2, \ldots, 104 \tag{1.18}$$

a L_x tak může být chápáno jako **tabulkový střední stav populace** ve věku x.

Předposlední proměnnou v úmrtnostních tabulkách je **pomocný ukazatel T_x.** Jedná se o celkovou dobu prožitou všemi členy hypotetické generace od věku x. Vyjádřit to lze vztahem:

$$T_x = \sum_{y=x}^{105+} L_y. \tag{1.19}$$

Triviálně platí $T_{105} = L_{105}$ a zbylé hodnoty T_x lze získat rekurzivně za pomoci L_x.

Poslední sloupec úmrtnostních tabulek obsahuje **střední délku života e_x.** Tuto proměnnou lze popsat jako průměrný počet let, které ještě jedinec ve věku x z hypotetické generace prožije. Vztahem lze definovat jako:

$$e_x = \frac{T_x}{l_x}. \tag{1.20}$$

Zmiňme ještě jeden pojem, a tím je **(tabulková, generační, centrální) úmrtnostní míra m_x^{TG},** která se spočítá jako podíl tabulkového počtu zemřelých a tabulkového počtu žijících:

$$m_x^{TG} = \frac{d_x}{L_x}. \tag{1.21}$$

Nelze si nevšimnout, že jsme zavedli tři druhy centrální úmrtnostní míry, které vždy popisují velmi podobný vztah. V případě m_x^C se jedná o teoretickou úmrtnostní míru, v případě m_x^T se jedná o empirickou úmrtnostní míru, která byla spočítána na základě získaných dat. Proměnná m_x^{TG} je rovněž empirickou úmrtnostní mírou, která však vznikla na základě výpočtu hypotetické generace. Pro jednoduchost budeme používat zkráceně symbol m_x, který v případě potřeby budeme blíže specifikovat, o jaký konkrétní typ úmrtnostní míry jde.

Všechny parametry z úmrtnostních tabulek jsou závislé na věku. V následujících kapitolách budeme k těmto proměnným přidávat navíc druhý index t, který bude specifikovat, pro jaký rok byla úmrtnostní tabulka vytvořena. Značení parametrů tak bude $m_{x,t}, d_{x,t}, D_{x,t}$ apod.

2. Leeův–Carterův model

Jeden z nejznámějších demografických modelů (viz Lee a Carter (1992)) navrhli v roce 1992 Ronald Demos Lee a Lawrence Rielly Carter. Tento model je relativně jednoduchý a hojně využívaný. Jejich model je založen na mírách úmrtnosti, které jsou specifické pro jednotlivé věky. Logaritmy těchto měr úmrtností Lee s Carterem modelují jako lineární funkci nepozorovaného, časově specifického indexu intenzity s parametry závisejícími na věku. Svůj model Lee a Carter aplikovali na data amerických měr úmrtnosti z let 1900-1989 a predikovali úmrtnost až do roku 2065.

2.1 Popis modelu

Základní komponenta Leeova–Carterova modelu je centrální míra úmrtnosti $m_{x,t}$, která je specifická pro věk x a čas t.

Nyní vezmeme logaritmus matice $\mathbf{m}$ o rozměrech $(m \times n)$, který odhadneme pomocí dvou vektorů $\mathbf{a}$ a $\mathbf{b}$, specifických pro jednotlivé věky $x = x_1, \ldots, x_m$ a dále pomocí nepozorovatelného indexu $\mathbf{k}$, jenž se liší s časem t. Tvar závislosti maticově[1] vyjádříme následovně:

$$\ln \mathbf{m} = \mathbf{A} + \mathbf{b}\mathbf{k}^T + \quad , \tag{2.1}$$

což můžeme ekvivalentně vyjádřit rovněž jako

$$\ln m_{x,t} = a_x + b_x k_t + \quad _{x,t}, \qquad x = x_1, \ldots, x_m, \quad t = t_1, \ldots, t_n. \tag{2.2}$$

Pro odvozování některých odhadů uvedeme ještě vztah se sumami, který rovněž platí a plyne z vyjádření 2.1:

$$\sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} \ln m_{x,t} = \sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} a_x + \sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} b_x k_t + \sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} \quad _{x,t}. \tag{2.3}$$

Rovnice se také někdy udává ve tvaru exponenciály:

$$\mathbf{m} = e^{\mathbf{A} + \mathbf{b}\mathbf{k}^T + \quad}, \tag{2.4}$$

kterou lze obdobně popsat vztahem:

$$m_{x,t} = e^{a_x + b_x k_t + \quad _{x,t}}, \qquad x = x_1, \ldots, x_m, \quad t = t_1, \ldots, t_n. \tag{2.5}$$

Rovněž platí i vztah se sumami:

$$\sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} m_{x,t} = \sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} e^{a_x + b_x k_t + \quad _{x,t}}. \tag{2.6}$$

Ve všech případech chybový člen $_{x,t}$ je bílý šum.[2] Parametr $\mathbf{k}$ zde značí specifický index, který se vyvíjí v čase t a udává stupeň úmrtnosti. Prvek $e^{\mathbf{a}}$ zde

[1] $\mathbf{A}$ je matice $m \times n$, která obsahuje v každém sloupci vektor $\mathbf{a}$. Parametry a_x zapisujeme do matice kvůli korektnosti maticových dimenzí.

[2] Bílý šum je definován jako posloupnost nekorelovaných náhodných veličin s nulovou střední hodnotou a kladným konečným rozptylem σ^2. Značíme $\mathrm{WN}(0, \sigma^2)$.

vyjadřuje celkový tvar úmrtností napříč věky a **b** zaznamenává rychlost poklesu měr v závislosti na změně indexu **k**.

Naším úkolem je dále odhadnout vektory **a, b, k**. Takto vytvořený model je však podurčený, neboť snadno dokážeme najít nekonečně mnoho vektorů **a, b, k** splňujících rovnici 2.1. Je proto nutné přidat k modelu nějaká omezení na hledané vektory. Lee s Carterem volili následující restrikce:

$$\sum_{x=x_1}^{x_m} b_x = 1, \qquad \sum_{t=t_1}^{t_n} k_t = 0, \tag{2.7}$$

kde $x_1, \ldots, x_m$ značí konkrétní uvažované věky a $t_1, \ldots, t_n$ konkrétní uvažované časy.

Omezení 2.7 implikují, že a_x odpovídají průměru logaritmických měr $m_{x,t}$ přes čas t. To vyvodíme například z rovnice 2.3, kde místo $_{x,t}$ budeme brát střední hodnotu této proměnné. Odtud pro dané x snadno získáme, že $a_x = \frac{1}{n} \sum_{t=t_1}^{t_n} \ln m_{x,t}$.

$$\sum_{t=t_1}^{t_n} \ln m_{x,t} = \sum_{t=t_1}^{t_n} a_x + \sum_{t=t_1}^{t_n} b_x k_t, \qquad x = x_1, \ldots, x_m,$$

$$\sum_{t=t_1}^{t_n} \ln m_{x,t} = n \cdot a_x + b_x \sum_{t=t_1}^{t_n} k_t, \qquad x = x_1, \ldots, x_m,$$

$$a_x = \frac{1}{n} \sum_{t=t_1}^{t_n} \ln m_{x,t}, \qquad x = x_1, \ldots, x_m,$$

přičemž poslední rovnost plyne právě z předpokladu 2.7.

2.2 Odhad modelu

Takto konstruovaný model 2.1 s podmínkami 2.7 je nyní zapotřebí odhadnout. Pro odhad vektorů parametrů **a, b, k** nelze použít běžná regresní metoda, protože na pravé straně rovnice 2.1 postrádáme regresory. Je však i tak spoustu jiných možností, jak model odhadnout a některé z nich si v následujících sekcích ukážeme.

2.2.1 SVD a metoda nejmenších čtverců

V literatuře se asi nejčastěji setkáme s metodou singulárního rozkladu SVD (např. Good (1969)). Pomocí zmíněného SVD nalezneme řešení pro odhad metodou nejmenších čtverců. Leeova–Carterova metoda aplikuje tento postup na matici rozdílů logaritmických měr úmrtností **m** a průměrů měr specifických pro jednotlivě věky **A**.

Model budeme odhadovat ve dvou fázích. Nejprve odhadneme vektory parametrů **a, b, k** a následně přeodhadneme vektor parametrů **k**. V každé fázi na konci vždy provedeme normalizační úpravu, aby byly splněny podmínky 2.7.

1. FÁZE

Uvažme, že známe centrální míry úmrtností pro všechny věky $x = x_1, \ldots, x_m$ a pro všechny časy $t = t_1, \ldots, t_n$. Chceme naleznout odhady neznámých vektorů parametrů $\mathbf{a}, \mathbf{b}, \mathbf{k}$, jež budeme značit $\hat{\mathbf{a}}, \hat{\mathbf{b}}, \hat{\mathbf{k}}$. Nalezneme je metodou nejmenších čtverců, neboli chceme minimalizovat

$$O_{LS}(\mathbf{a}, \mathbf{b}, \mathbf{k}) = \sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} [\ln m_{x,t} - a_x - b_x k_t]^2.$$

Odhad $\hat{\mathbf{a}}$ získáme ještě bez využití SVD, stačí nám k tomu pouze zderivovat O_{LS} podle a_x a položit derivaci rovnou nule.

$$\frac{\partial}{\partial a_x} O_{LS}(\mathbf{a}, \mathbf{b}, \mathbf{k}) = -2 \cdot \sum_{t=t_1}^{t_n} [\ln m_{x,t} - a_x - b_x k_t] = 0, \qquad x = x_1, \ldots, x_m,$$

odkud dostáváme, díky 2.7, vektorový tvar

$$\hat{\mathbf{a}} = \frac{1}{n} \sum_{t=t_1}^{t_n} \ln \mathbf{m}_{\cdot,t}. \tag{2.8}$$

Pro získání odhadů $\hat{\mathbf{b}}$ a $\hat{\mathbf{k}}$ využijeme SVD.

Věta 1 (Singulární rozklad matice)**.** *Pro každou matici* $\mathbf{M}$ *typu* $m \times n$ *s hodností* r *existují ortonormální báze* $(\mathbf{v}_1, \ldots \mathbf{v}_n)$ *prostoru* $\mathbb{R}^n$ *a ortonormální báze* $(\mathbf{u}_1, \ldots \mathbf{u}_m)$ *prostoru* $\mathbb{R}^m$ *tak, že*

$$\mathbf{M} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \ldots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T. \tag{2.9}$$

Důkaz. Důkaz nalezneme například ve skriptech Barto a Tůma (2017). □

Ve větě 1 značí $\sigma_1, \ldots, \sigma_r$ druhé odmocniny vlastních čísel matice $\mathbf{M}^T\mathbf{M}$ a platí, že $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$.

Důsledkem této věty je, že pokud chceme matici $\mathbf{M}$ hodnosti r aproximovat maticí nižší hodnosti s ($s < r$) metodou nejmenších čtverců, získáme odhad

$$\hat{\mathbf{M}} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \ldots + \sigma_s \mathbf{u}_s \mathbf{v}_s^T. \tag{2.10}$$

Velmi oblíbené a zároveň často využívané je aproximovat matici $\mathbf{M}$ maticí hodnosti 1:

$$\mathbf{M} \approx \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T. \tag{2.11}$$

Lee a Carter aplikovali SVD na matici rozdílů logaritmických měr úmrtností $\mathbf{m}$ a průměrů měr specifických pro jednotlivě věky $\mathbf{A}$. Tento rozdíl již dokážeme spočítat, proto označme

$$M_{x,t} = \ln m_{x,t} - \hat{a}_x, \qquad x = x_1, \ldots, x_m, \quad t = t_1, \ldots, t_n,$$

prvky matice $\mathbf{M}$.

Nyní potřebujeme získat odhady $\hat{\mathbf{b}}, \hat{\mathbf{k}}$, které opět získáme minimalizací

$$O_{LS}(\mathbf{b}, \mathbf{k}) = \sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} [M_{x,t} - b_x k_t]^2.$$

Výše jsme zmínili, že díky SVD je nejlepší aproximací matice $\mathbf{M}$ metodou nejmenších čtverců matice $\sigma_1 \mathbf{u_1} \mathbf{v}_1^T$. S přihlédnutím na první podmínku v 2.7 získáme odhady

$$\hat{\mathbf{b}} = \frac{\mathbf{u}_1}{\sum_{i=1}^{m} u_{1i}}, \qquad \hat{\mathbf{k}} = \sigma_1 (\sum_{i=1}^{m} u_{1i}) \mathbf{v}_1, \tag{2.12}$$

s předpokladem, že

$$\sum_{i=1}^{m} u_{1i} = 0.$$

Odhady, které jsme nalezli výše uvedenou metodou, však nesplňují normalizační podmínky 2.7. Musíme tedy provést potřebnou úpravu pro jejich splnění. Položme

$$\begin{aligned} a_x^{Fin} &= \hat{a}_x + \hat{b}_x \bar{k}, \qquad x = x_1, \ldots, x_m, \\ k_t^{Fin} &= (\hat{k}_t - \bar{k}) \hat{b}_\bullet, \qquad t = t_1, \ldots, t_n, \\ b_x^{Fin} &= \frac{\hat{b}_x}{\hat{b}_\bullet}, \qquad x = x_1, \ldots, x_m, \end{aligned} \tag{2.13}$$

kde $\hat{b}_\bullet = \sum_{x=x_1}^{x_m} \hat{b}_x$ a $\bar{k} = \frac{1}{n} \sum_{t=t_1}^{t_n} k_t$.

2. FÁZE

Takto nalezené odhady Lee s Carterem ve své práci Lee a Carter (1992) nepovažují za finální a parametr $\mathbf{k}$ předodhadnou. Důvodem, proč se autoři nespokojí s aktuálními odhady, je, že aktuální odhady $\mathbf{k}$ minimalizují chyby od logaritmů úmrtnostních měr, zatímco my bychom preferovali minimalizaci chyb od měr úmrtností. Minimalizace chyb od logaritmů úmrtnostních měr způsobuje, že nízké úmrtnosti mladší části populace dostávají stejnou váhu, jako vysoké úmrtnosti u starších generací, přestože vliv úmrtnosti mladých lidí přispívá do celkové úmrtnosti oproti starším lidem pouze nepatrně.

Pro předodhadnutí parametru $\mathbf{k}$ se vychází ze vztahu

$$m_{x,t} = \frac{D_{x,t}}{N_{x,t}}, \qquad x = x_1, \ldots, x_m, \quad t = t_1, \ldots, t_n. \tag{2.14}$$

Pokud vezmeme místo $m_{x,t}$ jeho odhad $exp(\hat{a}_x + \hat{b}_x \hat{k}_t)$, rovnost bude zatížená chybou. Odhad parametru $\mathbf{k}$ následně získáme z rovnice:

$$\sum_{x=x_1}^{x_m} D_{x,t} = \sum_{x=x_1}^{x_m} N_{x,t} e^{\hat{a}_x + \hat{b}_x k_t}, \qquad t = t_1, \ldots, t_n. \tag{2.15}$$

Analytické řešení této rovnice neexistuje, proto budeme iterativně minimalizovat rozdíl obou stran rovnice 2.15. Lze použít například Newtonův–Raphsonův algoritmus, pro podrobnější popis této metody viz např. Ben-Israel (1966).

Po předodhadnutí parametru $\mathbf{k}$ je ještě potřeba opět provést normalizační úpravu, která je však zcela analogická té v 1. fázi, viz 2.13.

2.2.2 Poissonova metoda s bilineárními prediktory

Další možností, jak odhadnout potřebné parametry, je uvažovat počet úmrtí $D_{x,t}$ jako proměnnou s Poissonovým rozdělením, jejíž střední hodnotu odhadneme nelineárními (bilineárními) parametry.

1. FÁZE

Tato metoda, popsaná v článku Brouhns a kol. (2002) a Alho (2000), vychází z rovnosti:

$$\mathsf{E}(D_{x,t}) = N_{x,t}\exp(\eta_{x,t}), \qquad x = x_1, \ldots, x_m, \quad t = t_1, \ldots, t_n, \tag{2.16}$$

kde

$$\eta_{x,t} = a_x + b_x k_t.$$

Pro odhad parametrů zde však nemůžeme použít Poissonův zobecněný lineární model, protože $\eta_{x,t}$ obsahuje nelineární parametry. Díky Brouhns a kol. (2002) můžeme parametry odhadnout optimalizací Poissonovy věrohodnosti. Algoritmus je založen na konvergenci deviance

$$\mathrm{Dev}(\mathbf{D}, \hat{\mathbf{D}}) = \sum_{t=t_1}^{t_n} \sum_{x=x_1}^{x_m} 2 \left(D_{x,t} \log \frac{D_{x,t}}{\hat{D}_{x,t}} - (D_{x,t} - \hat{D}_{x,t}) \right), \tag{2.17}$$

kde

$$\hat{D}_{x,t} = N_{x,t}\exp(\hat{a}_x + \hat{b}_x\hat{k}_t)$$

a $\mathbf{D}$ resp. $\hat{\mathbf{D}}$ jsou matice o rozměrech $(m \times n)$, jejichž prvky jsou $D_{x,t}$ resp. $\hat{D}_{x,t}$.

V prvním kroku algoritmu nastavíme počáteční hodnoty pro odhady $\hat{\mathbf{a}}$, $\hat{\mathbf{b}}$, $\hat{\mathbf{k}}$, spočítáme $\hat{\mathbf{D}}$ a následně spočítáme $\mathrm{Dev}(\mathbf{D}, \hat{\mathbf{D}})$.

V druhém kroku začneme aktualizovat odhady $\hat{\mathbf{a}}$, $\hat{\mathbf{b}}$, $\hat{\mathbf{k}}$. Nejprve zaktualizujeme $\hat{\mathbf{a}}$:

$$\hat{a}_x^{new} = \hat{a}_x + \frac{\sum_{t=t_1}^{t_n} (D_{x,t} - \hat{D}_{x,t})}{\sum_{t=t_1}^{t_n} \hat{D}_{x,t}}, \qquad x = x_1, \ldots, x_m \tag{2.18}$$

a přepočítáme $\hat{\mathbf{D}}$. Pokračujeme s aktualizací odhadu $\hat{\mathbf{k}}$:

$$\hat{k}_t^{new} = \hat{k}_t + \frac{\sum_{x=x_1}^{x_m} (D_{x,t} - \hat{D}_{x,t})\hat{b}_x}{\sum_{x=x_1}^{x_m} \hat{D}_{x,t}\hat{b}_x^2}, \qquad t = t_1, \ldots, t_n, \tag{2.19}$$

provedeme normalizaci tak, aby $\sum_{t=t_1}^{t_n} \hat{k}_t = 0$ a opět přepočítáme $\hat{\mathbf{D}}$. Zbývá zaktualizovat odhad $\hat{\mathbf{b}}$:

$$\hat{b}_x^{new} = \hat{b}_x + \frac{\sum_{t=t_1}^{t_n} (D_{x,t} - \hat{D}_{x,t})\hat{k}_t}{\sum_{t=t_1}^{t_n} \hat{D}_{x,t}\hat{k}_t^2}, \qquad x = x_1, \ldots, x_m, \tag{2.20}$$

znovu přepočítáme $\hat{\mathbf{D}}$ a rovněž spočítáme Dev($\mathbf{D}, \hat{\mathbf{D}}$). Vzorce pro aktualizace parametrů vychází z Newtonova–Raphsonova iteračního algoritmu, který je použit na soustavu rovnic, již jsme získali položením parciálních derivací logaritmické věrohodnostní funkce rovné nule.

Ve třetím kroku algoritmus končí, pokud Dev($\mathbf{D}, \hat{\mathbf{D}}$) konverguje, v opačném případě se vracíme do kroku 2. To, zda-li Dev($\mathbf{D}, \hat{\mathbf{D}}$) konverguje, rozhodneme podle toho, zda úbytek v devianci nepřekročil předem zvolenou hodnotu. V opačném případě, kdy úbytek deviance překročí předem zvolenou hodnotu, odhady nejsou finální a vracíme se do kroku 2.

Na závěr je třeba zkontrolovat, zda odhady $\hat{\mathbf{b}}$ a $\hat{\mathbf{k}}$ splňují normalizační podmínky 2.7. V případě, že tomu tak není, je třeba provést úpravu 2.13.

Jako startovací hodnoty pro odhady vektorů parametrů $\hat{\mathbf{a}}$, $\hat{\mathbf{b}}$ a $\hat{\mathbf{k}}$ se často používají SVD odhady. Díky těmto odhadům algoritmus typicky konverguje po pár iteracích.

2. FÁZE

Druhá fáze odhadování je zcela analogická jako v případě SVD a metody nejmenších čtverců.

2.2.3 Alternativní (regresní) metoda

Další možnost, jak odhadnout parametry, navrhli Lee s Carterem ve své práci Lee a Carter (1992). Opět uvažujme, že známe centrální míry úmrtností pro všechny věky $x = x_1, \ldots, x_m$ a pro všechny časy $t = t_1, \ldots, t_n$. Jak jsme zmínili výše, z restrikcí 2.7 opět plyne, že vektor parametrů $\mathbf{a}$ odpovídá průměru logaritmických měr $\mathbf{m}$ přes čas t.

Vyjdeme-li nyní z rovnice 2.1, kde za náhodnou veličinu $\varepsilon_{x,t}$ bereme její střední hodnotu, získáme odhad pro vektor parametrů $\mathbf{k}$:

$$\begin{aligned} \ln \mathbf{m} &= \mathbf{A} + \mathbf{b}\mathbf{k}^T + \varepsilon_{x,t} \\ \sum_{x=x_1}^{x_m} \ln \mathbf{m}_{x,\cdot} &= \sum_{x=x_1}^{x_m} \dot{\mathbf{a}}_x + \sum_{x=x_1}^{x_m} b_x \mathbf{k}^T \\ \hat{\mathbf{k}}^T &= \sum_{x=x_1}^{x_m} (\ln \mathbf{m}_{x,\cdot} - \dot{\mathbf{a}}_{\mathbf{x}}), \end{aligned} \tag{2.21}$$

kde $\dot{\mathbf{a}}_x = (a_x, \ldots, a_x)$ je vektor o velikosti n, jehož všechny složky jsou shodné prvky a_x. V poslední rovnosti jsme využili prvního předpokladu 2.7.

Zbývá odhadnout vektor $\mathbf{b}$, který však již zvládneme odhadnout za pomoci lineární regrese. Platí, že pokud máme soustavu rovnic ve tvaru

$$\mathsf{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}, \tag{2.22}$$

kde $\mathbf{Y}$ je vektor velikosti n pozorování, $\mathbf{X}$ je matice $n \times k$ známých hodnot a $\boldsymbol{\beta}$ je vektor velikosti k neznámých parametrů, který chceme odhadnout, pak odhad metodou nejmenších čtverců vyjádříme jako

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{2.23}$$

Aplikujeme lineární regresi na náš případ. Pro každé $x = x_1, \ldots, x_m$ minimalizujeme čtvercovou chybu rovnice

$$\ln \mathbf{m} - \mathbf{A} = \mathbf{b}\mathbf{k}^T.$$

Pro pevné x označme $y_t = \ln m_{x,t} - a_x$, vektor těchto y_t označme $\mathbf{Y}$. Vektor $\mathbf{k}$ uvažujeme jako matici $\mathbf{X}$ a b_x koresponduje v této notaci s β.

Z lineární regrese, konkrétně z 2.23 pro regresní přímku, dostáváme, že:

$$\hat{\mathbf{b}} = \frac{\sum_{t=t_1}^{t_n} k_t(\ln \mathbf{m}_{\cdot,t} - \mathbf{a})}{\sum_{t=t_1}^{t_n} k_t^2}. \tag{2.24}$$

2.3 Predikce

Když máme model odhadnutý, budeme chtít pomocí něj predikovat míry úmrtnosti i v budoucnu. Podíváme-li se na náš model 2.1, zjistíme, že jediný parametr, který závisí na čase t, je $\mathbf{k}$.

Naším úkolem je tedy predikovat časovou řadu $\mathbf{k}$, což budeme provádět Boxovou–Jenkinsovou metodologií (Box a Jenkins (1970)). Pomocí této metodologie vybereme vhodný ARIMA model řádu (p, d, q). Následně snadno dopočteme míry úmrtnosti i v budoucnosti.

Nejčastěji se nejlepším modelem jeví *náhodná procházka s driftem* neboli model ARIMA$(0, 1, 0)$. V tomto případě zbývá odhadnout pouze parametr α, kde se může použít odhad

$$\hat{\alpha} = \frac{k_{t_n} - k_{t_1}}{t_n - t_1}, \tag{2.25}$$

což plyne přímo z definice modelu a odhadu střední hodnoty $\mathsf{E}(k_t - k_{t-1})$ průměrem vypočtených hodnot $k_t - k_{t-1}$:

$$\begin{aligned} k_t - k_{t-1} &= \alpha + \varepsilon_t \\ \mathsf{E}(k_t - k_{t-1}) &= \mathsf{E}(\alpha + \varepsilon_t) \\ \hat{\alpha} &= \frac{1}{t_n - t_1} \sum_{t=t_2}^{t_n} k_t - k_{t-1} \\ \hat{\alpha} &= \frac{k_{t_n} - k_{t_1}}{t_n - t_1}. \end{aligned}$$

Nicméně tuto úvahu nelze převzít, vždy je potřeba nejprve provést vlastní analýzu řady a na základě této analýzy vybrat vhodný model.

V případě modelu ARIMA(0, 1, 0) dopočteme budoucí hodnoty $\tilde{k}_{t_{n+s}}$ následovně:

$$\begin{aligned} \tilde{k}_{t_{n+1}} &= \hat{k}_{t_n} + \hat{\alpha}, \\ \tilde{k}_{t_{n+s}} &= \tilde{k}_{t_{n+s-1}} + \hat{\alpha}, \qquad s = 2, \ldots, S, \\ \tilde{k}_{t_{n+s}} &= \hat{k}_{t_n} + s\hat{\alpha}, \qquad s = 1, \ldots, S, \end{aligned} \tag{2.26}$$

kde S je maximální počet roků, pro které chceme nalézt budoucí hodnoty.

Jakmile máme predikovaný parametr $\tilde{k}_t$, snadno dopočteme budoucí hodnoty $\tilde{m}_{x,t_{n+s}}$ podle našeho modelu 2.1:

$$\tilde{m}_{x,t_{n+s}} = e^{\hat{a}_x + \hat{b}_x \tilde{k}_{t_{n+s}}}, \qquad x = x_1, \ldots, x_m, \quad s = 1, \ldots, S. \tag{2.27}$$

Vztah 2.27 lze díky 2.26 upravit do podoby:

$$\begin{aligned} \tilde{m}_{x,t_{n+s}} &= e^{\hat{a}_x + \hat{b}_x \tilde{k}_{t_{n+s}}} \\ &= e^{\hat{a}_x + \hat{b}_x (\hat{k}_{t_n} + s\hat{\alpha})} \\ &= \hat{m}_{x,t_n} e^{\hat{b}_x s\hat{\alpha}} \\ &= \hat{m}_{x,t_n} e^{\hat{b}_x (\tilde{k}_{t_{n+s}} - \hat{k}_{t_n})}. \end{aligned}$$

Když máme odhadnuté budoucí míry úmrtnosti, jsme už pouze krok od dopočtení budoucích tabulkových počtů zemřelých. Nechť platí předpoklad o konstantní intenzitě úmrtnosti 1.1. Pak můžeme vyjádřit tabulkový počet dožívajících se do věku x v čase t rekurzivně následovně:

$$\tilde{l}_{x+1,t} = \tilde{l}_{x,t} e^{-\tilde{m}_{x,t}}, \qquad x = x_1, \ldots x_{m-1}, \quad t = t_{n+1}, \ldots, t_{n+S}. \tag{2.28}$$

Tabulkový počet zemřelých mezi roky x a $x+1$ pak dostaneme jako

$$\tilde{d}_{x,t} = \tilde{l}_{x,t} - \tilde{l}_{x+1,t}, \qquad x = x_1, \ldots x_{m-1}, \quad t = t_{n+1}, \ldots, t_{n+S}. \tag{2.29}$$

3. Renshawův–Habermanův model

V této kapitole si představíme další známý demografický model, který navazuje na Leeův–Carterův model. Jedná se o metodu, kterou popsali ve svém článku Renshaw a Haberman (2003) Artur E. Renshaw a Steven Haberman. Tito matematici vycházeli z Leeova–Carterova modelu (LC) a při analýze dat z let 1950-1998 pro úmrtnostní míry mužů v Anglii a Walesu si povšimli nedostatků LC modelu ve věkové skupině 20-39 let, kde výrazně stoupl počet sebevražd v závislosti na viru HIV a nemoci AIDS. Myšlenka tedy byla rozšířit LC model o druhou sadu singulárních vektorů a pokusit se tak zachytit různé vlivy úmrtí různých věkových skupin, které nejsou brány v potaz v LC modelu, ale způsobují výrazné rozdíly napříč populací. Model lze zapsat následovně s tím, že uvažujeme analogické značení jako v LC modelu:

$$\ln \mathbf{m} = \mathbf{A} + \mathbf{b}^{(1)}(\mathbf{k}^{(1)})^T + \mathbf{b}^{(2)}(\mathbf{k}^{(2)})^T + \quad , \tag{3.1}$$

nebo ekvivalentně ve tvaru exponenciály:

$$\mathbf{m} = \exp(\mathbf{A} + \mathbf{b}^{(1)}(\mathbf{k}^{(1)})^T + \mathbf{b}^{(2)}(\mathbf{k}^{(2)})^T + \quad). \tag{3.2}$$

Tento model je také podurčený jako při LC modelování. Je třeba přidat normalizační podmínky, které budou analogické rovnostem 2.7:

$$\sum_{x=x_1}^{x_n} b_x^{(i)} = 1, \qquad \sum_{t=t_1}^{t_n} k_t^{(i)} = 0, \quad i = 1, 2. \tag{3.3}$$

Zatímco konstrukce modelu je přímočará a snadná, možný problém může nastat při predikcích budoucích hodnot. V této kapitole si popíšeme dva různé způsoby odhadů parametrů a predikcí. První je založena na SVD metodě. Druhá metoda, jež byla představena v článku Brouhns a kol. (2002), využívá heteroskedastickou Poissonovu strukturu. Počet úmrtí je zde modelován v závislosti na bilineárních prediktorech tak, aby struktura přesně odpovídala LC modelu. Pro odhad modelů nelze použít zobecněný lineární model, což Brouhns a spol. řeší optimalizací Poissonovy věrohodnostní funkce.

3.1 Odhad modelu

Stejně tak, jako tomu bylo v LC modelu, potřebujeme nyní odhadnout vektory parametrů $\mathbf{a}, \mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \mathbf{k}^{(1)}, \mathbf{k}^{(2)}$.

3.1.1 SVD a metoda nejmenších čtverců

Zde se postupuje analogicky jako při odhadování parametrů LC modelu. Nejprve nalezneme odhad vektoru $\mathbf{a}$ metodou nejmenších čtverců. Minimalizujeme

$$O_{LS}(\mathbf{a}, \mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \mathbf{k}^{(1)}, \mathbf{k}^{(2)}) = \sum_{x=x_1}^{x_n} \sum_{t=t_1}^{t_n} [\ln m_{x,t} - a_x - b_x^{(1)} k_t^{(1)} - b_x^{(2)} k_t^{(2)}]^2. \tag{3.4}$$

Odhad $\hat{\mathbf{a}}$ získáme zderivováním 3.4 a položením této derivace rovné nule:

$$\frac{\partial}{\partial a_x} O_{LS}(\mathbf{a}, \mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \mathbf{k}^{(1)}, \mathbf{k}^{(2)}) = -2 \cdot \sum_{t=t_1}^{t_n} [\ln m_{x,t} - a_x - b_x^{(1)} k_t^{(1)} - b_x^{(2)} k_t^{(2)}] = 0,$$

pro $x = x_1, \ldots, x_m$, z čehož díky druhé podmínce 3.3 plyne

$$\hat{\mathbf{a}} = \frac{1}{n} \sum_{t=t_1}^{t_n} \ln \mathbf{m}_{\cdot,t}. \tag{3.5}$$

Dále potřebujeme odhadnout vektory parametrů $\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \mathbf{k}^{(1)}, \mathbf{k}^{(2)}$, k čemuž využijeme SVD, konkrétně větu 1. Tuto větu aplikujeme na matici $\mathbf{M}$ - matice rozdílů logaritmických měr úmrtností $\mathbf{m}$ a průměrů měr specifických pro jednotlivé věky $\mathbf{A}$ stejně, jako při LC metodě pouze s rozdílem, že matici $\mathbf{M}$ aproximujeme maticí řádu 2:

$$\mathbf{M} \approx \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T. \tag{3.6}$$

Chceme získat potřebné odhady metodou nejmenších čtverců:

$$O_{LS}(\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \mathbf{k}^{(1)}, \mathbf{k}^{(2)}) = \sum_{x=x_1}^{x_n} \sum_{t=t_1}^{t_n} [M_{x,t} - b_x^{(1)} k_t^{(1)} - b_x^{(2)} k_t^{(2)}]^2.$$

Z věty 1 víme, že odhad matice $\mathbf{M}$ metodou nejmenších čtverců maticí řádu 2 je 3.6, a proto

$$\sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T = \mathbf{b}^{(1)} (\mathbf{k}^{(1)})^T + \mathbf{b}^{(2)} (\mathbf{k}^{(2)})^T. \tag{3.7}$$

Společně s první podmínkou z 3.3 dostáváme

$$\hat{\mathbf{b}}^{(i)} = \frac{\mathbf{u}_i}{\sum_{k=1}^{m} u_{ik}}, \qquad \hat{\mathbf{k}}^{(i)} = \sigma_i \Big(\sum_{k=1}^{m} u_{ik}\Big) \mathbf{v}_i, \qquad i = 1, 2, \tag{3.8}$$

s předpokladem, že

$$\sum_{k=1}^{m} u_{ik} = 0, \qquad i = 1, 2.$$

Získané odhady nesplňují normalizační podmínky 3.3, proto je třeba provést analogické úpravy jako v 2.13. Zbývá přeodhadnout vektor parametrů $\mathbf{k}^{(1)}$, což se provádí analogicky jako v LC modelu. Vycházíme ze vztahu 2.14. Pokud vezmeme místo $m_{x,t}$ jeho odhad $exp(\hat{a}_x + \hat{b}_x^{(1)} (\hat{k}_t^{(1)})^T + \hat{b}_x^{(2)} (\hat{k}_t^{(2)})^T)$, rovnost bude zatížená chybou. Odhad vektoru parametrů $\mathbf{k}^{(1)}$ následně získáme z rovnice:

$$\sum_{x=x_1}^{x_n} D_{x,t} = \sum_{x=x_1}^{x_n} N_{x,t} e^{\hat{a}_x + \hat{b}_x^{(1)} k_t^{(1)} + \hat{b}_x^{(2)} \hat{k}_t^{(2)}}, \qquad t = t_1, \ldots, t_n. \tag{3.9}$$

Analytické řešení této rovnice neexistuje, proto se k problému musí přistoupit iterativně, nabízí se například Newtonův–Raphsonův algoritmus.

Po přeodhadnutí vektoru parametrů $\mathbf{k}^{(1)}$ je ještě potřeba opět provést analogickou normalizační úpravu té z LC modelu.

3.1.2 Poissonova metoda s bilineárními prediktory

Tato metoda byla popsána v LC modelu (2.2.2) a zde se postupuje analogicky. Vychází se z rovnosti 2.16, kde

$$\eta_{x,t} = a_x + b_x^{(1)} k_t^{(1)} + b_x^{(2)} k_t^{(2)}.$$

Parametry odhadneme rovněž optimalizací Poissonovy věrohodnosti v závislosti na konvergenci deviance 2.17, kde

$$\hat{D}_{x,t} = N_{x,t} \exp(\hat{a}_x + \hat{b}_x^{(1)} \hat{k}_t^{(1)} + \hat{b}_x^{(2)} \hat{k}_t^{(2)}). \tag{3.10}$$

V prvním kroku algoritmu nastavíme počáteční hodnoty pro odhady $\hat{\mathbf{a}}$, $\hat{\mathbf{b}}^{(1)}$, $\hat{\mathbf{b}}^{(2)}$, $\hat{\mathbf{k}}^{(1)}$, $\hat{\mathbf{k}}^{(2)}$, spočítáme $\hat{\mathbf{D}}$ a následně spočítáme Dev$(\mathbf{D}, \hat{\mathbf{D}})$.

V druhém kroku začneme aktualizovat odhady. Nejprve zaktualizujeme $\hat{\mathbf{a}}$. To provedeme zcela identicky jako v 2.18 a přepočítáme $\hat{\mathbf{D}}$. Pokračujeme s aktualizací odhadu $\hat{\mathbf{k}}^{(1)}$:

$$(\hat{k}_t^{(i)})^{new} = \hat{k}_t^{(i)} + \frac{\sum_{x=x_1}^{x_m} (D_{x,t} - \hat{D}_{x,t}) \hat{b}_x^{(i)}}{\sum_{x=x_1}^{x_m} \hat{D}_{x,t} (\hat{b}_x^{(i)})^2}, \qquad i = 1, 2, \quad t = t_1, \ldots, t_n, \tag{3.11}$$

provedeme normalizaci tak, aby $\sum_{t=t_1}^{t_n} \hat{k}_t^{(i)} = 0$ a opět přepočítáme $\hat{\mathbf{D}}$. Dále zaktualizujeme $\hat{\mathbf{b}}^{(1)}$:

$$(\hat{b}_x^{(i)})^{new} = \hat{b}_x^{(i)} + \frac{\sum_{t=t_1}^{t_n} (D_{x,t} - \hat{D}_{x,t}) \hat{k}_t^{(i)}}{\sum_{t=t_1}^{t_n} \hat{D}_{x,t} (\hat{k}_t^{(i)})^2}, \qquad i = 1, 2, \quad x = x_1, \ldots, x_m, \tag{3.12}$$

znovu přepočteme $\hat{\mathbf{D}}$. Nyní aktualizujeme $\hat{\mathbf{k}}^{(2)}$ podle 3.11, provedeme normalizaci a přepočítáme $\hat{\mathbf{D}}$. Na závěr zaktualizujeme $\hat{\mathbf{b}}^{(2)}$ podle 3.12, přepočítáme $\hat{\mathbf{D}}$ a Dev$(\mathbf{D}, \hat{\mathbf{D}})$. Vzorce pro aktualizace parametrů vychází opět z iteračního Newtonova–Raphsonova algoritmu, který je použit na soustavu rovnic, jejž jsme získali položením parciálních derivací logaritmické věrohodnostní funkce rovné nule.

Ve třetím kroku algoritmus končí, pokud Dev$(\mathbf{D}, \hat{\mathbf{D}})$ konverguje, v opačném případě se vracíme do kroku 2. Kritérium pro konvergenci je stejné jako v sekci pro LC metodu 2.2.2.

Na závěr je třeba zkontrolovat, zda odhady $\hat{\mathbf{b}}^{(i)}$ a $\hat{\mathbf{k}}^{(i)}$ pro $i = 1{,}2$ splňují normalizační podmínky 3.3. V případě, že tomu tak není, je třeba provést analogické úpravy jako ve 2.13.

Zbývá předodhadnout parametr $\hat{\mathbf{k}}^{(1)}$. Způsob, jakým se to provádí, je totožný jako při SVD a metodě nejmenších čtverců (3.1.1).

3.2 Predikce

Po odhadnutí parametrů modelu přichází na řadu predikce budoucích hodnot. Postup je zde opět analogický jako v případě predikce v LC modelu. Při pohledu na náš model 3.2 zjišťujeme, že na čase závisí pouze $\mathbf{k^{(1)}}$ a $\mathbf{k^{(2)}}$. Jakmile odhadneme tyto časové řady, pak už snadno dopočteme míry úmrtnosti. Renshaw s Habermanem ve své práci Renshaw a Haberman (2003) navrhují uvažovat $\mathbf{k^{(1)}}$, $\mathbf{k^{(2)}}$ jako dvě oddělené nezávislé časové řady, které odhadneme pomocí Boxovy-Jenkinsovy metodologie (Box a Jenkins (1970)). Pro každou časovou řadu $\mathbf{k^{(i)}}$, $i = 1, 2$ postupujeme zcela analogicky jako v LC modelu (2.3).

V případě, že uvažujeme často nejvhodnější model ARIMA(0, 1, 0), což je ale vždy třeba otestovat, získáme odhady

$$\begin{aligned}
\tilde{k}^{(i)}_{t_{n+1}} &= \hat{k}_{t_n} + \hat{\alpha} \qquad i = 1, 2, \\
\tilde{k}^{(i)}_{t_{n+s}} &= \tilde{k}_{t_{n+s-1}} + \hat{\alpha} \qquad i = 1, 2, \quad s = 2, \ldots, S \\
\tilde{k}^{(i)}_{t_{n+s}} &= \hat{k}_{t_n} + s\hat{\alpha}, \qquad i = 1, 2, \quad s = 1, \ldots, S.
\end{aligned} \tag{3.13}$$

Z 3.13 snadno dopočteme $\tilde{m}_{x,t_{n+s}}$:

$$\tilde{m}_{x,t_{n+s}} = exp(\hat{a}_x + \hat{b}^{(1)}_x \tilde{k}^{(1)}_{t_{n+s}} + \hat{b}^{(2)}_x \tilde{k}^{(2)}_{t_{n+s}}), \qquad x = x_1, \ldots, x_m, \quad s = 1, \ldots S, \tag{3.14}$$

což lze ekvivalentně zapsat jako

$$\tilde{m}_{x,t_{n+s}} = \hat{m}_{x,t_n} exp(\sum_{i=1}^{2} \hat{b}^{(i)}_x (\tilde{k}^{(i)}_{t_{n+s}} - \hat{k}^{(i)}_{t_n})), \qquad x = x_1, \ldots, x_m, \quad s = 1, \ldots S.$$

Na závěr dopočteme tabulkový počet zemřelých mezi roky x a $x+1$ analogicky jako v Leeově–Carterově modelu (viz 2.28 a 2.29).

4. Analýza kompozičních dat

Poslední metoda, kterou v této práci představíme, byla publikována v roce 2019 v díle Shang a Haberman (2019) matematiků Han Lin Shanga a Stevena Habermana. Hlavní změna oproti předchozím metodám spočívá v tom, že analyzujeme, modelujeme a předpovídáme tabulkový počet zemřelých závislý na roce a věku $d_{x,t}$, zatímco v Leeově–Carterově či Renshawově–Habermanově metodě analyzujeme centrální míry úmrtnosti $m_{x,t}$. Toto není velký zásah do struktury metod, neboť pokud známe tabulkové počty zemřelých, můžeme z nich vypočítat míry úmrtnosti a naopak, viz vztahy v kapitole 1.

Metoda popsaná v Shang a Haberman (2019) je postavena na kompoziční analýze dat (CoDa), která se aplikuje na kompoziční data.

Definice 1. *Kompoziční data jsou definována jako vektor* $\mathbf{d} = (d_1, \ldots, d_k)$ *velikosti* k*, jehož všechny složky jsou kladné a jejich součet je roven konstantě, neboli:*

$$d_i > 0 \qquad i = 1, \ldots, k, \qquad \sum_{i=1}^{k} d_i = c. \tag{4.1}$$

Konstanta c se obvykle nastavuje na 1, 100 000 či 1 000 000, my budeme v našem případě klást $c = 100\ 000$, jinými slovy budeme předpokádat, že každý rok se narodí 100 000 jedinců, což nejlépe koresponduje s vývojem populace v České republice, viz Český statistický úřad (2023).

Jim Oeppen (Oeppen a kol. (2008)) popsal způsob, jak aplikovat CoDa přístup na předpovídání počtu úmrtí. Začínáme tím, že vytvoříme matici počtů úmrtí $\mathbf{D}$. Poznamenejme, že v díle Shang a Haberman (2019) autoři používají indexaci matice transponovaně. My, v rámci jednotnosti s předchozími sekcemi, pokračujeme v námi zavedené indexaci. Následně vycentrujeme sloupce této matice a transformujeme data pomocí logaritmu. Tento krok je třeba udělat, protože chceme transformovat data do prostoru reálných čísel, zatímco prozatímní data jsou omezená - nabývají hodnot od 0 do c. Nabízí se samozřejmě řada možných transformací, mezi nimiž se nejhojněji využívají transformace logaritmické.

Dále přichází na řadu dekompozice takto upravené matice dat pomocí metody SVD, aproximace matice maticí nižšího řádu a odhad parametrů. Když máme odhadnuté parametry, potřebujeme použít inverzní logaritmickou transformaci a následně pak inverzní operaci k centrování, čímž se dostaneme zpět k našim kompozičním datům.

4.1 Model a odhad parametrů

V této sekci detailněji popíšeme, jak metoda popsaná v Shang a Haberman (2019) funguje. Algoritmus můžeme rozdělit do následujících sedmi bodů:

1. Nejprve vytvoříme matici počtů úmrtí $\mathbf{D}$, o rozměru $(m \times n)$. Klademe zde předpoklad, že každý rok se narodí 100 000 lidí, tedy součet hodnot v každém sloupci matice je roven této konstantě. Prvky matice $\mathbf{D}$ tak odpovídají tabulkovému počtu zemřelých $d_{x,t}$.

2. Dále je třeba standardizovat parametry $d_{x,t}$. Definujme geometrický průměr počtů úmrtí přes čas t pro zvolený věk x. Můžeme zapsat jako:

$$\alpha_x = \exp\left(\frac{1}{n}\sum_{t=t_1}^{t_n} \ln d_{x,t}\right), \tag{4.2}$$

pro $x = x_1, \ldots x_m$. Vezměme nyní sloupec matice $\mathbf{D}$ a vydělme jeho prvky geometrickým průměrem definovaným v 4.2. Takto upravené prvky standardizujeme pomocí celkového součtu sloupců, značíme $f_{x,t}$:

$$f_{x,t} = \frac{d_{x,t}/\alpha_x}{\sum_{x=x_1}^{x_m}(d_{x,t}/\alpha_x)}, \qquad x = x_1, \ldots, x_m, \quad t = t_1, \ldots, t_n. \tag{4.3}$$

Takto vytvořené prvky uspořádáme do matice $\mathbf{F}$. Tím jsme zaručili, že součet prvků v jednotlivých sloupcích matice je vždy roven 1.

3. V tomto kroku chceme transformovat data do prostoru reálných čísel. Na návrh Johna Aitchisona v dílech Aitchison (1982) a Aitchison (1986) využíváme centrovanou log-poměrovou transformaci (centred log-ratio transformation). Nejdříve spočítáme geometrický průměr upravených dat $f_{x,t}$ přes věk pro pevný čas t :

$$g_t = \exp\left(\frac{1}{m}\sum_{x=x_1}^{x_m} \ln f_{x,t}\right), \qquad t = t_1, \ldots, t_n. \tag{4.4}$$

Nyní můžeme přistoupit k samotné transformaci, která má následující podobu:

$$z_{x,t} = \ln\left[\frac{f_{x,t}}{g_t}\right], \qquad x = x_1, \ldots, x_m, \quad t = t_1, \ldots, t_n. \tag{4.5}$$

Takto transformovaná data $z_{x,t}$ můžeme znovu uskupit do matice $\mathbf{Z}$. Dosáhli jsme toho, že prvky této matice jsou prvky prostoru reálných čísel, což jsme požadovali.

4. Na matici $\mathbf{Z}$ aplikujeme analýzu hlavních komponent. Rozložíme prvky matice na součet produktů odhadnutých hlavních komponent $\varphi_{x,l}$ a jejich odhadnutých faktorových skórů $\beta_{l,t}$:

$$z_{x,t} = \sum_{l=1}^{L} \beta_{l,t}\varphi_{x,l} + \varepsilon_{x,t}, \qquad x = x_1, \ldots, x_m, \quad t = t_1, \ldots, t_n, \tag{4.6}$$

kde L značí počet hlavních komponent rozkladu a chybový člen modelu $\varepsilon_{x,t}$ pro věk x v čase t je bílý šum. Díky tomuto rozkladu dokážeme snížit dimenzi dat a pomocí singulárního rozkladu, popsaného ve větě 1, odhadneme hlavní komponenty $\hat{\varphi}_{x,l}$ a faktorové skóry $\hat{\beta}_{l,t}$:

$$\hat{\beta}_l = \mathbf{u}_l \qquad \hat{\varphi}_l = \sigma_l \mathbf{v}_l, \qquad l = 1, \ldots, L, \tag{4.7}$$

kde $\hat{\beta}_l$ je vektor o velikosti n a $\hat{\varphi}_l$ je vektor o velikosti m.

Pomocí rozkladu 4.6 budeme zároveň schopni předpovídat budoucí hodnoty $z_{x,t}$. Otázkou však zůstává, kolik hlavních komponent a faktorových

skórů pro odhad matice $\mathbf{Z}$ použít. Protože hlavní komponenty zachycují množství variability v datech a jsou uspořádány podle velikosti vysvětlované variability, nabízí se určit hladinu rozptylu, kterou chceme mít hlavními komponentami vysvětlenou, a vybrat minimální množství komponent, které tuto hladinu pokryjí. Horváth a Kokoszka (2012) navrhují zvolit hladinu $\delta = 0.85$. Počet komponent L můžeme určit kritériem:

$$L = \arg\min_{L:L\geq 1} \left[\sum_{l=1}^{L} \lambda_l \Big/ \sum_{l=1}^{r} \lambda_l \geq \delta \right], \tag{4.8}$$

kde λ_i značí i-té největší vlastní číslo matice $\mathbf{Z}^T\mathbf{Z}$, a číslo r značí celkový počet vlastních čísel.

5. Pokud máme zparametrizovaná transformovaná data a parametry odhadnuty, můžeme predikovat budoucí hodnoty. Stejně tak, jako tomu bylo v Leeově–Carterově či Renshawově–Habermanově modelu, tak i zde závisí na čase pouze jeden typ parametrů, a to faktorová skóre $\hat{\beta}_{l,t}$. Tyto parametry budeme predikovat jako jednorozměrné časové řady, neboť odhady parametrů $\hat{\beta}_{l,t}$ a $\hat{\beta}_{k,t}$ jsou nekorelované pro $k = l$ (viz Hyndman a Ullah (2007)). Existuje několik možností, jak predikovat jednorozměrné časové řady, autoři článku Shang a Haberman (2019) navrhují exponenciální vyhlazování (ETS). Detailněji se budeme způsoby predikce zabývat v následující sekci 4.2.

   Předpokládejme, že známe hodnoty všech parametrů do času t_n a budoucí hodnoty $\tilde{\beta}_{l,t_n+h}$, kde $h > 0$ udává na kolik let dopředu predikujeme. Budoucí hodnoty $\tilde{z}_{l,t_n+h}$ získáme vztahem:

$$\tilde{z}_{x,t_n+h} = \sum_{l=1}^{L} \tilde{\beta}_{l,t_n+h}\hat{\varphi}_{x,l}, \qquad x = x_1, \ldots, x_m, \quad h = 1, \ldots, H, \tag{4.9}$$

   kde H je horizont, pro který chceme predikovat budoucí hodnoty.

6. Když máme odhadnuté budoucí hodnoty $\tilde{z}_{x,t_n+h}$, je třeba transformovat je zpět, abychom získali budoucí hodnoty tabulkových počtů zemřelých $\tilde{d}_{x,t_n+h}$. Nejprve provedeme zpětnou centrovanou log-poměrovou transformaci, z které získáme parametry $\tilde{f}_{x,t_n+h}$. Platí:

$$\tilde{f}_{x,t_n+h} = \left[\frac{\exp(\tilde{z}_{x,t_n+h})}{\sum_{x=x_1}^{x_m} \exp(\tilde{z}_{x,t_n+h})} \right], \qquad x = x_1, \ldots, x_m, \quad h = 1, \ldots, H, \tag{4.10}$$

   neboť díky vztahům z bodů 2 a 3 z této sekce plyne:

$$\begin{aligned}
\left[\frac{\exp(z_{x,t})}{\sum_{x=x_1}^{x_m} \exp(z_{x,t})} \right] &= \left[\frac{\exp(\ln \frac{f_{x,t}}{g_t})}{\sum_{x=x_1}^{x_m} \exp(\ln \frac{f_{x,t}}{g_t})} \right] \\
&= \frac{\frac{f_{x,t}}{g_t}}{\frac{1}{g_t} \left[\sum_{x=x_1}^{x_m} f_{x,t} \right]} \\
&= \left[\frac{f_{x,t}}{\sum_{x=x_1}^{x_m} f_{x,t}} \right] = f_{x,t}.
\end{aligned}$$

Poslední rovnost platí díky vztahu

$$\sum_{x=x_1}^{x_m} f_{x,t} = \sum_{x=x_1}^{x_m} \frac{d_{x,t}/\alpha_x}{\sum_{x=x_1}^{x_m} d_{x,t}/\alpha_x} = \frac{\sum_{x=x_1}^{x_m} d_{x,t}/\alpha_x}{\sum_{x=x_1}^{x_m} d_{x,t}/\alpha_x} = 1.$$

7. Z parametrů $\tilde{f}_{x,t_n+h}$ už pouhou jednou úpravou dostaneme požadované budoucí hodnoty tabulkových počtů zemřelých $\tilde{d}_{x,t_n+h}$. Platí:

$$\tilde{d}_{x,t_n+h} = \frac{\tilde{f}_{x,t_n+h}\alpha_x}{\sum_{x=x_1}^{x_m} \tilde{f}_{x,t_n+h}\alpha_x}, \qquad x = x_1, \ldots, x_m, \quad h = 1, \ldots, H, \tag{4.11}$$

protože:

$$\begin{aligned}
\frac{f_{x,t}\alpha_x}{\sum_{x=x_1}^{x_m} f_{x,t}\alpha_x} &= \frac{\frac{d_{x,t}/\alpha_x}{\sum_{x=x_1}^{x_m} d_{x,t}/\alpha_x}\alpha_x}{\sum_{x=x_1}^{x_m} \frac{d_{x,t}/\alpha_x}{\sum_{x=x_1}^{x_m} d_{x,t}/\alpha_x}\alpha_x} \\
&= \frac{\frac{d_{x,t}}{\sum_{x=x_1}^{x_m} d_{x,t}/\alpha_x}}{\frac{1}{\sum_{x=x_1}^{x_m} d_{x,t}/\alpha_x}\sum_{x=x_1}^{x_m} d_{x,t}} \\
&= \frac{d_{x,t}}{\sum_{x=x_1}^{x_m} d_{x,t}} = \frac{d_{x,t}}{100000}.
\end{aligned}$$

Poslední rovnost platí díky předpokladu, který jsme vyslovili v bodu 1 této sekce. Je tedy zřejmé, že finální odhady tabulkových počtů zemřelých budeme ještě muset vynásobit námi zvolenou konstantou $c = 100\ 000$.

4.2 Predikce

V předchozí sekci jsme popsali algoritmus a zparametrizovali jsme hlavní model. Zbývá stále více prodiskutovat možnosti predikcí faktorových skórů $\tilde{\beta}_{l,t_n+h}$, parametrů $\tilde{z}_{x,t_n+h}$ a taktéž intervalové predikce.

4.2.1 Bodová předpověď

Chceme se zabývat bodovou predikcí parametrů $\hat{z}_{x,t}$, které jsme v předchozí sekci rozložili na sumu produktů hlavních komponent $\hat{\varphi}_{x,l}$ a jejich faktorových skórů $\hat{\beta}_{l,t}$. Na čase však závisí pouze faktorové skóry $\hat{\beta}_{l,t}$. Pro pevné l máme jednorozměrnou časovou řadu, kterou můžeme predikovat pomocí Boxovy–Jenkinsovy metodologie Box a Jenkins (1970) nalezením vhodného ARIMA(p, d, q) modelu, popsané také v 2.3.

Další možností je využít ETS modely, náhodnou procházku s driftem nebo náhodnou procházku bez driftu.

Exponenciální vyhlazování

Exponenciální vyhlazování, též ETS (Hyndman a Khandakar (2008)), je metoda, která se používá k predikci časových řad. Hlavní myšlenka je taková, že odhadovaná hodnota je modelována jako lineární kombinace pozorované hodnoty

a p°edchozí odhadované hodnoty. Tím m·oeme docílit toho, oe nedávné hodnoty zasahují do odhadu více neo hodnoty získané d°íve v minulosti.

Zkratka ETS vychází ze t°í slov - chyba (error), trend a sezónnost. Exponenciální vyhlazování totio modeluje tyto t°i veli£iny a máme n¥kolik mooností, jak to m·oeme provést. Chybu lze modelovat aditivn¥ (A) nebo multiplikativn¥ (M). Trend m·oeme rovn¥o modelovat aditivn¥ (A), multiplikativn¥ (M) nebo ho lze neuvaovat (N). K aditivnímu a multiplikativnímu trendu máme vody je²t¥ jednu alternativu 'damped' (A_d), (M_d), která uvauje parametr navíc. Sezónnost je moné modelovat aditivn¥ (A), multiplikativn¥ (M) nebo je moné ji neuvaovat (N). Celkem tedy $E \in \{A, M\}$, $T \in \{N, A, A_d, M, M_d\}$ a $S \in \{N, A, M\}$, co nám dává na výb¥r z 30 model·, nebo´ r·zné p°ístupy lze kombinovat.

ETS modely pro £asové °ady y_t jsou zaloené na hladin¥ °ady ℓ_t, trendu °ady b_t a sezónnosti s_{t-m}, kde m je délka sezónnosti. Hladina ℓ_t je lineární kombinace pozorované £asové °ady o£i²t¥né od sezónnosti a p°edchozí hodnoty hladiny s trendem. Trend je lineární kombinace p°íbytku hladiny a p°edchozí hodnoty trendu. Sezónnost je lineární kombinace pozorované £asové °ady o£i²t¥né od hladiny a trendu a p°edchozí hodnoty ve stejném sezónním období.

Bodová p°edpov¥¤ je sloením (se£tením, vynásobením £i kombinací) sloek hladiny, trendu a sezónnosti. P°ehlednou tabulku v²ech ETS model· nalezneme v díle Hyndman a Khandakar (2008) na stran¥ 4. Bodová predikce pro jednotlivé modely je vody stejná, a´ uo chybu modelujeme aditivn¥, nebo multiplikativn¥, proto se v literatu°e Hyndman a Khandakar (2008) rozli²uje pouze trend a sezónnost.

Ozna£me stavový vektor $x_t = (\ell_t, b_t, s_t, s_{t-1}, \ldots, s_{t-m+1})^T$. Obecn¥ lze v²echny ETS modely zapsat ve tvaru:

$$\begin{aligned} y_t &= w(x_{t-1}) + r(x_{t-1})\,\varepsilon_t, \\ x_t &= f(x_{t-1}) + g(x_{t-1})\,\varepsilon_t, \end{aligned} \tag{4.12}$$

kde ε_t je gaussovský bílý ²um[1] a funkce $w(x_{t-1})$ je podmín¥ná st°ední hodnota $\mu_t = E[y_t | x_{t-1}]$.

Vhodn¥ zvolené funkce $f(x_{t-1})$, $g(x_{t-1})$ a $r(x_{t-1})$ ur£ují tvar ETS modelu. Pokud uvaujeme aditivní chybu, pak $r(x_{t-1}) = 1$; pokud naopak uvaujeme multiplikativní strukturu chyby, pak $r(x_{t-1}) = \mu_t$.

Kdy máme de nované výše zmín¥né rekurzivní modely, je zapot°ebí je²t¥ odhadnout po£áte£ní hodnoty - vektor $x_0 = (\ell_0, b_0, s_0, s_{-1}, \ldots, s_{-m+1})$ a parametry α, β, γ, ϕ. Ty budeme odhadovat metodou maximální v¥rohodnosti. Pro podrobnosti ohledn¥ odhad· metodou maximální v¥rohodnosti odkazujeme na £lánky Hyndman a Khandakar (2008) a Ord a kol. (1997).

Nyní jsme schopni zkonstruovat v²echny modely a zbývá vybrat ten nejvhodn¥j²í. Sou£ástí ETS metody £asto bývá vybrání modelu podle Akaikeho informa£ního kritéria (AIC), které je de novano jako:

$$AIC = -2\ln[L(\theta)] + 2(\theta + q), \tag{4.13}$$

kde $L(\theta)$ je v¥rohodnostní funkce, θ je vektor parametr·, které je pot°eba odhadnout a q je po£et parametr·, které mají být odhadnuty.

[1] Gaussovský bílý ²um je klasický bílý ²um, kde navíc platí, oe kaodá náhodná veli£ina má normální rozd¥lení s nulovou st°ední hodnotou a kladným kone£ným rozptylem σ^2.

AIC není samoz°ejm¥ jediným moºným výb¥rovým kritériem, zvolit lze i nap°íklad Bayesovo informa£ní kritérium (BIS). My v²ak budeme vyuºívat AIC a nejlep²í model bude ten, jehoº AIC hodnota bude nejni²í.

Náhodná procházka

Náhodná procházka s driftem je £asová °ada $X_t; t = 0; 1; : : : n$; která je denována jako

$$X_{t+1} = \quad + X_t + \quad _{t+1}; \qquad t = 0; : : : n \quad 1; \tag{4.14}$$

kde je konstanta, která ozna£uje drift a $_{t+1}$ je gaussovský bílý ²um, který chápeme jako chybový £len v £ase $t+1$:

Budoucí hodnoty náhodné procházky s driftem v £ase $n+h$ jsou dány vztahem:

$$\hat{X}_{n+h} = E[X_{n+h} | X_1; : : : ; X_n] = h \quad + X_n; \qquad h = 1; : : : ; H: \tag{4.15}$$

Rozptyl budoucích hodnot spo£ítáme

$$var[\hat{X}_{n+h}] = var[X_{n+h} | X_1; : : : ; X_n] = h \quad ^2; \tag{4.16}$$

kde ^{2} je rozptyl gaussovského bílého ²umu $f_t g$: Pokud uvaºujeme náhodnou procházku bez driftu, znamená to, ºe parametr $= 0$: Díky 4.15 a 4.16 platí:

$$\begin{aligned} \hat{X}_{n+h} &= X_n; \\ var[\hat{X}_{n+h}] &= h \quad ^2: \end{aligned} \tag{4.17}$$

4.2.2 Intervalová predikce

asto mohou být uºite£né také intervalové predikce. Jedná se o interval, ve kterém s danou pravd¥podobností leºí skute£ná hodnota budoucího pozorování. V na²em p°ípad¥ se tedy budeme snaºit odpov¥d¥t na otázku, kde bude leºet s pravd¥podobností 0.95 budoucí hodnota po£tu úmrtí v roce t a ve v¥ku x?

Existuje mnoho metod, kterými lze najít predik£ní intervaly. Jednou z nich je i metoda bootstrap, která se hodí pouºít v p°ípadech, kdy nemáme dostatek dat nebo kdyº je obtíºné £i nemoºné odvodit teoretické statistiky náhodných veli£in. V £lánku Shang a Haberman (2019) je navrºena konkrétní verze neparametrické bootstrap metody, kterou zde více popí²eme.

Bootstrap metoda

Základní my²lenka bootstrap metody je generování náhodných výb¥r· s opakováním ze stávajících dat. Tím získáme velké mnoºství dat, pomocí nichº nalezneme intervalový odhad budoucích hodnot.

Definice 2. M¥jme posloupnost prvk· $f x_i g_{i=1}^n$: Náhodný výb¥r s vracením o velikosti m z posloupnosti $f x_i g_{i=1}^n$ se rozumí vektor $(y_1; : : : ; y_m)$; kde y_j je náhodná hodnota x_i z posloupnosti $f x_i g_{i=1}^n$ pro $j = 1; : : : ; m$:

Máme dva zdroje chyby - chyba vzniklá p°i rozkladu matice Z na hlavní komponenty a chyba vzniklá p°i predikci. My chceme bootstrapem predikovat parametry $z_{x;t_n+h}$: Z formule 4.9 vidíme, ºe musíme predikovat nové hodnoty pro $\tilde{}_{l;t_n+h}$:

Upravme zde je²t¥ zna£ení. Budoucí hodnoty jsou toti° predikce, které jsou podmín¥né známými hodnotami. Hodnoty v £ase $t_n + h$; $h = 1; : : : ; H$ podmín¥né hodnotami do £asu t_n budeme zna£it $\tilde{}_{l;t_n+h|t_n}$ a platí:

$$\tilde{}_{l;t_n+h|t_n} = E[\ _{t_n+h} |\ _{t_1}; : : : ;\ _{t_n}]: \qquad (4.18)$$

Postup pro sestrojení bootstrap odhad· je následující:

1. M¥jme £asovou °adu faktorových skór $f\hat{}_{l;t_1}; : : : ; \hat{}_{l;t_n} g$ pro $l = 1; : : : ; L$: De nujme chybu predikce h krok· dop°edu v £ase t pro l té faktorové skóre jako

$$\hat{}_{l;h;t} = \hat{}_{l;t} \quad \tilde{}_{l;t|t\ h}; \qquad h = 1; : : : ; H; \quad t = t_{h+1}; : : : ; t_n: \qquad (4.19)$$

   Bootstrap hodnoty parametru $\hat{}^{(b)}_{l;h;}$ získáme náhodným výb¥rem s vracením z $f\hat{}_{l;h;t} g_{t=t_{h+1}}^{t_n}$:

2. Kdy° máme spo£ítány bootstrap hodnoty parametru $\hat{}^{(b)}_{l;h;}$; bootstrap hodnoty parametru $\tilde{}^{(b)}_{l;t_n+h|t_n}$ získáme pouhým dosazením ze vztahu:

$$\tilde{}^{(b)}_{l;t_n+h|t_n} = \tilde{}_{l;t_n+h|t_n} + \hat{}^{(b)}_{l;h;}\ ; \qquad b = 1; : : : ; B; \qquad (4.20)$$

   kde $B = 1000$; $\tilde{}_{l;t_n+h|t_n}$ je budoucí hodnota £asové °ady získaná standardním predik£ním algoritmem a $\hat{}^{(b)}_{l;h;}$ jsou bootstrap hodnoty z bodu 1. Tím jsme pokryli první zdroj chyby.

3. Nyní se musíme vypo°ádat s chybami $_{x;n+h}$ vzniklými p°i p°edpov¥di budoucích hodnot. Platí

$$\hat{}_{x;t} = z_{x;t} \sum_{l=1}^{L} \hat{}_{l;t}\ \hat{}_{x;l}; \qquad x = x_1; : : : ; x_m; \quad t = t_1; : : : ; t_n: \qquad (4.21)$$

   Bootstrap hodnoty parametru $\hat{}^{(b)}_{x;t_n+h}$ pro $h = 1; : : : ; H$ získáme náhodným výb¥rem s vracením z $f\hat{}_{x;t} g_{t=t_1}^{t_n}$:

4. Pomocí dvou nezávislých bootstrap metod jsme získali bootstrap hodnoty $\tilde{}^{(b)}_{l;t_n+h|t_n}$ a $\hat{}^{(b)}_{x;t_n+h}$ pro $b = 1; : : : ; B$: Bootstrap hodnoty $z^{(b)}_{x;t_n+h|t_n}$ dopo£ítáme pouze dosazením do vzorce:

$$z^{(b)}_{x;t_n+h|t_n} = \sum_{l=1}^{L} \tilde{}^{(b)}_{l;t_n+h|t_n}\ \hat{}_{x;l} + \hat{}^{(b)}_{x;t_n+h}; \qquad b = 1; : : : ; B; \qquad (4.22)$$

   kde $B = 1000$:

5. S takto vytvo°enými hodnotami $z^{(b)}_{x;t_n+h|t_n}$ provedeme stejné inverzní transformace 6 a 7, £ím° získáme bootstrap hodnoty budoucích tabulkových po£t· zem°elých $d^{(b)}_{x;t_n+h|t_n}$:

6. Nyní chceme vytvo°it intervalový odhad pro budoucí hodnoty tabulkového po£tu zem°elých s pravd¥podobností pokrytí 95 %. Pot°ebujeme tedy nalézt dolní a horní meze pro tento interval, které nalezneme jako 0.025 a 0.975 kvantily z $f d^{(1)}_{x;t_n+h|t_n}; d^{(2)}_{x;t_n+h|t_n}; : : : ; d^{(B)}_{x;t_n+h|t_n} g$ pro $x = x_1; : : : ; x_m$ a pro zvolené $h = 1; : : : ; H$:

Bootstrap metoda pro LC a RH modely

Tato verze bootstrapu lze aplikovat rovn¥º na LC model a RH model. Budeme vyuºívat stejné zna£ení, které jsme zavedli p°i bootstrap metod¥ pro CoDa metodu. Analogicky aplikujeme boostrap metodu:

1. M¥jme £asovou °adu index· $\{\hat{k}_1^{(i)}, \hat{k}_2^{(i)}, \ldots, \hat{k}_n^{(i)}\}$ pro $i = 1, 2$. V p°ípad¥ LC modelu pouze pro $i = 1$. Ozna£me chybu predikce h krok· dop°edu v £ase t pro i tý index $k_t^{(i)}$ jako

$$\hat{\varepsilon}_{i,h,t} = \hat{k}_t^{(i)} - \hat{k}_{t|t-h}^{(i)}, \qquad h = 1, \ldots, H; \quad t = t_{h+1}, \ldots, t_n. \tag{4.23}$$

   Bootstrap hodnoty parametru $\hat{\varepsilon}_{i,h,}^{(b)}$ získáme náhodným výb¥rem s vracením z $\{\hat{\varepsilon}_{i,h,t}\}_{t=t_{h+1}}^{t_n}$.

2. Kdyº máme spo£ítány bootstrap hodnoty parametru $\hat{\varepsilon}_{i,h,}^{(b)}$, bootstrap hodnoty parametru $[k_{t_n+h|t_n}^{(i)}]^{(b)}$ získáme pouhým dosazením ze vztahu:

$$[k_{t_n+h|t_n}^{(i)}]^{(b)} = k_{t_n+h|t_n}^{(i)} + \hat{\varepsilon}_{i,h,}^{(b)}, \qquad b = 1, \ldots, B, \tag{4.24}$$

   kde $B = 1000$, $k_{t_n+h|t_n}^{(i)}$ je budoucí hodnota £asové °ady získaná standardním predik£ním algoritmem a $\hat{\varepsilon}_{i,h,}^{(b)}$ je bootstrap hodnota z bodu 1. Tím jsme pokryli první zdroj chyby.

3. Nyní se musíme vypo°ádat s chybami ε_{x,t_n+h} vzniklými p°i p°edpov¥di budoucích hodnot. Platí

$$\hat{\varepsilon}_{x,t} = \ln m_{x,t} - \ln \hat{m}_{x,t}, \qquad x = x_1, \ldots, x_m; \quad t = t_1, \ldots, t_n. \tag{4.25}$$

   Bootstrap hodnoty parametru $\hat{\varepsilon}_{x,t_n+h}^{(b)}$ pro $h = 1, \ldots, H$ získáme náhodným výb¥rem s vracením z $\{\hat{\varepsilon}_{x,t}\}_{t=t_1}^{t_n}$.

4. Pomocí dvou nezávislých bootstrap metod jsme získali bootstrap hodnoty $[k_{t_n+h|t_n}^{(i)}]^{(b)}$ a $\hat{\varepsilon}_{x,t_n+h}^{(b)}$ pro $b = 1, \ldots, B$. Bootstrap hodnoty $\ln \hat{m}_{x,t_n+h|t_n}^{(b)}$ dopo£ítáme pouze dosazením do vzorce:

$$\ln \hat{m}_{x,t_n+h|t_n}^{(b)} = \hat{a}_x + \sum_{i=1}^{2} [k_{t_n+h|t_n}^{(i)}]^{(b)} \hat{b}_x^{(i)} + \hat{\varepsilon}_{x,t_n+h}^{(b)}, \qquad b = 1, \ldots, B, \tag{4.26}$$

   kde $B = 1000$.

5. Na záv¥r spo£ítáme bootstrap hodnoty budoucích tabulkových po£t· zem°elých $d_{x,t_n+h|t_n}^{(b)}$ analogicky dle 2.28 a 2.29.

6. Sestrojení dolní a horní meze pro intervalový odhad je pak totoºné jako v CoDa metod¥.

5. Kvalita model·

Modely popsané v p°ede²lých kapitolách je zapot°ebí porovnat a zjistit jejich kvalitu. Zde si popí²eme n¥kolik kritérií, které lze vyuºívat k m¥°ení kvality model·, porovnání model· a k výb¥ru nejvhodn¥j²ího modelu.

Vhodným prost°edkem pro gra cké posouzení toho, jak dob°e model vysv¥tluje data, jsou reziduální grafy. Nej£ast¥ji se pouºívají grafy rezidua versus vyrovnané hodnoty, kde poºadujeme náhodné rozloºení bod· v grafu. Pokud tomu tak není a pozorujeme v grafu n¥jaký vzor, znamená to, ºe v datech je n¥jaká systematická chyba, kterou model nedokázal vysv¥tlit. N¥kdy m·ºe být vhodn¥j²í pouºít graf se standardizovanými reziduy.

Kvalitu modelu m·ºeme rovn¥º m¥°it pomocí r·zných kritérií. Jedním z nich je koe cient determinace R^2:

De nice 3. Nech´ $Y_1; : : : ; Y_n$ jsou empirická data a $\hat{Y}_1; : : : ; \hat{Y}_n$ odhady jejich st°edních hodnot na základ¥ modelu M. Koe cient determinace R^2 de nujeme jako pom¥r variability vysv¥tlené modelem v·£i celkové variabilit¥ empirických dat:

$$R^2 = 1 - \frac{SS_e}{SS_T} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}; \qquad (5.1)$$

kde $\bar{Y} = \sum_{i=1}^{n} (Y_i)/n$.

Koe cient determinace nabývá hodnot $[0;1]$; kde model s vy²²í hodnotou R^2 lépe popisuje data neº model s niº²í hodnotou R^2 aplikovaný na matici tabulkových po£t· zem°elých má pak následující tvar:

$$R^2 = 1 - \frac{\sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} (d_{x;t} - \hat{d}_{x;t})^2}{\sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} (d_{x;t} - \bar{d}_x)^2}; \qquad (5.2)$$

kde $\bar{d}_x$ je pr·m¥r tabulkových po£t· zem°elých ve v¥kux p°es £ast.

Modely tedy lze porovnat podle R^2. Za lep²í model povaºujeme ten, jenº má hodnotu R^2 vy²²í. Dal²í moºností, jak ur£it který model je lep²í, m·ºe být m¥°ení chyby predikce. Shang a Haberman (2019) navrhují pouºít st°ední absolutní procentuální chybu MAPE.

De nice 4. Nech´ $Y_1; : : : ; Y_n; : : : ; Y_{n+h}$ jsou empirická data a $\hat{Y}_1; : : : ; \hat{Y}_n$ odhady jejich st°edních hodnot na základ¥ modelu M a $\hat{Y}_{n+1}; : : : ; \hat{Y}_{n+h}$ jsou modelové predikce budoucích hodnot. St°ední absolutní procentuální chybu modelu MAPE (Mean Absolute Percentage Error) de inujeme jako

$$\text{MAPE} = \frac{100}{h} \sum_{i=1}^{h} \frac{Y_{n+i} - \hat{Y}_{n+i}}{Y_{n+i}} : \qquad (5.3)$$

Postup je takový, ºe z pozorovaných dat odebereme h posledních pozorování a na základ¥ zbylých dat se vytvo°í predikce pro tato vynechaná pozorování. Poté se predikované hodnoty porovnají s odpovídajícími skute£nými hodnotami a ur£í se velikost chyby MAPE. Jako lep²í model ozna£íme ten model, jenº má chybu MAPE men²í.

Aplikujme nyní chybu MAPE do tvaru, ve kterém ji budeme chtít použít na naše data. Z dat budeme odebírat posledních 20 pozorování, tedy $n = 20$. Naše data jsou zde tabulkové položky zemřelých $d_{x;t}$ a ty máme dostupné pro všechny věky $x = x_1; \ldots; x_m$. Celkem tedy máme mh pozorování a MAPE má tvar:

$$\text{MAPE} = \frac{100}{mh} \sum_{i=1}^{h} \sum_{x=x_1}^{x_m} \frac{d_{x;t_n+i} - \hat{d}_{x;t_n+i}}{d_{x;t_n+i}}. \tag{5.4}$$

Dalším způsobem, jak modely porovnat, je zaměřit se na přesnost intervalových odhadů. Gneiting a Raftery (2007) navrhli intervalové skóre S, které si popíšeme.

Hlavní myšlenka je obodovat kvalitu intervalového odhadu. Ten by měl být v optimálním případě co nejužší a měl by obsahovat skutečnou hodnotu. Uvažujme predikční interval s pravděpodobností pokrytí $100(1-\alpha)$, který je symetrický, tedy spodní respektive horní mez intervalu ($\hat{d}^{l}_{x;t_n+i}$ resp. $\hat{d}^{u}_{x;t_n+i}$) je $\alpha/2$ respektive $1-\alpha/2$ kvantil. Gneiting a Raftery (2007) měří kvalitu odhadu predikčního intervalu s pravděpodobností pokrytí $100(1-\alpha)$ pro věk x v čase $t_n + i$ jako délku intervalu + penalizaci v případě, že predikční interval nepokrývá skutečnou hodnotu:

$$\begin{aligned} S_{\alpha;x;i}\left[\hat{d}^{l}_{x;t_n+i}; \hat{d}^{u}_{x;t_n+i}; d_{x;t_n+i}\right] &= \left(\hat{d}^{u}_{x;t_n+i} - \hat{d}^{l}_{x;t_n+i}\right) \\ &+ \frac{2}{\alpha}\left(\hat{d}^{l}_{x;t_n+i} - d_{x;t_n+i}\right) 1_{\left(d_{x;t_n+i} < \hat{d}^{l}_{x;t_n+i}\right)} \\ &+ \frac{2}{\alpha}\left(d_{x;t_n+i} - \hat{d}^{l}_{x;t_n+i}\right) 1_{\left(d_{x;t_n+i} > \hat{d}^{u}_{x;t_n+i}\right)}; \end{aligned} \tag{5.5}$$

kde $1_{a>b}$ je funkce indikátoru, jež nabývá pouze hodnot 0 a 1 a je dána předpisem:

$$1_{(a>b)} = \begin{cases} 1 & \text{pokud } a > b; \\ 0 & \text{pokud } a \leq b. \end{cases}$$

Když máme zavedené intervalové skóre pro predikční interval s pravděpodobností pokrytí $100(1-\alpha)$ pro věk x v čase $t_n + i$, můžeme definovat intervalové skóre podle Gneiting a Raftery (2007):

Definice 5. Nechť $d_{x;t_1}; \ldots; d_{x;t_n}; \ldots; d_{x;t_n+h}$ jsou empirická data a $\hat{d}_{x;t_1}; \ldots; \hat{d}_{x;t_n}$ odhady jejich středních hodnot na základě modelu M a $\hat{d}_{x;t_n+1}; \ldots; \hat{d}_{x;t_n+h}$ jsou modelové predikce budoucích hodnot pro $x = x_1; \ldots; x_m$. Intervalové skóre s pravděpodobností pokrytí $100(1-\alpha)$ modelu M je definováno jako:

$$S = \frac{1}{mh} \sum_{i=1}^{h} \sum_{x=x_1}^{x_m} S_{\alpha;x;i}\left[\hat{d}^{l}_{x;t_n+i}; \hat{d}^{u}_{x;t_n+i}; d_{x;t_n+i}\right]; \tag{5.6}$$

kde $S_{\alpha;x;i}\left[\hat{d}^{l}_{x;t_n+i}; \hat{d}^{u}_{x;t_n+i}; d_{x;t_n+i}\right]$ je intervalové skóre s pravděpodobností pokrytí $100(1-\alpha)$ pro věk x v čase $t_n + i$ definované v 5.5.

6. Ocen¥ní ºivotních smluv

Predikce tabulkového po£tu zem°elých m·ºe mít °adu uplatn¥ní jako nap°íklad ocen¥ní ºivotních smluv. To se m·ºe hodit p°edev²ím poji²´ovacím institucím £i penzijním fond·m. Pro výpo£et ceny ºivotní smlouvy je nezbytnou sou£ástí znalost úmrtnosti populace v budoucnu.

V této kapitole se zam¥°íme na teoretický výpo£et cen ºivotních smluv v závislosti na predikovaných hodnotách tabulkových po£t· zem°elých, které jsme získali v kapitolách 2, 3 a 4. Konkrétn¥ se budeme zabývat smlouvami, u nichº je zaplaceno jednorázové pojistné a výplaty pojistného pln¥ní probíhají na ro£ní bázi:

- Do£asný okamºitý d·chod s výplatami na konci období (Temporary life annuity in arrear)
- Doºivotní okamºitý d·chod s výplatami na konci období (Whole life annuity in arrear)
- Odloºený do£asný d·chod s výplatami na konci období (Deferred temporary life annuity in arrear)
- Odloºený doºivotní d·chod s výplatami na konci období (Deferred whole life annuity in arrear)

Pro výpo£et t¥chto d·chod· budeme pot°ebovat vyjád°it pravd¥podobnost p°eºití jedince ve v¥kux dal²ích let. Platí:

$$p_x = \prod_{i=1}^{Y} p_{x+i-1} = \prod_{i=1}^{Y} \left(1 - \frac{d_{x+i-1}}{l_{x+i-1}}\right) = \prod_{i=1}^{Y} \left(1 - \frac{d_{x+i-1}}{100000 - \sum_{j=0}^{x+i-2} d_j}\right): \quad (6.1)$$

Cenu v²ech ºivotních produkt· vypo£teme jako st°ední sou£asnou hodnotu budoucích výplat d·chod·. K tomu je²t¥ pot°ebujeme ur£it zp·sob diskontování. Vzhledem k tomu, ºe pravd¥podobnosti úmrtí máme vypo£tené pouze po rocích, nabízí se pouºít diskrétní diskontování. Nicmén¥ spojité diskontování bývá p°esn¥j²í a lze také aplikovat na diskrétní data. Proto pouºijeme spojité diskontování pomocí diskontního faktoru, který se spo£ítá jako exponenciální funkce s exponentem, jenº se rovná záporné hodnot¥ úrokové míry( ) násobené £asem, exp(( )); stejn¥ jako tomu ud¥lali Shang a Haberman (2019). Dále uvaºujme konstantní úrokovou míru ( ) = pro v²echna = 1; : : : ;n:

První zmín¥nou ºivotní smlouvou je do£asný okamºitý d·chod s výplatami na konci období. Jedná se o produkt, kde poji²t¥ný zaplatí jednorázové pojistné NSP (Net Single Premium), za coº bude dostávat pojistné pln¥ní B (Bene t) za kaºdý následující doºitý rok ihned od data platnosti smlouvy po p°edem stanovenou dobu n.

Uvaºujme, ºe K zna£í po£et doºitých let od za£átku platnosti smlouvy, neboli ºe smrt nastala b¥hemK + 1: roku. Sou£asné hodnoty výplat pojistných pln¥ní pak jsou:

$$PV(B) = \begin{cases} 0 & \text{pokud } K = 0; \\ B[e^{-1} + e^{-2} + \cdots + e^{-K}] & \text{pokud } K = 1; : : : ; n; \\ B[e^{-1} + e^{-2} + \cdots + e^{-n}] & \text{pokud } K = n+1; : : : \end{cases}$$

Pojistné NSP stanovíme jako

$$a_{x:\overline{n}|} = E[PV(B)] = B \sum_{=1}^{n} e \quad p_x; \qquad (6.2)$$

kde B je p°edem domluvená výše pojistného pln¥ní vyplacená kaºdý rok ve stanoveném období, p_x je pravd¥podobnost p°eºití de novaná v 6.1 a $ex(p \quad)$ je diskontní faktor p°edpokládající konstantní úrokovou míru:

Druhou zmín¥nou ºivotní smlouvou je doºivotní okamºitý d·chod s výplatami na konci období, coº je produkt, který se od do£asného okamºitého d·chodu li²í pouze tím, ºe výplaty pojistného pln¥ní neprobíhají po p°edem stanovenou omezenou dobu n, ale doºivotn¥. Sou£asné hodnoty výplat pojistných pln¥ní mají podobu:

$$PV(B) = \begin{cases} 0 & \text{pokud } K = 0; \\ B[e^{1} + e^{2} + \quad + e^{K}] & \text{pokud } K = 1;:: \end{cases}$$

Jednorázové pojistné NSP tak stanovíme jako

$$a_x = E[PV(B)] = B \sum_{=1}^{1} e \quad p_x; \qquad (6.3)$$

s tím, ºe vyuºíváme stejné zna£ení jako p°i výpo£tu $a_{x:\overline{n}|}$:

Zbylé dva zmín¥né produkty jsou alternativy k prvním dv¥ma s tím, ºe výplata neprobíhá okamºit¥ po uzav°ení smlouvy, nýbrº aº po p°edem stanované dob¥ m.

Sou£asné hodnoty výplat odloºeného do£asného d·chodu s výplatami na konci období jsou:

$$PV(B) = \begin{cases} 0 & \text{pokud } K = 0;:::;m; \\ Be^{m}[e^{1} + e^{2} + \quad + e^{K}] & \text{pokud } K = m+1;:::;m+n; \\ Be^{m}[e^{1} + e^{2} + \quad + e^{n}] & \text{pokud } K = m+n+1;:: \end{cases}$$

Pojistné NSP spo£ítáme jako:

$$ {}_{m|}a_{x:\overline{n}|} = E[PV(B)] = B \left[\sum_{=1}^{n} e^{(m+\ )} p_{x+m} \ {}_m p_x \right] = Be^{m}\, {}_m p_x \sum_{=1}^{n} e \quad p_{x+m}; \qquad (6.4)$$

kde op¥t vyuºíváme stejné zna£ení jako p°i výpo£tu $a_{x:\overline{n}|}$:

Sou£asné hodnoty výplat odloºeného doºivotního d·chodu s výplatami na konci období jsou:

$$PV(B) = \begin{cases} 0 & \text{pokud } K = 0;:::;m; \\ Be^{m}[e^{1} + e^{2} + \quad + e^{K}] & \text{pokud } K = m+1;:: \end{cases}$$

Pojistné NSP spo£ítáme jako

$$ {}_{m|}a_x = E[PV(B)] = B \left[\sum_{=1}^{1} e^{(m+\ )} p_{x+m} \ {}_m p_x \right] = Be^{m}\, {}_m p_x \sum_{=1}^{1} e \quad p_{x+m}; \qquad (6.5)$$

kde op¥t vyuºíváme stejné zna£ení jako p°i výpo£tu $a_{x:\overline{n}|}$:

7. Aplikace na reálná data

V této kapitole aplikujeme teoretické poznatky, které jsme p°edstavili v p°edchozích kapitolách, na reálná data mu^{o}· a ^{o}en v ƒeské republice. Na webové stránce ƒeský statistický ú°ad (2023) jsou dostupné úmrtnostní tabulky pro ƒR od roku 1920 a^{o} po sou£asnost (2021). Na tato data aplikujeme LC, RH a PCA (Analýza kompozi£ních dat) modely, provedeme predikce pro roky 2022-2051, porovnáme modely a odhadneme budoucí pravd¥podobnosti p°e^{o}it$_{x}$ a spo£ítáme ceny sou£asných i budoucích ^{o}ivotních produkt·. Celkem budeme uva^{o}ovat 13 model· - z toho t°ikrát LC model s r·znými metodami pro odhady parametr· (SVD, Poissonova metoda, alternativná metoda), dvakrát RH model (SVD, Poissonova metoda) a osmkrát PCA model s r·znými metodami pro predikci (ARIMA modely, ETS modely, náhodná procházka s driftem a náhodná procházka bez driftu) a s dv¥ma r·znými metodami pro ur£ení po£tu hlavních komponent$t$.

V celé kapitole budeme p°edpokládat konstatní intenzitu úmrtnosti a také, ^{o}e st°ední stav populaceP_x se rovná centrální expoziciN_x:

7.1 P°edstavení dat

ƒeský statistický ú°ad archivuje úmrtnostní tabulky od roku 1920. V t¥chto tabulkách jsou dostupné úmrtnostní míry a tabulkové po£ty zem°elých pro v¥kové skupiny s jednoletým intervalem, a to pro skupiny 0 a^{o} 105+, kde poslední skupina zahrnuje v²echny v¥ky v¥t²í nebo rovno 105. Jak jsme se dozv¥d¥li v podsekci 1.2, úmrtnostní tabulky jsou vypo£teny na základ¥ po£tu úmrtíD_x a st°edního stavu populaceP_x: ƒSÚ m¥l k dispozici hodnotyD_x a P_x v letech 1920-1923 pro v¥ky 0-94, v letech 1924-2006 pro v¥ky 0-99 a v letech 2007-2021 pro v¥ky 0-105. Úmrtnostní míry, které poskytuje ƒeský statistický ú°ad (2023), jsou vyhlazené, jak jsme zmínili v sekci 1.2.

V LC a RH modelu vychází odhady parametr· z úmrtnostních m¥r$m_{x;t}$; v p°ípad¥ SVD metody se v druhé £ásti odhad· parametrk_t vyu^{o}ívá rovn¥^{o} po£t· úmrtí $D_{x;t}$ a st°edních stav· populace$P_{x;t}$: PCA metoda je zalo^{o}ena na znalosti tabulkových po£t· úmrtí $d_{x;t}$:

Na obrázcích 7.1 a 7.2 máme p°edstaveny tabulkové po£ty zem°elých pro vybrané v¥ky. Na základ¥ t¥chto graf· jsme se rozhodli dále pokra£ovat pouze s daty z let 1970-2021.

Na obrázcích 7.3 a 7.4 máme gra cky vyobrazeny úmrtnostní míry pro vybrané roky pro mu^{o}e a ^{o}eny. Úmrtnostní míry nejprve klesají, co^{o} je zp·sobeno novorozeneckou úmrtností, která je vy²²í ne^{o} úmrtnost u malých d¥tí obecn¥. Po zbytek v¥k· je patrný exponenciální r·st úmrtnostních m¥r.

Pokud budeme porovnávat zm¥nu úmrtnostních m¥r v £ase, u ^{o}en na obrázku 7.4 je viditelný pokles hodnot úmrtnostních m¥r od v¥ku 60. U mu^{o}· na obrázku 7.3 pozorujeme obdobný trend.

Obrázky 7.5 a 7.6 nám gra cky p°edstavují tabulkové po£ty zem°elých$d_{x;t}$ pro vybrané roky pro mu^{o}e a ^{o}eny. I tyto parametry nejprve klesají z d·vodu vy²²í úmrtnosti novorozenc·. P°ibli^{o}n¥ od v¥ku 10 po£et zem°elých roste, výrazné zrychlení r·stu pozorujeme okolo v¥ku 50-60. Naopak okolo v¥ku 70-90, v závislosti na pohlaví a roku, po£et zem°elých klesá, nebo´ jedinc· v t¥chto

v¥cích v populaci je obecn¥ mén¥.

Oba obrázky 7.5 a 7.6 rovn¥º zachycují trend prodluºování délky ºivota - s rostoucím £asem se vrchol k°ivky zvy²uje a posouvá se do pozd¥j²ích v¥k·.

Pokud budeme porovnávat obrázek 7.5 se 7.6, m·ºeme si v²imnout, ºe ºeny obecn¥ umírají ve vy²²ím v¥ku.

Obrázek 7.1: Porovnání tabulkových po£t· zem°elých muº· ve v¥ku 25, 50, 75 a 100 let.

Obrázek 7.2: Porovnání tabulkových po£t· zem°elých ºen ve v¥ku 25, 50, 75 a 100 let.

Obrázek 7.3: Porovnání úmrtnostních m¥r muº· z let 1970, 1980, 1990, 2000, 2010 a 2020.

Obrázek 7.4: Porovnání úmrtnostních m¥r ºen z let 1970, 1980, 1990, 2000, 2010 a 2020.

Obrázek 7.5: Porovnání tabulkových po£t· zem°elých muº· z let 1970, 1980, 1990, 2000, 2010 a 2020.

Obrázek 7.6: Porovnání tabulkových po£t· zem°elých ºen z let 1970, 1980, 1990, 2000, 2010 a 2020.

7.2 Odhady parametr· model·

Pro odhady parametr· model· vyuºíváme data od roku 1970 do 2021 pro v¥ky 0 aº 105+. Platí tedy, ºe $n = 51$ a $m = 106$:

7.2.1 LC model

SVD metoda

Pro odhad parametr· jsme vyuºili v R softwaru funkce svd() . Odhadnuté parametry jsme znormalizovali pomocí normaliza£ních podmínek 2.7, p°eodhadli jsme parametr k_t vy°e²ením rovnice 2.15 Newtonovým Raphsonovým itera£ním algoritmem a na záv¥r jsme odhady znovu znormalizovali. Výsledné odhady jsou k náhlednutí na obrázku 7.7 pro muºe a na obrázku 7.8 pro ºeny.

Obrázek 7.7: Finální odhady parametr· v LC-SVD modelu pro muºe.

Obrázek 7.8: Finální odhady parametr· v LC-SVD modelu pro ºeny.

Obrázky 7.9 a 7.10 ukazují, jak se li²í odhadnuté úmrtnostní míry od úmrtnostních m¥r pozorovaných v letech 1970, 1990, 2005 a 2020 mu^{o}· a ^{o}en.

Obrázek 7.9: Porovnání pozorovaných a odhadnutých m¥r úmrtností pro mu^{o}e LC-SVD modelem.

Obrázek 7.10: Porovnání pozorovaných a odhadnutých m¥r úmrtností pro ^{o}eny LC-SVD modelem.

Poissonova metoda

V Poissonov¥ metod¥ jsme zvolili po£áte£ní odhady parametr· na základ¥ SVD rozkladu v R software op¥t pomocí funkcesvd() a normaliza£ních podmínek 2.7. Dále jsme aktualizovali odhady parametr· pomocí itera£ních vzorc· 2.18, 2.19, 2.20, dokud úbytek hodnoty deviance $D(v; \hat{D})$ nebyl men²í neº 100. Následn¥ jsme odhady parametr· op¥t znormalizovali a provedli p°eodhadnutí parametru k_t zcela analogicky jako v sekci 7.2.1. Výsledné odhady jsou k náhlednutí na obrázku 7.11 pro muºe a na obrázku 7.12 pro ºeny.

Obrázek 7.11: Finální odhady parametr· v LC-Poisson modelu pro muºe.

Obrázek 7.12: Finální odhady parametr· v LC-Poisson modelu pro ºeny.

Obrázky 7.13 a 7.14 op¥t ukazují, jak se li²í odhadnuté úmrtnostní míry od úmrtnostních m¥r pozorovaných v letech 1970, 1990, 2005 a 2020 mu^{o}· a ^{o}en.

Oproti modelu s SVD metodou si v²ímáme zna£ných nep°esností pro nejvy²²í v¥ky. To je zp·sobené nedostupností dat pro v¥kové skupiny 100-105+ a^{o} do roku 2006. Poissonova metoda vyu^{o}ívá k odhadování parametr· po£ty úmrtí D_x a centrální expozici N_x a práv¥ nedostupnost t¥chto dat pro období 1970-2006 zp·sobuje zmín¥né nep°esnosti.

Obrázek 7.13: Porovnání pozorovaných a odhadnutých m¥r úmrtností pro mu^{o}e LC-Poisson modelem.

Obrázek 7.14: Porovnání pozorovaných a odhadnutých m¥r úmrtností pro ^{o}eny LC-Poisson modelem.

Alternativní metoda

Alternativní regresní metoda byla popsána v sekci 2.2.3. Výpo£et v R-Studiu probíhal u^{o} pouze na základ¥ vzorc· 2.8, 2.21 a 2.24. Výsledné odhady jsou gracky zobrazeny na obrázku 7.15 pro mu^{o}e a na obrázku 7.16 pro ^{o}eny.

Obrázek 7.15: Finální odhady parametr· v LC-alternativním modelu pro mu^{o}e.

Obrázek 7.16: Finální odhady parametr· v LC-alternativním modelu pro ^{o}eny.

Obrázky 7.17 a 7.18 op¥t ukazují, jak se li²í odhadnuté úmrtnostní míry od úmrtnostních m¥r pozorovaných v letech 1970, 1990, 2005 a 2020 mu^{o}· a ^{o}en.

Obrázek 7.17: Porovnání pozorovaných a odhadnutých m¥r úmrtností pro mu^{o}e LC-alternativním modelem.

Obrázek 7.18: Porovnání pozorovaných a odhadnutých m¥r úmrtností pro ^{o}eny LC-alternativním modelem.

7.2.2 RH model

SVD metoda

Zp·sob odhad· parametr· v R-Studiu je zde analogický jako v sekci 7.2.1. Výsledné odhady parametr·$a; b^{(1)}; b^{(2)}; k^{(1)}; k^{(2)}$ lze nahlédnout na obrázku 7.19 pro mu^{o}e a na obrázku 7.20 pro ^{o}eny.

Obrázek 7.19: Finální odhady parametr· v RH-SVD modelu pro mu^{o}e.

Obrázek 7.20: Finální odhady parametr· v RH-SVD modelu pro ^{o}eny.

Obrázky 7.21 a 7.22 op¥t ukazují, jak se li²í odhadnuté úmrtnostní míry od úmrtnostních m¥r pozorovaných v letech 1970, 1990, 2005 a 2020 mu^{o}· a ^{o}en.

Obrázek 7.21: Porovnání pozorovaných a odhadnutých m¥r úmrtností pro mu^{o}e RH-SVD modelem.

Obrázek 7.22: Porovnání pozorovaných a odhadnutých m¥r úmrtností pro ^{o}eny RH-SVD modelem.

Poissonova metoda

I zde je zp·sob odhad· parametr· v R-Studiu analogický odhad·m Poissonovou metodou v LC modelu (7.2.1). Výsledné odhady parametr· jsou k dispozici na obrázcích 7.23 pro muºe a na obrázku 7.24 pro ºeny.

Obrázek 7.23: Finální odhady parametr· v RH-Poisson modelu pro muºe.

Obrázek 7.24: Finální odhady parametr· v RH-Poisson modelu pro ºeny.

Obrázky 7.25 a 7.26 op¥t ukazují, jak se li²í odhadnuté úmrtnostní míry od úmrtnostních m¥r pozorovaných v letech 1970, 1990, 2005 a 2020 mu^{o}· a ^{o}en.

Stejn¥ jako v LC modelu i zde zaznamenáváme zna£né nep°esnosti v modelu pro nejvy²²í v¥ky. D·vod je op¥t nedostupnost dat D_x a N_x z let 1970-2006, které Poisson·v algoritmus vyu^{o}ívá pro odhad parametr·.

Obrázek 7.25: Porovnání pozorovaných a odhadnutých m¥r úmrtností pro mu^{o}e RH-Poisson modelem.

Obrázek 7.26: Porovnání pozorovaných a odhadnutých m¥r úmrtností pro ^{o}eny RH-Poisson modelem.

7.2.3 PCA model

V PCA modelu jsme nejprve p°etransformovali data $d_{x;t}$ pomocí vzorc· 4.2, 4.3, 4.4 a 4.5. Na základ¥ transformovaných dat $z_{x;t}$ jsme podle rovnice 4.8 ur£ili po£et hlavních komponent, které budeme odhadovat. V obou p°ípadech, jak pro mu^{o}e, tak pro ^{o}eny, nám vy²lo $L = 1$; nebo´ $_1$ zachytila 96.03 % variability v datech u mu^o·, resp. 94.48 % u ^{o}en. Následn¥ jsme odhadli parametry $\hat{}_{1;t}$ a $\hat{}_{x;1}$ pomocí funkce svd() a vzorc· 4.7.

Zajímavé by rovn¥^{o} mohlo být uva^{o}ovat více hlavních komponent a faktorových skór· pro odhad transformovaných dat $z_{x;t}$: Rozhodli jsme se proto modelovat data také pro $L = 5$; kde hlavní komponenty vystihly 98.75 % variability v datech u mu^o· a 97.99 % u ^{o}en.

Obrázek 7.27 poskytuje gra cký náhled na odhadnuté parametry $\hat{}_{1;t}$ a $\hat{}_{x;1}$ pro p°ípad $L = 1$:

Obrázek 7.27: Finální odhady parametr· v PCA modelu s jednou hlavní komponentou pro mu^{o}e a ^{o}eny.

Obrázky 7.28 a 7.29 poskytují porovnání odhadnutých parametr· $\hat{z}_{x;t}$ na základ¥ PCA modelu s 1 hlavní komponentou od pozorovaných transformovaných dat $z_{x;t}$.

Obrázky 7.30 a 7.31 p°edkládají porovnání odhadnutých parametr· $\hat{z}_{x;t}$ na základ¥ PCA modelu s 5 hlavními komponentami od pozorovaných transformovaných dat $z_{x;t}$. Pokud tyto grafy porovnáme s obrázky 7.28 a 7.29, v²imneme si výrazného vylep²ení.

Obrázek 7.28: Porovnání pozorovaných transformovaných dat $z_{x;t}$ a odhadnutých parametrů $\hat{z}_{x;t}$ pro muže PCA modelem s 1 hlavní komponentou.

Obrázek 7.29: Porovnání pozorovaných transformovaných dat $z_{x;t}$ a odhadnutých parametrů $\hat{z}_{x;t}$ pro ženy PCA modelem s 1 hlavní komponentou.

Obrázek 7.30: Porovnání pozorovaných transformovaných dat $z_{x;t}$ a odhadnutých parametrů $\hat{z}_{x;t}$ pro muže PCA modelem s 5 hlavními komponentami.

Obrázek 7.31: Porovnání pozorovaných transformovaných dat $z_{x;t}$ a odhadnutých parametrů $\hat{z}_{x;t}$ pro ženy PCA modelem s 5 hlavními komponentami.

7.3 Predikce

Na základ¥ odhadnutých model· budeme predikovat budoucí hodnoty tabulkových po£t· úmrtí $d_{x;t}$; pro roky $t = 2022; : : : ; 2051$. Horizont predikce je $H = 30$:

7.3.1 LC model

V LC modelu je na £ase závislý pouze jeden parametr a tím je, na n¥jº budeme nahlíºet jako na £asovou °adu. Tu budeme analyzovat pomocí ARIMA model·. V R Studiu jsme pro nalezení nejlep²ího ARIMA modelu vyuºili funkci auto.arima() . Predikce parametru k byla získána funkcíforecast() a dopo£tení budoucích hodnot $m_{x;t_{n+h}}$ a $d_{x;t_{n+h}}$ je pak p°ímo£aré za p°edpokladu, ºe uvaºujeme konstantní intenzitu úmrtnosti.

SVD metoda

Nejvhodn¥j²ím modelem pro LC model v kombinaci s SVD metodou pro odhad parametr· je ARIMA(1,1,1) bez driftu pro data muº· a ARIMA(0,1,0) s driftem pro data ºen. Obrázek 7.32 poskytuje gra cký pohled na sou£asné a budoucí hodnoty parametru k v£etn¥ 95% intervalových odhad·, dále pak pro ilustraci také pozorované a odhadnuté budoucí hodnoty úmrtnostních m¥r v£etn¥ 95% interval· spolehlivosti pro vybranou v¥kovou skupinu (75 let) muº· a ºen.

Obrázek 7.32: Odhadnuté a budoucí hodnoty parametru k v£etn¥ 95% intervalu spolehlivosti a pozorované a odhadnuté budoucí hodnoty úmrtnostních m¥r v£etn¥ 95% intervalu spolehlivosti - LC model SVD metoda. Vlevo grafy pro muºe, vpravo pro ºeny.

Poissonova metoda

Stejn¥ jako p°i pou$itím SVD metody je i p°i pou$ití Poissonovy metody nejlep-²ím modelem pro mu$ská data model ARIMA(1,1,1) bez driftu a pro $enská data model ARIMA(0,1,0) s driftem. Na obrázku 7.33 jsou vykresleny sou£asné a budoucí hodnoty parametruk v£etn¥ 95% intervalových odhad·, dále pak pro ilustraci také pozorované a odhadnuté budoucí hodnoty úmrtnostních m¥r v£etn¥ 95% interval· spolehlivosti pro vybranou v¥kovou skupinu (75 let) mu$· a $en.

Obrázek 7.33: Odhadnuté a budoucí hodnoty parametru k v£etn¥ 95% intervalu spolehlivosti a pozorované a odhadnuté budoucí hodnoty úmrtnostních m¥r v£etn¥ 95% intervalu spolehlivosti - LC model Poissonova metoda. Vlevo grafy pro mu$e, vpravo pro $eny.

Alternativní metoda

V LC modelu s vyu$itím alternativní metody pro odhad parametr· vy²ly nejlépe modely ARIMA(0,1,0) s driftem pro mu$e a ARIMA(1,1,0) s driftem pro $eny. Na obrázku 7.34 jsou vykresleny sou£asné a budoucí hodnoty parametru k v£etn¥ 95% intervalových odhad·, dále pak pro ilustraci také pozorované a odhadnuté budoucí hodnoty úmrtnostních m¥r v£etn¥ 95% interval· spolehlivosti pro vybranou v¥kovou skupinu (75 let) mu$· a $en.

Obrázek 7.34: Odhadnuté a budoucí hodnoty parametru κ v£etn¥ 95% intervalu spolehlivosti a pozorované a odhadnuté budoucí hodnoty úmrtnostních m¥r v£etn¥ 95% intervalu spolehlivosti - LC model alternativní metoda. Vlevo grafy pro mu^{o}e, vpravo pro ^{o}eny.

7.3.2 RH model

V RH modelu na £ase závisí parametry $\kappa^{(1)}$ a $\kappa^{(2)}$. Tyto parametry jsme modelovali jako nezávislé £asové °ady pomocí ARIMA model·. Stejn¥ tak jako v LC modelu jsme v R Studiu pro výb¥r nejlep²ího ARIMA modelu zvolili funkci auto.arima() a pro predikci parametr· $\kappa^{(1)}$ a $\kappa^{(2)}$ funkci forecast(). Dopo£tení budoucích hodnot $m_{x;t_{n+h}}$ a $d_{x;t_{n+h}}$ je pak p°ímo£aré za p°edpokladu, ^{o}e uva^{o}ujeme konstantní intenzitu úmrtnosti.

SVD metoda

V RH modelu s vyu^{o}itím SVD metody pro odhad parametr· jsme vybrali modely ARIMA(0,1,0) s driftem pro £asovou °adu $\kappa^{(1)}$ pro mu^{o}e, ARIMA(1,0,0) s nulovým pr·m¥rem pro $\kappa^{(2)}$ pro mu^{o}e, ARIMA(1,1,0) s driftem pro £asovou °adu $\kappa^{(1)}$ pro ^{o}eny a ARIMA(0,0,0) s nulovým pr·m¥rem pro $\kappa^{(2)}$ pro ^{o}eny. Sou£asné a budoucí hodnoty parametr· $\kappa^{(1)}$ a $\kappa^{(2)}$ jsou k dispozici na obrázku 7.35.

Pro ilustraci op¥t p°idáváme obrázek 7.36 s pozorovanými a odhadnutými budoucími hodnotami úmrtnostních m¥r v£etn¥ 95% interval· spolehlivosti pro vybranou v¥kovou skupinu (75 let) mu^o· a ^{o}en.

Obrázek 7.35: Odhadnuté a budoucí hodnoty parametrů $\kappa^{(1)}$ a $\kappa^{(2)}$ včetně 95% intervalu spolehlivosti - RH model, SVD metoda. Vlevo grafy pro muže, vpravo pro ženy.

Obrázek 7.36: Pozorované a odhadnuté budoucí hodnoty úmrtnostních měr včetně 95% intervalu spolehlivosti - RH model, SVD metoda. První graf pro muže 75 let, druhý pro ženy 75 let.

Poissonova metoda

V RH modelu s vyuºitím Poissonovy metody pro odhad parametr· jsme vybrali modely ARIMA(0,1,0) s driftem pro $k^{(1)}$ pro muºe, ARIMA(1,0,0) s nulovým pr·m¥rem pro $k^{(2)}$ pro muºe, ARIMA(1,1,0) s driftem pro £asovou °adu $k^{(1)}$ pro ºeny a ARIMA(0,0,0) s nulovým pr·m¥rem pro $k^{(2)}$ pro ºeny. Sou£asné a budoucí hodnoty parametr· $k^{(1)}$ a $k^{(2)}$ jsou k dispozici na obrázku 7.37.

Pro ilustraci op¥t p°idáváme obrázek 7.38 s pozorovanými a odhadnutými budoucími hodnotami úmrtnostních m¥r v£etn¥ 95% interval· spolehlivosti pro vybranou v¥kovou skupinu (75 let) muº· a ºen.

Obrázek 7.37: Odhadnuté a budoucí hodnoty parametr· $k^{(1)}$ a $k^{(2)}$ v£etn¥ 95% intervalu spolehlivosti - RH model, Poissonova metoda. Vlevo grafy pro muºe, vpravo pro ºeny.

Obrázek 7.38: Pozorované a odhadnuté budoucí hodnoty úmrtnostních m¥r v£etn¥ 95% intervalu spolehlivosti - RH model, Poissonova metoda. První graf pro muºe 75 let, druhý pro ºeny 75 let.

7.3.3 PCA model

V PCA modelu závisí na £ase pouze parametry $\gamma_{i;t}$: Na tyto parametry op¥t budeme nahlíºet jako na nezávislé £asové °ady. Máme £ty°i r·zné zp·soby, jak predikovat budoucí hodnoty. První moºnost je op¥t nalézt vhodný ARIMA model, coº provedeme v R Studiu op¥t pomocí funkce auto.arima() . Dal²í moºností je modelovat £asovou °adu jako náhodnou procházku s driftem £i bez driftu. V t¥chto p°ípadech budeme v R softwaru pouºívat funkci Arima() . Poslední moºnost je predikovat na základ¥ ETS metody - v tom p°ípad¥ nalezneme vhodný ETS model pomocí funkce ets() [1].

Ve v²ech p°ípadech budoucí hodnoty spo£ítáme funkcí forecast() . Dopo£tení budoucích hodnot $d_{x;t_{n+h}}$ je pak p°ímo£aré.

1 hlavní komponenta

Nejprve se zam¥°íme na modely, kde jsme pouºili pro ur£ení metodu popsanou v 4.8. Zde máme pouze jeden parametr $\gamma_{1;t}$ závislý na £ase. Jako nejvhodn¥j²í ARIMA model pro muºe se zdá být model ARIMA(0,1,0) s driftem a pro ºeny ARIMA(0,1,1) s driftem. V p°ípad¥ ETS model· jsme jako nejlep²í model, jak pro muºe, tak pro ºeny, nalezli ETS(A,A,N), tedy model bez sezónnosti, který uvaºuje aditivní trend a aditivní chybu.

Odhadnutý parametr $\gamma_{1;t}$ v £etn¥ v²ech 4 zp·sob· predikce nalezneme na obrázku 7.39 pro muºská data a na obrázku 7.40 pro data ºen. Pro ilustraci uvádíme i obrázky 7.41 a 7.42, kde si m·ºeme prohlédnout odhadnuté budoucí hodnoty $d_{x;t}$ pro vybrané v¥kové kategorie (25, 50, 75 a 100 let). Na obrázcích pro muºe splývají predikce ARIMA model· s náhodnou procházkou s driftem, nebo´ jsme v rámci ARIMA model· vybrali práv¥ náhodnou procházku s driftem.

Obrázek 7.39: Budoucí hodnoty $\gamma_{1;t_{n+h}}$ muº· na základ¥ PCA modelu, v závislosti na predik£ní metod¥.

[1] Funkce ets() v R Studiu v p°ípad¥ ro£ních dat neuvaºuje modely se sezónností.

Obrázek 7.40: Budoucí hodnoty $_{1;t_{n+h}}$ ºen na základ¥ PCA modelu, v závislosti na predik£ní metod¥

Obrázek 7.41: Budoucí hodnoty$\varnothing_{x;t_{n+h}}$ muº· na základ¥ PCA modelu v závislosti na predik£ní metod¥ pro vybrané v¥ky 25, 50, 75 a 100 let.

Obrázek 7.42: Budoucí hodnoty $d_{x;t_{n+h}}$ ºen na základ¥ PCA modelu v závislosti na predik£ní metod¥ pro vybrané v¥ky 25, 50, 75 a 100 let.

5 hlavních komponent

Poj¤me se nyní podívat, jaké modely jsme vybrali pro jednotlivé £asové °ady $_{l;t}$ v p°ípad¥, kdy uvaºujeme $L = 5$ pro muºe i ºeny. Nejvhodn¥j²í ARIMA a ETS modely pro zmín¥né £asové °ady máme sepsány v tab²ulce 7.1.

l	ARIMA muºi	ARIMA ºeny	ETS muºi	ETS ºeny
1	(0,1,0)+d	(0,1,1)+d	(A,A,N)	(A,A,N)
2	(1,0,0)	(1,0,0)	(A,N,N)	(A,Ad,N)
3	(0,1,1)	(0,0,0)	(A,N,N)	(A,N,N)
4	(1,0,2)	(0,0,0)	(A,N,N)	(A,N,N)
5	(0,0,0)	(0,0,0)	(A,N,N)	(A,N,N)

Tabulka 7.1: Nejlep²í ARIMA a ETS modely pro PCA5 modely.

Obrázky 7.43 a 7.44 poskytují k nahlédnutí porovnání odhadnuté budoucí hodnoty $d_{x;t}$ na základ¥ v²ech £ty° r·zných predik£ních metod pro vybrané v¥kové skupiny (25, 50, 75 a 100 let).

[2] Symbol +d zna£í, ºe v modelu se uvaºuje drift, v opa£ném p°ípad¥ se uvaºuje model bez driftu.

Obrázek 7.43: Budoucí hodnoty $\hat{y}_{x;t_{n+h}}$ mu^{o}· na základ¥ PCA modelu s 5 hlavními komponentami v závislosti na predik£ní metod¥ pro vybrané v¥ky 25, 50, 75 a 100 let.

Obrázek 7.44: Budoucí hodnoty $\hat{y}_{x;t_{n+h}}$ ^{o}en na základ¥ PCA modelu s 5 hlavními komponentami v závislosti na predik£ní metod¥ pro vybrané v¥ky 25, 50, 75 a 100 let.

7.4 Porovnání model·

Kdy^{o} máme modely odhadnuté a napredikovali jsme i budoucí hodnoty, p°ichází na °adu modely porovnat a ur£it, který model se pro na²e data hodí nejvíce.

Na²e práce, jak ji^{o} název napovídá, se zabývá primárn¥ po£tem úmrtí, budeme tedy porovnávat, jak se li²í od pozorovaných dat$d_{x;t}$ námi odhadnuté parametry $\hat{d}_{x;t}$:

7.4.1 Koe cient determinace

Nejprve se zam¥°íme na to, jak dob°e modely kopírují pozorované hodnoty. K tomu pou^{o}ijeme koe cient determinaceR^2: V tabulce 7.2 máme hodnoty parametru R^2 pro v²echny uva^{o}ované modely.

	LC-SVD	LC-PO	LC-ALT	RH-SVD	RH-PO	PCA1	PCA5
mu^{o}i	0.9997	0.9997	0.9996	0.9997	0.9997	0.9996	0.9998
^{o}eny	0.9999	0.9999	0.9998	0.9998	0.9998	0.9998	0.9999

Tabulka 7.2: Koe cient determinace pro jednotlivé modely mu^{o}· a ^{o}en.

Z tabulky 7.2 je patrné, ^{o}e v²echny modely popisují data velmi dob°e, nebo´ jsou v²echny hodnoty vy²í ne^{o} 99.95 %. Nejvy²í hodnotu má PCA model s 5 hlavními komponentami pro mu^{o}ská data a LC-SVD model[3] pro data ^{o}en. Zajímavostí je, ^{o}e LC model má vy²í hodnoty ne^{o} RH model, tedy p°idání druhé sady singulárních vektor· nevylep²ilo kvalitu modelu v rámciR^2:

7.4.2 MAPE a intervalové skóre

Dal²ími kritérii pro ur£ení nejvhodn¥j²ího modelu jsou MAPE £i intervalové skóre S. Zde odebíráme posledních 20 pozorování, rozm¥ry dat tedy budou $n = 32$ a $m = 99$, nebo´ do roku 2006 máme dostupná data pouze do v¥ku 99.

Z tabulky 7.3, kde máme zaznamenány hodnoty MAPE a$S_{0:05}$ v²ech model·, lze vy£íst, ^{o}e nejlep²ím modelem pro mu^{o}e je dle MAPE model PCA s 5 hlavními komponentami, kde jsme pro predikci volili náhodnou procházku s driftem. Pro ^{o}eny je dle MAPE nejlep²í PCA model s jednou hlavní komponentou a náhodnou procházkou s driftem.

Pozorujeme, ^{o}e obecn¥ PCA modely poskytují nejp°esn¥j²í predikce, s výjimkou náhodné procházky bez driftu. P°idáním dal²ích komponent do modelu se predikce ^{o}en nezlep²í. U mu^{o}· se budoucí hodnoty zp°esní krom¥ PCA-ARIMA modelu, nejv¥t²í rozdíl je u náhodné procházky s driftem.

LC model obecn¥ zaostává za zbylými modely, nejh·°e na tom je LC model s Poissonovou metodou. Nejlep²í odhady z LC modelu má ten s alternativní metodou pro mu^{o}e a s SVD metodou pro ^{o}eny. U RH modelu pozorujeme zp°esn¥ní budoucích odhad· oproti LC modelu krom¥ SVD metody u ^{o}en.

Pokud budeme hodnotit pomocí intervalového skóre pro $= 0:05$ s tím, ^{o}e intervalové odhady jsme sestrojili na základ¥ bootstrap metod, získáme jako nejlep²í variantu PCA model s 5 hlavními komponentami s náhodnou procházkou

[3] U ^{o}en jsou hodnoty pro LC-SVD = 0.999888, LC-Po = 0.999873, pro PCA5 = 0.999856.

Model	MAPE muºi	MAPE ºeny	$S_{0:05}$ muºi	$S_{0:05}$ ºeny
LC - SVD	19.74	13.11	843.99	853.79
LC - PO	26.41	18.49	816.79	843.96
LC - ALT	17.26	13.58	825.97	838.33
RH - SVD	17.88	13.75	806.53	827.17
RH - PO	19.48	15.22	804.47	825.65
PCA1 - ARIMA	17.08	12.08	508.68	402.16
PCA1 - ETS	18.47	12.43	585.46	432.12
PCA1 - RWD	17.08	11.93	506.59	414.16
PCA1 - RW	33.18	28.65	554.78	277.95
PCA5 - ARIMA	17.11	12.10	500.22	393.08
PCA5 - ETS	18.19	12.87	591.38	481.73
PCA5 - RWD	14.78	12.30	647.54	454.10
PCA5 - RW	31.47	29.46	499.25	285.78

Tabulka 7.3: Porovnání modelu podle MAPE a intervalového skóre.

bez driftu pro muºe a PCA model s 1 hlavní komponentou s náhodnou procházkou bez driftu pro ºeny. Nejh·°e dopadl LC model s SVD metodou. RH model má p°esn¥j²í intervalové odhady v porovnání s LC modelem. PCA modely dopadly celkov¥ nejlépe s tím, ºe nepozorujeme výrazné zlep²ení p°idáním dal²ích hlavních komponent.

Vy²²í hodnoty intervalového skóre jsou zp·sobené p°edev²ím tím, ºe n¥kdy predik£ní interval nepokrývá skute£nou pozorovanou hodnotu. Nejvíce nep°esností pozorujeme v letech 2020 a 2021, kde po£ty zem°elých významn¥ stouply i kv·li pandemii Covid-19.

Celkov¥ volíme jako nální model pro muºe PCA model s p¥ti hlavními komponentami a náhodnou procházkou s driftem, protoºe je nejlep²í podle MAPE a hodnota intervalového skóre není p°íli² vysoká. U ºen volíme PCA model s jednou hlavní komponentou a náhodnou procházkou s driftem ze stejného d·vodu.

Na obrázcích 7.45 a 7.46 si m·ºeme prohlédnout pozorované hodnoty $y_{x,t}$ v£etn¥ jejich budoucích hodnot a intervalového bootstrap odhadu pro vybranou v¥kovou skupinu (75 let) získaných nálními modely.

Obrázek 7.45: Pozorované hodnoty $y_{x;t}$ v£etn¥ jejich budoucích hodnot (získaných PCA-RWD modelem s 5 hlavními komponentami) a intervalového bootstrap odhadu pro vybranou v¥kovou skupinu 75 let muº·.

Obrázek 7.46: Pozorované hodnoty $y_{x;t}$ v£etn¥ jejich budoucích hodnot (získaných PCA-RWD modelem s 1 hlavní komponentou) a intervalového bootstrap odhadu pro vybranou v¥kovou skupinu 75 let ºen.

Pro nální modely je²t¥ uve¤me rozd¥lení po£tu úmrtí v roce 2021 a v predikovaných rocích 2030, 2040 a 2050 (Obrázky 7.47 a 7.48).

Obrázek 7.47: Rozd¥lení po£tu úmrtí muº· v letech 2021, 2030, 2040 a 2050 pomocí PCA5-RWD modelu.

Obrázek 7.48: Rozd¥lení po£tu úmrtí ºen v letech 2021, 2030, 2040 a 2050 pomocí PCA1-RWD modelu.

7.5 Ocen¥ní ºivotních smluv

S vybranými nálními modely dopo£teme pravd¥podobnosti p°e ºit¶ $_x$ a ceny ºivotních smluv p°edstavených v kapitole 6. U odloºených smluv p°edpokládáme $m = 10$, u do£asných smluv beremen $= 20$. Nech´ dále platí pro pojistné pln¥ní $B = 10\,000$ a pro konstatní úrokovou míru $= 0;03$. Na obrázcích 7.49, 7.50, 7.51 a 7.52 máme k nahlédnutí vývoj cen ºivotních smluv v letech 2021, 2030, 2040 a 2050. Je patrné, ºe ceny produkt· budou r·st, coº je zp·sobeno tím, ºe o£ekáváme, ºe se populace bude doºívat vy²²ích v¥k·.

Obrázek 7.49: Do£asný okamºitý d·chod v letech 2021, 2030, 2040 a 2050.

Obrázek 7.50: Doºivotní okamºitý d·chod v letech 2021, 2030, 2040 a 2050.

Obrázek 7.51: Do£asný odlo^{o}ený d·chod v letech 2021, 2030, 2040 a 2050.

Obrázek 7.52: Do^{o}ivotní odlo^{o}ený d·chod v letech 2021, 2030, 2040 a 2050.

M¥jme nyní portfolio o 10 000 smlouvách do^{o}ivotního okam^{o}ité d·chodu, kde polovina jsou mu^{o}i a polovina ^{o}eny, a jedinci jsou rovnom¥rn¥ ve v¥cích 50, 55, 60, 65 a 70 let. Abychom ukázali d·le^{o}itost správné predikce po£tu úmrtí, nastíníme si, jaký dopad to m·^{o}e mít. V tabulce 7.4 máme uvedeny ceny daného produktu pro vybrané v¥kové skupiny v závislosti na nálních modelech (PCA5-RWD pro mu^{o}e a PCA1-RWD pro ^{o}eny) a na LC-SVD modelu v roce 2025. Celkový rozdíl na na²em portfoliu v roce 2025 by byl 63 179 277 K£.

V¥k	PCA5-RWD muºi	LC-SVD muºi	PCA1-RWD ºeny	LC-SVD ºeny
50	170 931	160 706	196 755	195 154
55	151 458	140 711	178 226	176 491
60	131 724	120 440	157 996	156 162
65	112 175	100 675	136 195	134 328
70	92 228	81 647	113 067	111 263

Tabulka 7.4: Cena smlouvy doºivotního okamºitého d·chodu pro vybrané v¥kové skupiny v závislosti na nálních modelech a LC-SVD modelu.

Záv¥r

V této práci jsme se zabývali modelováním úmrtnostních m¥r a po£tu úmrtí. Nejprve jsme zavedli základní pojmy demogra ckého modelu a úmrtnostních tabulek. Následn¥ jsme se v¥novali Leeovu Carterovu modelu, kde jsme si model denovali, uvedli jsme t°i r·zné metody pro odhad parametr· a popsali jsme, jak pomocí ARIMA model· predikovat budoucí hodnoty. Obdobn¥ jsme nabyli teoretické poznatky o Renshawovu-Habermanovu modelu, kde mnoho vlastností plyne z Leeova-Carterova modelu. P°edstavili jsme si rovn¥º metodu od Shanga s Habermanem, jeº vyuºívá analýzu kompozi£ních dat. Zde jsme si popsali také neparametrické bootstrap metody pro intervalové odhady v²ech model·.

V praktické £ásti jsme na £eská data úmrtí muº· a ºen aplikovali v²echny modely, odhadli jsme je a p°edpov¥d¥li budoucí hodnoty. Zjistili jsme, ºe v²echny modely popisovaly data velmi dob°e, nebo´ ºádný z model· nem¥l hodnotu koe - cientu determinace niº²í neº 99.96 %. Nejv¥t²í nep°esnosti model· byly pro v¥ky 100-105, coº bylo zp·sobené chyb¥jícími daty pro tyto v¥ky aº do roku 2006. Nejv¥t²í rozdíly jsme pozorovali u model· s Poissonovou metodou, nebo´ tento algoritmus pracoval s nedostupnými daty D_x a N_x: Z d·vodu nedostupnosti n¥kterých dat a zárove¬ nedostatku dat pro tyto v¥kové skupiny bychom výsledky model· t¥chto v¥kových skupin nebrali p°íli² váºn¥.

Na základ¥ st°ední absolutní procentuální chyby jsme na²li nejlep²í modely - pro data muº· model PCA s p¥ti hlavními komponentami, kde pro predikci jsme vyuºili náhodnou procházku s driftem. Pro ºeny vy²el nejlépe PCA model s jednou hlavní komponentou a s metodou náhodné procházky s driftem pro predikování.

Zjistili jsme, ºe nový model vyuºívající analýzy kompozi£ních dat poskytuje velmi dobré výsledky v porovnání s Leeovým Carterovým a Renshawovým-Habermanovým modelem. Renshaw·v Haberman·v model vylep²uje predikce Leeova Carterova modelu, vyjma modelu s Poissonovou metodou pro ºeny. P°idáním dal²ích hlavních komponent do PCA modelu se predikce zlep²ily pouze na datech pro muºe krom¥ modelu s ARIMA modelem pro predikci, nejvýrazn¥j²í zlep²ení pozorujeme u náhodné procházky s driftem.

Nejlep²í model podle intervalového skóre je pro muºe trochu p°ekvapiv¥ PCA model s p¥ti hlavními komponentami s náhodnou procházku bez driftu a pro ºeny PCA model s jednou hlavní komponentou a náhodnou procházkou bez driftu. PCA modely s náhodnou procházkou bez driftu mají ale velmi nep°esné bodové odhady, a proto i p°es nízké intervalové skóre nejsou vhodnými modely. Renshaw·v Haberman·v model je i zde obecn¥ lep²í neº Lee·v Carter·v model.

Jako nální modely jsme vybrali PCA model s p¥ti hlavními komponentami a náhodnou procházkou s driftem pro muºe a PCA model s jednou hlavní komponentou a s náhodnou procházkou s driftem pro ºeny. Na základ¥ t¥chto model· jsme ur£ili cenu ºivotních smluv - do£asný okamºitý d·chod, doºivotní okamºitý d·chod, do£asný odloºený d·chod a doºivotní odloºený d·chod v roce 2021 a v budoucnu. Pozorujeme, ºe cena t¥chto smluv poroste, coº je zp·sobené postupným prodluºováním délky ºivota populace.

Na záv¥r jsme na konkrétním portfoliu o 10 000 smlouvách ukázali d·leºitost p°esnosti budoucích po£t· úmrtí, kdyº rozdíl mezi predikcemi na základ¥ nálních model· a LC-SVD modelu £inil p°es 63 milion· korun.

Seznam pouºité literatury

Aitchison, J. (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44, 139 160. doi: 10.1111/j.2517-6161.1982.tb01195.x.

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Chapman and Hall, London.

Alho, J. (2000). the lee-carter method for forecasting mortality, with various extensions and applications , ronald lee, january 2000. North American Actuarial Journal, 4, 80 93. doi: 10.1080/10920277.2000.10595883.

Barto, L. a T·ma, J. (2017). Lineární algebra. URL https://www2.karlin.mff.cuni.cz/~barto/LinAlg/skripta_la5.pdf . On-line skripta MFF UK.

Ben-Israel, A. (1966). A newton-raphson method for the solution of systems of equations. Journal of Mathematical analysis and applications, 15(2), 243 252.

Box, G. E. P. a Jenkins, G. M. (1970). Time series analysis forcasting and control. Holden-Day, pages 126 173.

Brouhns, N., Denuit, M. a Vermunt, J. (2002). A poisson log-bilinear regression approach to the construction of projected life tables. Insurance: Mathematics and Economics, 31, 373 393. doi: 10.1016/S0167-6687(02)00185-3.

Cipra, T. (1990). Matematické metody demogra e a poji²t¥ní. SNTL. ISBN 80-03-00222-2.

Cipra, T. (1999). Pojistná matematika - teorie a praxe. Ekopress, Praha. ISBN 80-86119-17-3.

Gneiting, T. a Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102, 359 378. doi: 10.1198/016214506000001437.

Good, I. (1969). Some applications of the singular decomposition of a matrix. Technometrics, 11, 823 831. doi: 10.1080/00401706.1969.10490741.

Horváth, L. a Kokoszka, P. (2012). Inference for functional data with applications, volume 200. Springer Science & Business Media.

Hyndman, R. a Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for r. Journal of Statistical Software, 26. doi: 10.18637/jss.v027.i03.

Hyndman, R. a Ullah, S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. Computational Statistics & Data Analysis, 51, 4942 4956. doi: 10.1016/j.csda.2006.07.028.

Lee, R. a Carter, L. (1992). Modeling and forecasting u. s. mortality. J. Amer. Statist. Assoc, 87. doi: 10.2307/2290201.

Oeppen, J. a kol. (2008). Coherent forecasting of multiple-decrement life tables: a test using japanese cause of death data. URL https://dugi-doc.udg.edu/bitstream/handle/10256/742/Oeppen?sequence=1 .

Ord, K., Koehler, A. B. a Snyder, R. D. (1997). Estimation and prediction for a class of dynamic nonlinear statistical models. Journal of the American Statistical Association, 92(439), 1621 1629. doi: 10.1080/01621459.1997.10473684.

Renshaw, A. a Haberman, S. (2003). Lee-carter mortality forecasting with age-speci c enhancement. Insurance: Mathematics and Economics, 33, 255 272. doi: 10.1016/S0167-6687(03)00138-0.

Shang, H. L. a Haberman, S. (2019). Forecasting age distribution of death counts: an application to annuity pricing. Annals of Actuarial Science, 14, 1 20. doi: 10.1017/S1748499519000101.

ƒeský statistický ú°ad (2023). ƒeský statistický ú°ad. Online. URL https://www.czso.cz/ . Datum p°ístupu: 20. dubna 2023.

Seznam obrázk·