



**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 kapitulu, kde se modelují česká data úmrtnosti zvláště 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	3
Úvod	6
1 Úmrtnost a základní pojmy	7
1.1 Demografický model	7
1.2 Úmrtnostní tabulky	9
2 Leeův–Carterův model	12
2.1 Popis modelu	12
2.2 Odhad modelu	13
2.2.1 SVD a metoda nejmenších čtverců	13
2.2.2 Poissonova metoda s bilineárními prediktory	16
2.2.3 Alternativní (regresní) metoda	17
2.3 Predikce	18
3 Renshawův–Habermanův model	20
3.1 Odhad modelu	20
3.1.1 SVD a metoda nejmenších čtverců	20
3.1.2 Poissonova metoda s bilineárními prediktory	22
3.2 Predikce	23
4 Analýza kompozičních dat	24
4.1 Model a odhad parametrů	24
4.2 Predikce	27
4.2.1 Bodová předpověď	27
4.2.2 Intervalová predikce	29
5 Kvalita modelů	32
6 Ocenění životních smluv	34
7 Aplikace na reálná data	36
7.1 Představení dat	36
7.2 Odhady parametrů modelů	40
7.2.1 LC model	40
7.2.2 RH model	46
7.2.3 PCA model	50
7.3 Predikce	53
7.3.1 LC model	53
7.3.2 RH model	55
7.3.3 PCA model	58
7.4 Porovnání modelů	62
7.4.1 Koefficient determinace	62
7.4.2 MAPE a intervalové skóre	62
7.5 Ocenění životních smluv	65

Závěr	68
Seznam použité literatury	69
Seznam obrázků	71

Seznam symbolů

x	<i>Diskrétní věk</i> Nabývá hodnot x_1, \dots, x_m .
t	<i>Diskrétní čas</i> Nabývá hodnot t_1, \dots, t_n .
T_x	<i>Zbývající délka života pro jedince ve věku x</i> Jedná se o spojitou náhodnou veličinu na pravděpodobnostním prostoru $(\Omega, \mathcal{F}, \mathbb{P})$.
${}_tq_x$	<i>Pravděpodobnost, že jedinec ve věku x zemře v následujících t rocích</i> Jedná se o pravděpodobnost $P(T_x \leq t)$, což také odpovídá distribuční funkci náhodné veličiny T_x . Můžeme značit též $F_x(t)$.
${}_tp_x$	<i>Pravděpodobnost, že jedinec ve věku x nezemře v následujících t rocích</i> Jedná se o pravděpodobnost $P(T_x > t)$, což také odpovídá funkci přežití $S_x(t)$.
η_x	<i>Intenzita úmrtnosti ve věku x</i>
$D_{x,t}$	<i>Počet úmrtí jedinců mezi věky x a $x+1$ v čase t</i>
$d_{x,t}$	<i>Tabulkový počet zemřelých jedinců z hypotetické generace mezi věky x a $x+1$ v čase t</i>
D	<i>Matice počtu úmrtí</i> 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 $(m \times n)$.
$P_{x,t}$	<i>Střední stav populace ve věku x v čase t</i> Jedná se o počet jedinců ve věku x k 1. červenci roku t .
$l_{x,t}$	<i>Tabulkový počet dožívajících se věku x v čase t</i>
$L_{x,t}$	<i>Tabulkový počet žijících</i> Jedná se o součet dob všech jedinců prožitých mezi věky x a $x+1$.
$N_{x,t}$	<i>Centrální expozice ve věkovém intervalu $(x, x+1)$ v čase t</i> Jedná se o součet všech dob strávených jednotlivými jedinci populace v daném kalendářním roce t ve věkovém intervalu $(x, x+1)$.
a_x	<i>Neznámý parametr závislý na věku x, v LC modelu odhadnut průměrem úmrtnostních měř</i>
a	<i>Vektor neznámých parametrů a_x závislých na věku x</i> Tato proměnná je specifická pro každý věk, má rozměr $(m \times 1)$ a platí $\mathbf{a} = (a_{x_1}, \dots, a_{x_m})^T$.

A	<i>Matice neznámých parametrů a_x závislých na věku x</i> Jedná se o matici, která je složena z vektorů $(\mathbf{a}_1, \dots, \mathbf{a}_n)$. Matice má rozměr $(m \times n)$.
\mathbf{a}_x	<i>Vektor shodných a_x</i> Platí $\mathbf{a}_x = (a_x, \dots, a_x)^T$ a tento vektor má rozměr $(n \times 1)$.
b_x	<i>Neznámý parametr závislý na věku x vyjadřující citlivost na změnu v časovém indexu</i>
b	<i>Vektor neznámých parametrů b_x závislých na věku x</i> Tato proměnná je specifická pro každý věk, má rozměr $(m \times 1)$ a platí $\mathbf{b} = (b_{x_1}, \dots, b_{x_m})^T$.
k_t	<i>Časová řada úrovně úmrtnosti v čase t</i>
k	<i>Vektor časové řady úrovně úmrtnosti</i> Tato proměnná je specifická pro každý uvažovaný čas, má proto rozměr $(n \times 1)$ a platí $\mathbf{k} = (k_{t_1}, \dots, k_{t_n})^T$.
$m_{x,t}$	<i>Centrální míra úmrtnosti pro věk x v čase t</i>
m	<i>Matice centrálních měř úmrtností</i> Prvky této matice jsou $m_{x,t}$, což odpovídá centrální míře úmrtnosti pro věk x v čase t . Matice má rozměr $(m \times n)$.
$\mathbf{m}_{\cdot,t}$	<i>Vektor centrálních měř úmrtnosti v čase t</i> Platí $\mathbf{m}_{\cdot,t} = (m_{x_1,t}, \dots, m_{x_m,t})^T$.
$\mathbf{m}_{x,\cdot}$	<i>Vektor centrálních měř úmrtnosti ve věku x</i> Platí $\mathbf{m}_{x,\cdot} = (m_{x,t_1}, \dots, m_{x,t_n})^T$.
$\epsilon_{x,t}$	<i>Chybový člen modelu pro věk x v čase t</i>
ϵ	<i>Matice chybových členů modelu</i>
$\phi_{x,l}$	<i>l-tá hlavní komponenta pro věk x</i>
$\beta_{l,t}$	<i>Faktorové skóre l-té hlavní komponenty pro čas t</i>
$z_{x,t}$	<i>Transponovaná kompoziční data</i>
\hat{p}	<i>Modelový odhad parametru p</i>
\tilde{p}	<i>Budoucí hodnota parametru p</i>
\bar{p}	<i>Aritmetický průměr parametrů p_i</i>
$\frac{\partial}{\partial x_i} f(x)$	<i>Parciální derivace funkce $f(x)$ podle x_i</i>
$\frac{d}{dx} f(x), f'(x)$	<i>Derivace funkce $f(x)$</i>
O_{LS}	<i>Součet čtvercových chyb modelu</i>
$L(\beta)$	<i>Věrohodnostní funkce</i>

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

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 měř 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 modelem (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 bankovníctví, 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 $(\Omega, \mathcal{F}, \mathbb{P})$, která značí **zbývající délku života**. Pro tuto náhodnou veličinu T_x zavedme distribuční funkci

$$F_x(t) = P(T_x \leq t), \quad t \geq 0. \quad (1.1)$$

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

$$f_x(t) = F'_x(t) \quad t \geq 0, \quad (1.2)$$

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** $S_x(t)$:

$$S_x(t) = 1 - F_x(t) = P(T_x > t), \quad t \geq 0. \quad (1.3)$$

Nyní zavedeme aktuárské značení¹ pro již výše zmíněné pravděpodobnosti $F_x(t)$ a $S_x(t)$. Označme

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

¹Pokud 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:

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

kde (FA) jsme využili v předposlední rovnosti.

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

$$\begin{aligned} e_x^o &= \mathbb{E}(T_x) = \int_0^\infty t f_x(t) dt = [t F_x(t)]_{t=0}^\infty - \int_0^\infty F_x(t) dt = \int_0^\infty 1 dt - \int_0^\infty F_x(t) dt \\ &= \int_0^\infty (1 - F_x(t)) dt = \int_0^\infty S_x(t) dt = \int_0^\infty {}_t p_x dt, \end{aligned} \tag{1.5}$$

kde ve třetí rovnosti jsme využili metodu per-partes.

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

$$\mu_x = \lim_{h \rightarrow 0_+} \frac{P(T_x \leq h)}{h} = \lim_{h \rightarrow 0_+} \frac{h q_x}{h}. \tag{1.6}$$

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

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

$${}_t p_x = \exp\left(-\int_0^t \mu_{x+s} ds\right). \tag{1.8}$$

Platí:

$$\begin{aligned} \mu_{x+t} &= \lim_{h \rightarrow 0_+} \frac{P(T_{x+t} \leq h)}{h} = \lim_{h \rightarrow 0_+} \frac{P(T_x - t \leq h | T_x > t)}{h} \\ &= \lim_{h \rightarrow 0_+} \frac{P(t < T_x \leq t + h)}{h P(T_x > t)} = \lim_{h \rightarrow 0_+} \frac{F_x(t + h) - F_x(t)}{h} \frac{1}{{}_t p_x} \\ &= \frac{F'_x(t)}{{}_t p_x} = \frac{f_x(t)}{{}_t p_x}, \end{aligned}$$

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:

$$\mu_{x+t} = \frac{F'_x(t)}{{}_t p_x} = \frac{1}{{}_t p_x} \frac{d}{dt} {}_t q_x = \frac{1}{{}_t p_x} \frac{d}{dt} (1 - {}_t p_x) = -\frac{1}{{}_t p_x} \frac{d}{dt} {}_t p_x = -\frac{d}{dt} \ln({}_t p_x), \tag{1.9}$$

neboť

$$\frac{d}{dt} \ln({}_t p_x) = \frac{1}{{}_t p_x} \frac{d}{dt} {}_t p_x.$$

Nyní již z 1.9 snadno získáme vztah 1.8.

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

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

Protože pravděpodobnosti ${}_t p_x$ a ${}_t q_x$ 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 zbývajícím délce života, a

$$0 \leq S_x < 1.$$

Aproximační předpoklady pak jsou:

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

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

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

a pro vztah mezi intenzitou úmrtnosti a pravděpodobností q_x platí:

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

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 měř 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 $x = 0, 1, 2, \dots, 105 +$, 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 expozicí² 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:

$$m_x^T = \frac{D_x}{P_x}. \quad (1.13)$$

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 specifčnosti 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:

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

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-

²Centrální 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 rozložena 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:

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

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:

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

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 $x + 1$, přispějí poměrem prožité doby v daném intervalu. Platí:

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

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

$$L_x = \frac{1}{2}(l_x + l_{x+1}) \quad x = 2, \dots, 104 \quad (1.18)$$

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:

$$T_x = \sum_{y=x}^{105+} L_y. \quad (1.19)$$

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:

$$e_x = \frac{T_x}{l_x}. \quad (1.20)$$

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:

$$m_x^{TG} = \frac{d_x}{L_x}. \quad (1.21)$$

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 nejnámějších 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 úmrtnosti Lee s Carterem modelují jako lineární funkci nepozorovaného, časově specifického indexu intenzity s parametry závisujícími 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 $x = x_1, \dots, x_m$ a dále pomocí nepozorovatelného indexu \mathbf{k} , jenž se liší s časem t . Tvar závislosti maticově¹ vyjádříme následovně:

$$\ln \mathbf{m} = \mathbf{A} + \mathbf{b}\mathbf{k}^T + \boldsymbol{\epsilon}, \quad (2.1)$$

což můžeme ekvivalentně vyjádřit rovněž jako

$$\ln m_{x,t} = a_x + b_x k_t + \epsilon_{x,t}, \quad x = x_1, \dots, x_m, \quad t = t_1, \dots, t_n. \quad (2.2)$$

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

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

Rovnice se také někdy udává ve tvaru exponenciály:

$$\mathbf{m} = e^{\mathbf{A} + \mathbf{b}\mathbf{k}^T + \boldsymbol{\epsilon}}, \quad (2.4)$$

kterou lze obdobně popsat vztahem:

$$m_{x,t} = e^{a_x + b_x k_t + \epsilon_{x,t}}, \quad x = x_1, \dots, x_m, \quad t = t_1, \dots, t_n. \quad (2.5)$$

Rovněž platí i vztah se sumami:

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

Ve všech případech chybový člen $\epsilon_{x,t}$ 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.

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

vyjadřuje celkový tvar úmrtností napříč věky a \mathbf{b} zaznamenává rychlost poklesu měr v závislosti na změně indexu \mathbf{k} .

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

$$\sum_{x=x_1}^{x_m} b_x = 1, \quad \sum_{t=t_1}^{t_n} k_t = 0, \quad (2.7)$$

kde x_1, \dots, x_m značí konkrétní uvažované věky a t_1, \dots, 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 $\epsilon_{x,t}$ budeme brát střední hodnotu této proměnné. Odtud pro dané x snadno získáme, že $a_x = \frac{1}{n} \sum_{t=t_1}^{t_n} \ln m_{x,t}$.

$$\begin{aligned} \sum_{t=t_1}^{t_n} \ln m_{x,t} &= \sum_{t=t_1}^{t_n} a_x + \sum_{t=t_1}^{t_n} b_x k_t, & x = x_1, \dots, x_m, \\ \sum_{t=t_1}^{t_n} \ln m_{x,t} &= n \cdot a_x + b_x \sum_{t=t_1}^{t_n} k_t, & x = x_1, \dots, x_m, \\ a_x &= \frac{1}{n} \sum_{t=t_1}^{t_n} \ln m_{x,t}, & x = x_1, \dots, x_m, \end{aligned}$$

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 \mathbf{a} , \mathbf{b} , \mathbf{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 \mathbf{m} a průměrů měr specifických pro jednotlivě věky \mathbf{A} .

Model budeme odhadovat ve dvou fázích. Nejprve odhadneme vektory parametrů \mathbf{a} , \mathbf{b} , \mathbf{k} a následně přeodhadneme vektor parametrů \mathbf{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 $x = x_1, \dots, x_m$ a pro všechny časy $t = t_1, \dots, t_n$. 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.

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

odkud dostáváme, díky 2.7, vektorový tvar

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

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 hodnotí r existují ortonormální báze $(\mathbf{v}_1, \dots, \mathbf{v}_n)$ prostoru \mathbb{R}^n a ortonormální báze $(\mathbf{u}_1, \dots, \mathbf{u}_m)$ prostoru \mathbb{R}^m tak, že*

$$\mathbf{M} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \dots + \sigma_r \mathbf{u}_r \mathbf{v}_r^T. \quad (2.9)$$

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

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

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

$$\hat{\mathbf{M}} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \dots + \sigma_s \mathbf{u}_s \mathbf{v}_s^T. \quad (2.10)$$

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

$$\mathbf{M} \approx \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T. \quad (2.11)$$

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

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

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

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

s předpokladem, že

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

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

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

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} předhadnou. Důvodem, proč se autoři nespokojí s aktuálními odhady, je, že aktuální odhady \mathbf{k} minimalizují chyby od logaritmu úmrtnostních měř, zatímco my bychom preferovali minimalizaci chyb od měř úmrtností. Minimalizace chyb od logaritmu úmrtnostních měř 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 předhadnutí parametru \mathbf{k} se vychází ze vztahu

$$m_{x,t} = \frac{D_{x,t}}{N_{x,t}}, \quad x = x_1, \dots, x_m, \quad t = t_1, \dots, t_n. \quad (2.14)$$

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:

$$\sum_{x=x_1}^{x_m} D_{x,t} = \sum_{x=x_1}^{x_m} N_{x,t} e^{\hat{a}_x + \hat{b}_x k_t}, \quad t = t_1, \dots, t_n. \quad (2.15)$$

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 předhadnutí 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:

$$\mathbb{E}(D_{x,t}) = N_{x,t} \exp(\eta_{x,t}), \quad x = x_1, \dots, x_m, \quad t = t_1, \dots, t_n, \quad (2.16)$$

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

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

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 $\text{Dev}(\mathbf{D}, \hat{\mathbf{D}})$.

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

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

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

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

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

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

znovu přepočítáme $\hat{\mathbf{D}}$ a rovněž spočítáme $\text{Dev}(\mathbf{D}, \hat{\mathbf{D}})$. 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 $\text{Dev}(\mathbf{D}, \hat{\mathbf{D}})$ konverguje, v opačném případě se vracíme do kroku 2. To, zda-li $\text{Dev}(\mathbf{D}, \hat{\mathbf{D}})$ konverguje, rozhodneme podle toho, zda úbytek v devianci nepřekročil předem zvolenou hodnotu. V opačném případě, kdy úbytek deviance překročí předem zvolenou hodnotu, odhady jsou 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 $x = x_1, \dots, x_m$ a pro všechny časy $t = t_1, \dots, t_n$. 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 $\epsilon_{x,t}$ bereme její střední hodnotu, získáme odhad pro vektor parametrů \mathbf{k} :

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

kde $\hat{\mathbf{a}}_x = (a_x, \dots, a_x)$ 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

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

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 $x = x_1, \dots, x_m$ 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:

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

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

$$\hat{\alpha} = \frac{k_{t_n} - k_{t_1}}{t_n - t_1}, \quad (2.25)$$

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

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

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:

$$\begin{aligned} \tilde{k}_{t_n+1} &= \hat{k}_{t_n} + \hat{\alpha}, \\ \tilde{k}_{t_n+s} &= \tilde{k}_{t_n+s-1} + \hat{\alpha}, \quad s = 2, \dots, S, \\ \tilde{k}_{t_n+s} &= \hat{k}_{t_n} + s\hat{\alpha}, \quad s = 1, \dots, S, \end{aligned} \quad (2.26)$$

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:

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

Vztah 2.27 lze díky 2.26 upravit do podoby:

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

Když máme odhadnuté budoucí míry úmrtnosti, jsme už pouze krok od dočtení 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:

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

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

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

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 nedostatku 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:

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

nebo ekvivalentně ve tvaru exponenciály:

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

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

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

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

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

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

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

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:

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

Chceme získat potřebné odhady metodou nejmenších čtverců:

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

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

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

Společně s první podmínkou z 3.3 dostáváme

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

s předpokladem, že

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

Získané odhady nesplňují normalizační podmínky 3.3, proto je třeba provést analogické úpravy jako v 2.13. Zbývá předodhadnout 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:

$$\sum_{x=x_1}^{x_m} D_{x,t} = \sum_{x=x_1}^{x_m} N_{x,t} e^{\hat{a}_x + \hat{b}_x^{(1)} \hat{k}_t^{(1)} + \hat{b}_x^{(2)} \hat{k}_t^{(2)}}, \quad t = t_1, \dots, t_n. \quad (3.9)$$

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

Po předodhadnutí 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

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

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 $\text{Dev}(\mathbf{D}, \hat{\mathbf{D}})$.

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 $\hat{\mathbf{k}}^{(1)}$:

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

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

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

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 $\text{Dev}(\mathbf{D}, \hat{\mathbf{D}})$. Vzorce pro aktualizace parametrů vychází opět z iteračního Newtona–Raphsonova algoritmu, který je použit na soustavu rovnic, jež jsme získali položením parciálních derivací logaritnické věrohodnostní funkce rovné nule.

Ve třetím kroku algoritmus končí, pokud $\text{Dev}(\mathbf{D}, \hat{\mathbf{D}})$ 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 předhadnout 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

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

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

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

což lze ekvivalentně zapsat jako

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

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 $\mathbf{d} = (d_1, \dots, d_k)$ velikosti k , jehož všechny složky jsou kladné a jejich součet je roven konstantě, neboli:*

$$d_i > 0 \quad \forall i = 1, \dots, k, \quad \sum_{i=1}^k d_i = c. \quad (4.1)$$

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:

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

pro $x = x_1, \dots, x_m$. 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 $f_{x,t}$:

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

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 :

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

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

$$z_{x,t} = \ln\left(\frac{f_{x,t}}{g_t}\right), \quad x = x_1, \dots, x_m, \quad t = t_1, \dots, t_n. \quad (4.5)$$

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 $\phi_{x,l}$ a jejich odhadnutých faktorových skóre $\beta_{l,t}$:

$$z_{x,t} = \sum_{l=1}^L \beta_{l,t} \phi_{x,l} + \epsilon_{x,t}, \quad x = x_1, \dots, x_m, \quad t = t_1, \dots, t_n, \quad (4.6)$$

kde L značí počet hlavních komponent rozkladu a chybový člen modelu $\epsilon_{x,t}$ 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 $\hat{\phi}_{x,l}$ a faktorové skóre $\hat{\beta}_{l,t}$:

$$\hat{\beta}_l = \mathbf{u}_l \quad \hat{\phi}_l = \sigma_l \mathbf{v}_l, \quad l = 1, \dots, L, \quad (4.7)$$

kde $\hat{\beta}_l$ je vektor o velikosti n a $\hat{\phi}_l$ 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:

$$L = \arg \min_{L: L \geq 1} \left\{ \sum_{l=1}^L \lambda_l / \sum_{l=1}^r \lambda_l \geq \delta \right\}, \quad (4.8)$$

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 $k \neq l$ (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}_{x,t_n+h} získáme vztahem:

$$\tilde{z}_{x,t_n+h} = \sum_{l=1}^L \tilde{\beta}_{l,t_n+h} \hat{\phi}_{x,l}, \quad x = x_1, \dots, x_m, \quad h = 1, \dots, H, \quad (4.9)$$

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:

$$\tilde{f}_{x,t_n+h} = \frac{\exp(\tilde{z}_{x,t_n+h})}{\sum_{x=x_1}^{x_m} \exp(\tilde{z}_{x,t_n+h})}, \quad x = x_1, \dots, x_m, \quad h = 1, \dots, H, \quad (4.10)$$

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

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

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:

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

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 skóreů $\tilde{\beta}_{l,t_n+h}$, 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 $\hat{\phi}_{x,l}$ a jejich faktorových skóreů $\hat{\beta}_{l,t}$. Na čase však závisí pouze faktorové skóre $\hat{\beta}_{l,t}$. Pro pevné l máme jedno-rozměrnou 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 předchozí odhadované hodnoty. Tím můžeme docílit toho, že nedávné hodnoty zasahují do odhadu více než hodnoty získané dříve v minulosti.

Zkratka ETS vychází ze tří slov - chyba (error), trend a sezónnost. Exponenciální vyhlazování totiž modeluje tyto tři veličiny a máme několik možností, jak to můžeme provést. Chybu lze modelovat aditivně (A) nebo multiplikativně (M). Trend můžeme rovněž modelovat aditivně (A), multiplikativně (M) nebo ho lze neuvažovat (N). K aditivnímu a multiplikativnímu trendu máme vždy ještě jednu alternativu 'damped' (A_d), (M_d), která uvažuje parametr ϕ navíc. Sezónnost je možné modelovat aditivně (A), multiplikativně (M) nebo je možné ji neuvažovat (N). Celkem tedy $E \in \{A, M\}$, $T \in \{N, A, A_d, M, M_d\}$ a $S \in \{N, A, M\}$, což nám dává na výběr z 30 modelů, neboť různé přístupy lze kombinovat.

ETS modely pro časové řady y_t jsou založené na hladině řady ℓ_t , trendu řady b_t a sezónnosti s_{t-m} , kde m je délka sezónnosti. Hladina ℓ_t je lineární kombinace pozorované časové řady očištěné od sezónnosti a předchozí hodnoty hladiny s trendem. Trend je lineární kombinace příbytku hladiny a předchozí hodnoty trendu. Sezónnost je lineární kombinace pozorované časové řady očištěné od hladiny a trendu a předchozí hodnoty ve stejném sezónním období.

Bodová předpověď je složením (sečtením, vynásobením či kombinací) složek hladiny, trendu a sezónnosti. Přehlednou tabulku všech ETS modelů nalezneme v díle Hyndman a Khandakar (2008) na straně 4. Bodová predikce pro jednotlivé modely je vždy stejná, ať už chybu modelujeme aditivně, nebo multiplikativně, proto se v literatuře Hyndman a Khandakar (2008) rozlišuje pouze trend a sezónnost.

Označme stavový vektor $\mathbf{x}_t = (\ell_t, b_t, s_t, s_{t-1}, \dots, s_{t-m+1})^T$. Obecně lze všechny ETS modely zapsat ve tvaru:

$$\begin{aligned} y_t &= w(\mathbf{x}_{t-1}) + r(\mathbf{x}_{t-1})\epsilon_t, \\ \mathbf{x}_t &= f(\mathbf{x}_{t-1}) + g(\mathbf{x}_{t-1})\epsilon_t, \end{aligned} \quad (4.12)$$

kde ϵ_t je gaussovský bílý šum¹ a funkce $w(\mathbf{x}_{t-1})$ je podmíněná střední hodnota $\mu_t = \mathbb{E}[y_t | \mathbf{x}_{t-1}]$.

Vhodně zvolené funkce $f(\mathbf{x}_{t-1})$, $g(\mathbf{x}_{t-1})$ a $r(\mathbf{x}_{t-1})$ určují tvar ETS modelu. Pokud uvažujeme aditivní chybu, pak $r(\mathbf{x}_{t-1}) = 1$, pokud naopak uvažujeme multiplikativní strukturu chyby, pak $r(\mathbf{x}_{t-1}) = \mu_t$.

Když máme definované výše zmíněné rekurzivní modely, je zapotřebí ještě odhadnout počáteční hodnoty - vektor $f(\mathbf{x}_0)$ a parametry α , β^* , γ , ϕ . Ty budeme odhadovat metodou maximální věrohodnosti. Pro podrobnosti ohledně odhadů metodou maximální věrohodnosti odkazujeme na články Hyndman a Khandakar (2008) a Ord a kol. (1997).

Nyní jsme schopni zkonstruovat všechny modely a zbývá vybrat ten nejvhodnější. Součástí ETS metody často bývá vybrání modelu podle Akaikeho informačního kritéria (AIC), které je definováno jako:

$$\text{AIC} = -2 \ln[L(\Theta)] + 2(q), \quad (4.13)$$

kde $L(\Theta)$ je věrohodnostní funkce, Θ je vektor parametrů, které je potřeba odhadnout a q je počet parametrů, které mají být odhadnuty.

¹Gaussovský bílý šum je klasický bílý šum, kde navíc platí, že každá náhodná veličina má normální rozdělení s nulovou střední hodnotou a kladným konečným rozptylem σ^2 .

AIC není samozřejmě jediným možným výběrovým kritériem, zvolit lze i například Bayesovo informační kritérium (BIS). My však budeme využívat AIC a nejlepší model bude ten, jehož AIC hodnota bude nejnižší.

Náhodná procházka

Náhodná procházka s driftem je časová řada $X_t, t = 0, 1, \dots, n$, která je definována jako

$$X_{t+1} = \tau + X_t + \epsilon_{t+1}, \quad t = 0, \dots, n-1, \quad (4.14)$$

kde τ je konstanta, která označuje drift a ϵ_{t+1} je gaussovský bílý šum, který chápeme jako chybový člen v čase $t+1$.

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

$$\tilde{X}_{n+h} = \mathbb{E}[X_{n+h}|X_1, \dots, X_n] = h\tau + X_n, \quad h = 1, \dots, H. \quad (4.15)$$

Rozptyl budoucích hodnot spočítáme

$$\text{var}[\tilde{X}_{n+h}] = \text{var}[X_{n+h}|X_1, \dots, X_n] = h\sigma^2, \quad (4.16)$$

kde σ^2 je rozptyl gaussovského bílého šumu $\{\epsilon_t\}$. Pokud uvažujeme náhodnou procházku bez driftu, znamená to, že parametr $\tau = 0$. Díky 4.15 a 4.16 platí:

$$\begin{aligned} \tilde{X}_{n+h} &= X_n, \\ \text{var}[\tilde{X}_{n+h}] &= h\sigma^2. \end{aligned} \quad (4.17)$$

4.2.2 Intervalová predikce

Často mohou být užitečné také intervalové predikce. Jedná se o interval, ve kterém s danou pravděpodobností leží skutečná hodnota budoucího pozorování. V našem případě se tedy budeme snažit odpovědět na otázku, kde bude ležet s pravděpodobností 0.95 budoucí hodnota počtu úmrtí v roce t a ve věku x ?

Existuje mnoho metod, kterými lze najít predikční intervaly. Jednou z nich je i metoda bootstrap, která se hodí použít v případech, kdy nemáme dostatek dat nebo když je obtížné či nemožné odvodit teoretické statistiky náhodných veličin. V článku Shang a Haberman (2019) je navržena konkrétní verze neparametrické bootstrap metody, kterou zde více popíšeme.

Bootstrap metoda

Základní myšlenka bootstrap metody je generování náhodných výběrů s opakováním ze stávajících dat. Tím získáme velké množství dat, pomocí nichž nalezneme intervalový odhad budoucích hodnot.

Definice 2. *Mějme posloupnost prvků $\{x_i\}_{i=1}^n$. Náhodný výběr s vrácením o velikosti m z posloupnosti $\{x_i\}_{i=1}^n$ se rozumí vektor (y_1, \dots, y_m) , kde y_j je náhodná hodnota x_i z posloupnosti $\{x_i\}_{i=1}^n$ pro $j = 1, \dots, m$.*

Máme dva zdroje chyby - chyba vzniklá při rozkladu matice \mathbf{Z} na hlavní komponenty a chyba vzniklá při predikci. My chceme bootstrapem predikovat parametry \tilde{z}_{x,t_n+h} . Z formule 4.9 vidíme, že musíme predikovat nové hodnoty pro $\tilde{\beta}_{l,t_n+h}$.

Upravme zde ještě značení. Budoucí hodnoty jsou totiž predikce, které jsou podmíněné známými hodnotami. Hodnoty v čase $t_n + h$, $h = 1, \dots, H$ podmíněné hodnotami do času t_n budeme značit $\tilde{\beta}_{l,t_n+h|t_n}$ a platí:

$$\tilde{\beta}_{l,t_n+h|t_n} = \mathbb{E}[\beta_{t_n+h} | \beta_{t_1}, \dots, \beta_{t_n}]. \quad (4.18)$$

Postup pro sestavení bootstrap odhadů je následující:

1. Mějme časovou řadu faktorových skóru $\{\hat{\beta}_{l,t_1}, \dots, \hat{\beta}_{l,t_n}\}$ pro $l = 1, \dots, L$. Definujme chybu predikce h -kroků dopředu v čase t pro l -té faktorové skóre jako

$$\hat{\nu}_{l,h,t} = \hat{\beta}_{l,t} - \tilde{\beta}_{l,t|t-h}, \quad h = 1, \dots, H, \quad t = t_{h+1}, \dots, t_n. \quad (4.19)$$

Bootstrap hodnoty parametru $\hat{\nu}_{l,h,*}^{(b)}$ získáme náhodným výběrem s vracením z $\{\hat{\nu}_{l,h,t}\}_{t=t_{h+1}}^{t_n}$.

2. Když máme spočítány bootstrap hodnoty parametru $\hat{\nu}_{l,h,*}^{(b)}$, bootstrap hodnoty parametru $\tilde{\beta}_{l,t_n+h|t_n}^{(b)}$ získáme pouhým dosazením ze vztahu:

$$\tilde{\beta}_{l,t_n+h|t_n}^{(b)} = \tilde{\beta}_{l,t_n+h|t_n} + \hat{\nu}_{l,h,*}^{(b)}, \quad b = 1, \dots, B, \quad (4.20)$$

kde $B = 1000$, $\tilde{\beta}_{l,t_n+h|t_n}$ je budoucí hodnota časové řady získaná standardním predikčním algoritmem a $\hat{\nu}_{l,h,*}^{(b)}$ jsou bootstrap hodnoty z bodu 1. Tím jsme pokryli první zdroj chyby.

3. Nyní se musíme vypořádat s chybami $\epsilon_{x,n+h}$ vzniklými při předpovědi budoucích hodnot. Platí

$$\hat{\epsilon}_{x,t} = z_{x,t} - \sum_{l=1}^L \hat{\beta}_{l,t} \hat{\phi}_{x,l}, \quad x = x_1, \dots, x_m, \quad t = t_1, \dots, t_n. \quad (4.21)$$

Bootstrap hodnoty parametru $\hat{\epsilon}_{x,t_n+h}^{(b)}$ pro $h = 1, \dots, H$ získáme náhodným výběrem s vracením z $\{\hat{\epsilon}_{x,t}\}_{t=t_1}^{t_n}$.

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

$$\tilde{z}_{x,t_n+h|t_n}^{(b)} = \sum_{l=1}^L \tilde{\beta}_{l,t_n+h|t_n}^{(b)} \hat{\phi}_{x,l} + \hat{\epsilon}_{x,t_n+h}^{(b)}, \quad b = 1, \dots, B, \quad (4.22)$$

kde $B = 1000$.

5. S takto vytvořenými hodnotami $\tilde{z}_{x,t_n+h|t_n}^{(b)}$ provedeme stejné inverzní transformace 6 a 7, čímž získáme bootstrap hodnoty budoucích tabulkových počtů zemřelých $\tilde{d}_{x,t_n+h|t_n}^{(b)}$.

6. Nyní chceme vytvořit intervalový odhad pro budoucí hodnoty tabulkového počtu zemřelých s pravděpodobností pokrytí 95 %. Potřebujeme tedy nalézt dolní a horní meze pro tento interval, které nalezneme jako 0.025 a 0.975 kvantily z $\{\tilde{d}_{x,t_n+h|t_n}^{(1)}, \tilde{d}_{x,t_n+h|t_n}^{(2)}, \dots, \tilde{d}_{x,t_n+h|t_n}^{(B)}\}$ pro $x = x_1, \dots, x_m$ a pro zvolené $h = 1, \dots, H$.

Bootstrap metoda pro LC a RH modely

Tato verze bootstrapu lze aplikovat rovněž na LC model a RH model. Budeme využívat stejné značení, které jsme zavedli při bootstrap metodě pro CoDa metodu. Analogicky aplikujeme bootstrap metodu:

1. Mějme časovou řadu indexů $\{\hat{k}_1^{(i)}, \hat{k}_2^{(i)}, \dots, \hat{k}_n^{(i)}\}$ pro $i = 1, 2$. V případě LC modelu pouze pro $i = 1$. Definujme chybu predikce h -kroků dopředu v čase t pro i -tý index $k_t^{(i)}$ jako

$$\hat{\nu}_{i,h,t} = \hat{k}_t^{(i)} - \tilde{k}_{t|t-h}^{(i)}, \quad h = 1, \dots, H, \quad t = t_{h+1}, \dots, t_n. \quad (4.23)$$

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

2. Když máme spočítány bootstrap hodnoty parametru $\hat{\nu}_{i,h,*}^{(b)}$, bootstrap hodnoty parametru $[\tilde{k}_{t_n+h|t_n}^{(i)}]^{(b)}$ získáme pouhým dosazením ze vztahu:

$$[\tilde{k}_{t_n+h|t_n}^{(i)}]^{(b)} = \tilde{k}_{t_n+h|t_n}^{(i)} + \hat{\nu}_{i,h,*}^{(b)}, \quad b = 1, \dots, B, \quad (4.24)$$

kde $B = 1000$, $\tilde{k}_{t_n+h|t_n}^{(i)}$ je budoucí hodnota časové řady získaná standardním predikčním algoritmem a $\hat{\nu}_{i,h,*}^{(b)}$ je bootstrap hodnota z bodu 1. Tím jsme pokryli první zdroj chyby.

3. Nyní se musíme vypořádat s chybami ϵ_{x,t_n+h} vzniklými při předpovědi budoucích hodnot. Platí

$$\hat{\epsilon}_{x,t} = \ln m_{x,t} - \ln \hat{m}_{x,t}, \quad x = x_1, \dots, x_m, \quad t = t_1, \dots, t_n. \quad (4.25)$$

Bootstrap hodnoty parametru $\hat{\epsilon}_{x,t_n+h}^{(b)}$ pro $h = 1, \dots, H$ získáme náhodným výběrem s vracením z $\{\hat{\epsilon}_{x,t}\}_{t=t_1}^{t_n}$.

4. Pomocí dvou nezávislých bootstrap metod jsme získali bootstrap hodnoty $[\tilde{k}_{t_n+h|t_n}^{(i)}]^{(b)}$ a $\hat{\epsilon}_{x,t_n+h}^{(b)}$ pro $b = 1, \dots, B$. Bootstrap hodnoty $\ln \tilde{m}_{x,t_n+h|t_n}^{(b)}$ do počítáme pouze dosazením do vzorce:

$$\ln \tilde{m}_{x,t_n+h|t_n}^{(b)} = \hat{a}_x + \sum_{i=1}^2 [\tilde{k}_{t_n+h|t_n}^{(i)}]^{(b)} \hat{b}_x^{(i)} + \hat{\epsilon}_{x,t_n+h}^{(b)}, \quad b = 1, \dots, B, \quad (4.26)$$

kde $B = 1000$.

5. Na závěr spočítáme bootstrap hodnoty budoucích tabulkových počtů zemřelých $\tilde{d}_{x,t_n+h|t_n}^{(b)}$ analogicky dle 2.28 a 2.29.
6. Sestrojení dolní a horní meze pro intervalový odhad je pak totožné jako v CoDa metodě.

5. Kvalita modelů

Modely popsané v předešlých kapitolách je zapotřebí porovnat a zjistit jejich kvalitu. Zde si popíšeme několik kritérií, které lze využívat k měření kvality modelů, porovnání modelů a k výběru nejvhodnějšího modelu.

Vhodným prostředkem pro grafické posouzení toho, jak dobře model vysvětluje data, jsou reziduální grafy. Nejčastěji se používají grafy rezidua versus vyrovnané hodnoty, kde požadujeme náhodné rozložení bodů v grafu. Pokud tomu tak není a pozorujeme v grafu nějaký vzor, znamená to, že v datech je nějaká systematická chyba, kterou model nedokázal vysvětlit. Někdy může být vhodnější použít graf se standardizovanými reziduy.

Kvalitu modelu můžeme rovněž měřit pomocí různých kritérií. Jedním z nich je koeficient determinace R^2 :

Definice 3. *Nechť Y_1, \dots, Y_n jsou empirická data a $\hat{Y}_1, \dots, \hat{Y}_n$ odhady jejich středních hodnot na základě modelu \mathcal{M} . Koeficient determinace R^2 definujeme jako poměr variability vysvětlené modelem vůči celkové variabilitě empirických dat:*

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

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

Koeficient determinace nabývá hodnot $[0,1]$, kde model s vyšší hodnotou R^2 lépe popisuje data než model s nižší hodnotou. R^2 aplikovaný na matici tabulkových počtů zemřelých má pak následující tvar:

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

kde \bar{d}_x je průměr tabulkových počtů zemřelých ve věku x přes čas t .

Modely tedy lze porovnat podle R^2 . Za lepší model považujeme ten, jenž má hodnotu R^2 vyšší. Další možností, jak určit který model je lepší, může být měření chyby predikce. Shang a Haberman (2019) navrhují použít střední absolutní procentuální chybu MAPE.

Definice 4. *Nechť $Y_1, \dots, Y_n, \dots, Y_{n+h}$ jsou empirická data a $\hat{Y}_1, \dots, \hat{Y}_n$ odhady jejich středních hodnot na základě modelu \mathcal{M} a $\tilde{Y}_{n+1}, \dots, \tilde{Y}_{n+h}$ jsou modelové predikce budoucích hodnot. Střední absolutní procentuální chybu modelu MAPE (Mean Absolute Percentage Error) definujeme jako*

$$MAPE = \frac{100}{h} \sum_{i=1}^h \left| \frac{Y_{n+i} - \tilde{Y}_{n+i}}{Y_{n+i}} \right|. \quad (5.3)$$

Postup je takový, že z pozorovaných dat odebereme h posledních pozorování a na základě zbylých dat se vytvoří predikce pro tato vynechaná pozorování. Poté se predikované hodnoty porovnají s odpovídajícími skutečnými hodnotami a určí se velikost chyby MAPE. Jako lepší model označíme ten model, jenž má chybu MAPE menší.

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

$$\text{MAPE} = \frac{100}{mh} \sum_{i=1}^h \sum_{x=x_1}^{x_m} \left| \frac{d_{x,t_n+i} - \tilde{d}_{x,t_n+i}}{d_{x,t_n+i}} \right|. \quad (5.4)$$

Dalším způsobem, jak modely porovnat, je zaměřit se na přesnost intervalových odhadů. Gneiting a Raftery (2007) navrhli intervalové skóre $\bar{\mathcal{S}}_\gamma$, 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 $100(1 - \gamma)$, který je symetrický, tedy spodní respektive horní mez intervalu (\tilde{d}_{x,t_n+i}^l resp. \tilde{d}_{x,t_n+i}^u) je $\gamma/2$ respektive $1 - \gamma/2$ kvantil. Gneiting a Raftery (2007) měří kvalitu odhadu predikčního intervalu s pravděpodobností pokrytí $100(1 - \gamma)$ pro věk x v čase $t_n + i$ jako délku intervalu + penalizaci v případě, že predikční interval nepokrývá skutečnou hodnotu:

$$\begin{aligned} \mathcal{S}_{\gamma,x,i}[\tilde{d}_{x,t_n+i}^l, \tilde{d}_{x,t_n+i}^u, d_{x,t_n+i}] &= (\tilde{d}_{x,t_n+i}^u - \tilde{d}_{x,t_n+i}^l) \\ &+ \frac{2}{\gamma}(\tilde{d}_{x,t_n+i}^l - d_{x,t_n+i})\mathbb{1}_{(d_{x,t_n+i} < \tilde{d}_{x,t_n+i}^l)} \\ &+ \frac{2}{\gamma}(d_{x,t_n+i} - \tilde{d}_{x,t_n+i}^u)\mathbb{1}_{(d_{x,t_n+i} > \tilde{d}_{x,t_n+i}^u)}, \end{aligned} \quad (5.5)$$

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

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

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

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

$$\bar{\mathcal{S}}_\gamma = \frac{1}{mh} \sum_{i=1}^h \sum_{x=x_1}^{x_m} \mathcal{S}_{\gamma,x,i}[\tilde{d}_{x,t_n+i}^l, \tilde{d}_{x,t_n+i}^u, d_{x,t_n+i}], \quad (5.6)$$

kde $\mathcal{S}_{\gamma,x,i}[\tilde{d}_{x,t_n+i}^l, \tilde{d}_{x,t_n+i}^u, d_{x,t_n+i}]$ je intervalové skóre s pravděpodobností pokrytí $100(1 - \gamma)$ pro věk x v čase $t_n + i$ definované v 5.5.

6. Ocenění životních smluv

Predikce tabulkového počtu zemřelých může mít řadu uplatnění jako například ocenění životních smluv. To se může hodit především pojišťovacím institucím či penzijním fondům. Pro výpočet ceny životní smlouvy je nezbytnou součástí znalost úmrtnosti populace v budoucnu.

V této kapitole se zaměříme na teoretický výpočet cen životních smluv v závislosti na predikovaných hodnotách tabulkových počtů zemřelých, které jsme získali v kapitolách 2, 3 a 4. Konkrétně se budeme zabývat smlouvami, u nichž je zapláceno jednorázové pojistné a výplaty pojistného plnění probíhají na roční bázi:

- Dočasný okamžitý důchod s výplatami na konci období (Temporary life annuity in arrear)
- Doživotní okamžitý důchod s výplatami na konci období (Whole life annuity in arrear)
- Odložený dočasný důchod s výplatami na konci období (Deferred temporary life annuity in arrear)
- Odložený doživotní důchod s výplatami na konci období (Deferred whole life annuity in arrear)

Pro výpočet těchto důchodů budeme potřebovat vyjádřit pravděpodobnost přežití jedince ve věku x dalších τ let. Platí:

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

Cenu všech životních produktů vypočteme jako střední současnou hodnotu budoucích výplat důchodů. K tomu ještě potřebujeme určit způsob diskontování. Vzhledem k tomu, že pravděpodobnosti úmrtí máme vypočtené pouze po rocích, nabízí se použít diskrétní diskontování. Nicméně spojitě diskontování bývá přesnější a lze také aplikovat na diskrétní data. Proto použijeme spojitě diskontování pomocí diskontního faktoru, který se spočítá jako exponenciální funkce s exponentem, jenž se rovná záporné hodnotě úrokové míry $\eta(\tau)$ násobené časem τ , $\exp(-\eta(\tau)\tau)$, stejně jako tomu udělali Shang a Haberman (2019). Dále uvažujme konstantní úrokovou míru $\eta(\tau) = \eta$ pro všechna $\tau = 1, \dots, n$.

První zmíněnou životní smlouvou je dočasný okamžitý důchod s výplatami na konci období. Jedná se o produkt, kde pojištěný zaplatí jednorázové pojistné **NSP** (Net Single Premium), za což bude dostávat pojistné plnění B (Benefit) za každý následující dožitý rok ihned od data platnosti smlouvy po předem stanovenou dobu n .

Uvažujme, že K značí počet dožitých let od začátku platnosti smlouvy, neboli že smrt nastala během $K + 1$. roku. Současné hodnoty výplat pojistných plnění pak jsou:

$$\mathbf{PV}(B) = \begin{cases} 0 & \text{pokud } K = 0, \\ B[e^{-1\eta} + e^{-2\eta} + \dots + e^{-K\eta}] & \text{pokud } K = 1, \dots, n, \\ B[e^{-1\eta} + e^{-2\eta} + \dots + e^{-n\eta}] & \text{pokud } K = n + 1, \dots \end{cases}$$

Pojistné **NSP** stanovíme jako

$$a_{x:\overline{n}|} = \mathbb{E}[\mathbf{PV}(B)] = B \sum_{\tau=1}^n e^{-\eta\tau} {}_{\tau}p_x, \quad (6.2)$$

kde B je předem domluvená výše pojistného plnění vyplacená každý rok ve stanoveném období, ${}_{\tau}p_x$ je pravděpodobnost přežití definovaná v 6.1 a $\exp(-\eta\tau)$ je diskontní faktor předpokládající konstantní úrokovou míru η .

Druhou zmíněnou životní smlouvou je doživotní okamžitý důchod s výplatami na konci období, což je produkt, který se od dočasného okamžitého důchodu liší pouze tím, že výplaty pojistného plnění neprobíhají po předem stanovenou omezenou dobu n , ale doživotně. Současné hodnoty výplat pojistných plnění mají podobu:

$$\mathbf{PV}(B) = \begin{cases} 0 & \text{pokud } K = 0, \\ B[e^{-1\eta} + e^{-2\eta} + \dots + e^{-K\eta}] & \text{pokud } K = 1, \dots \end{cases}$$

Jednorázové pojistné **NSP** tak stanovíme jako

$$a_x = \mathbb{E}[\mathbf{PV}(B)] = B \sum_{\tau=1}^{\infty} e^{-\eta\tau} {}_{\tau}p_x, \quad (6.3)$$

s tím, že využíváme stejné značení jako při výpočtu $a_{x:\overline{n}|}$.

Zbylé dva zmíněné produkty jsou alternativy k prvním dvěma s tím, že výplata neprobíhá okamžitě po uzavření smlouvy, nýbrž až po předem stanované době m .

Současné hodnoty výplat odloženého dočasného důchodu s výplatami na konci období jsou:

$$\mathbf{PV}(B) = \begin{cases} 0 & \text{pokud } K = 0, \dots, m, \\ Be^{-m\eta}[e^{-1\eta} + e^{-2\eta} + \dots + e^{-K\eta}] & \text{pokud } K = m + 1, \dots, m + n, \\ Be^{-m\eta}[e^{-1\eta} + e^{-2\eta} + \dots + e^{-n\eta}] & \text{pokud } K = m + n + 1, \dots \end{cases}$$

Pojistné **NSP** spočítáme jako:

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

kde opět využíváme stejné značení jako při výpočtu $a_{x:\overline{n}|}$.

Současné hodnoty výplat odloženého doživotního důchodu s výplatami na konci období jsou:

$$\mathbf{PV}(B) = \begin{cases} 0 & \text{pokud } K = 0, \dots, m, \\ Be^{-m\eta}[e^{-1\eta} + e^{-2\eta} + \dots + e^{-K\eta}] & \text{pokud } K = m + 1, \dots \end{cases}$$

Pojistné **NSP** spočítáme jako

$${}_m|a_x = \mathbb{E}[\mathbf{PV}(B)] = B \left[\sum_{\tau=1}^{\infty} e^{-\eta(m+\tau)} {}_{\tau}p_{x+m} \right] {}_m p_x = Be^{-\eta m} {}_m p_x \sum_{\tau=1}^{\infty} e^{-\eta\tau} {}_{\tau}p_{x+m}, \quad (6.5)$$

kde opět využíváme stejné značení jako při výpočtu $a_{x:\overline{n}|}$.

7. Aplikace na reálná data

V této kapitole aplikujeme teoretické poznatky, které jsme představili v předchozích kapitolách, na reálná data mužů a žen v České republice. Na webové stránce Český statistický úřad (2023) jsou dostupné úmrtnostní tabulky pro ČR od roku 1920 až po současnost (2021). Na tato data aplikujeme LC, RH a PCA (Analýza kompozičních dat) modely, provedeme predikce pro roky 2022-2051, porovnáme modely a odhadneme budoucí pravděpodobnosti přežití p_x a spočítáme ceny současných i budoucích životních produktů. Celkem budeme uvažovat 13 modelů - z toho třikrát LC model s různými metodami pro odhady parametrů (SVD, Poissonova metoda, alternativní metoda), dvakrát RH model (SVD, Poissonova metoda) a osmkrát PCA model s různými metodami pro predikci (ARIMA modely, ETS modely, náhodná procházka s driftem a náhodná procházka bez driftu) a s dvěma různými metodami pro určení počtu hlavních komponent L .

V celé kapitole budeme předpokládat konstatní intenzitu úmrtnosti a také, že střední stav populace P_x se rovná centrální expozici N_x .

7.1 Představení dat

Český statistický úřad archivuje úmrtnostní tabulky od roku 1920. V těchto tabulkách jsou dostupné úmrtnostní míry a tabulkové počty zemřelých pro věkové skupiny s jednoletým intervalem, a to pro skupiny 0 až 105+, kde poslední skupina zahrnuje všechny věky větší nebo rovno 105. Jak jsme se dozvěděli v podsektci 1.2, úmrtnostní tabulky jsou vypočteny na základě počtu úmrtí D_x a středního stavu populace P_x . ČSÚ měl k dispozici hodnoty D_x a P_x v letech 1920-1923 pro věky 0-94, v letech 1924-2006 pro věky 0-99 a v letech 2007-2021 pro věky 0-105. Úmrtnostní míry, které poskytuje Český statistický úřad (2023), jsou vyhlazené, jak jsme zmínili v sekci 1.2.

V LC a RH modelu vychází odhady parametrů z úmrtnostních měř $m_{x,t}$, v případě SVD metody se v druhé části odhadů parametrů k_t využívá rovněž počtů úmrtí $D_{x,t}$ a středních stavů populace $P_{x,t}$. PCA metoda je založena na znalosti tabulkových počtů úmrtí $d_{x,t}$.

Na obrázcích 7.1 a 7.2 máme představeny tabulkové počty zemřelých pro vybrané věky. Na základě těchto grafů jsme se rozhodli dále pokračovat pouze s daty z let 1970-2021.

Na obrázcích 7.3 a 7.4 máme graficky vyobrazeny úmrtnostní míry pro vybrané roky pro muže a ženy. Úmrtnostní míry nejprve klesají, což je způsobeno novorozeneckou úmrtností, která je vyšší než úmrtnost u malých dětí obecně. Po zbytek věků je patrný exponenciální růst úmrtnostních měř.

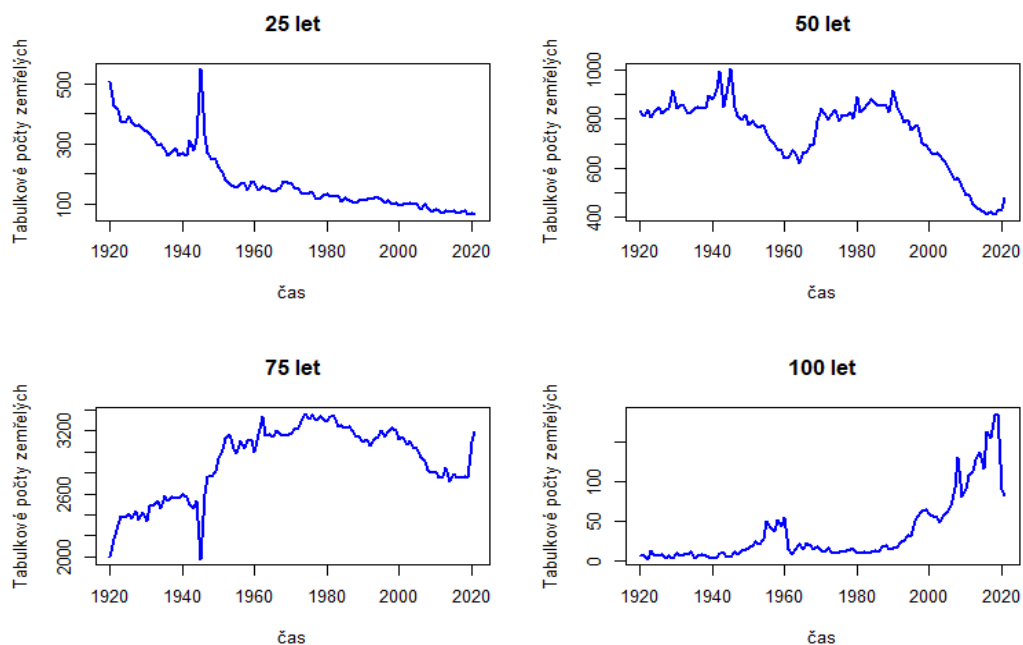
Pokud budeme porovnávat změnu úmrtnostních měř v čase, u žen na obrázku 7.4 je viditelný pokles hodnot úmrtnostních měř od věku 60. U mužů na obrázku 7.3 pozorujeme obdobný trend.

Obrázky 7.5 a 7.6 nám graficky představují tabulkové počty zemřelých $d_{x,t}$ pro vybrané roky pro muže a ženy. I tyto parametry nejprve klesají z důvodu vyšší úmrtnosti novorozenců. Přibližně od věku 10 počet zemřelých roste, výrazné zrychlení růstu pozorujeme okolo věku 50-60. Naopak okolo věku 70-90, v závislosti na pohlaví a roku, počet zemřelých klesá, neboť jedinců v těchto

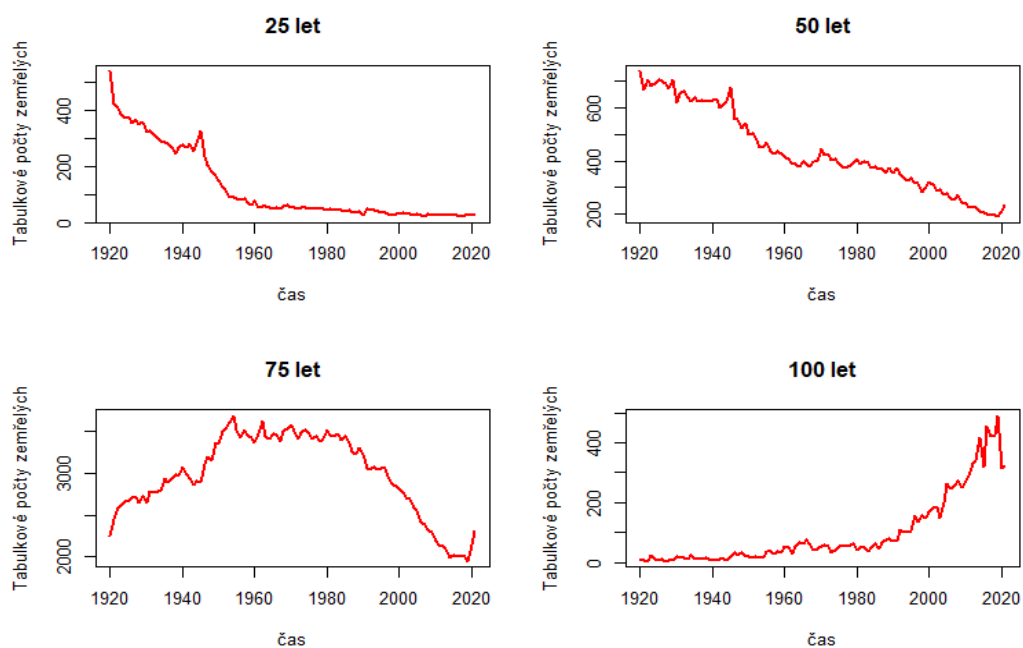
věcích v populaci je obecně méně.

Oba obrázky 7.5 a 7.6 rovněž zachycují trend prodlužování délky života - s rostoucím časem se vrchol křivky zvyšuje a posouvá se do pozdějších věků.

Pokud budeme porovnávat obrázek 7.5 se 7.6, můžeme si všimnout, že ženy obecně umírají ve vyšším věku.

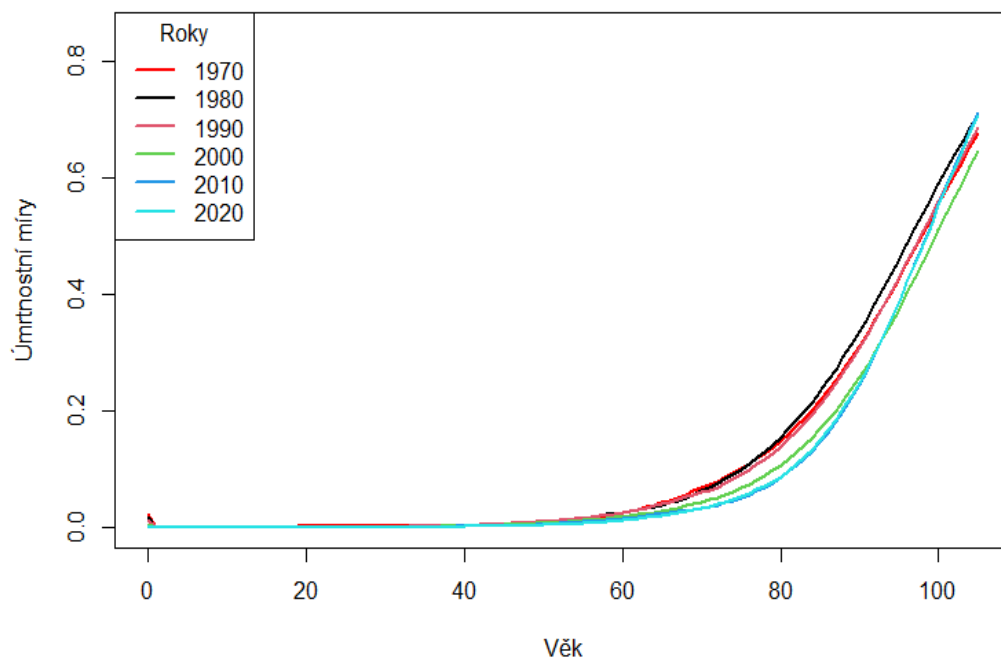


Obrázek 7.1: Porovnání tabulkových počtů zemřelých mužů ve věku 25, 50, 75 a 100 let.



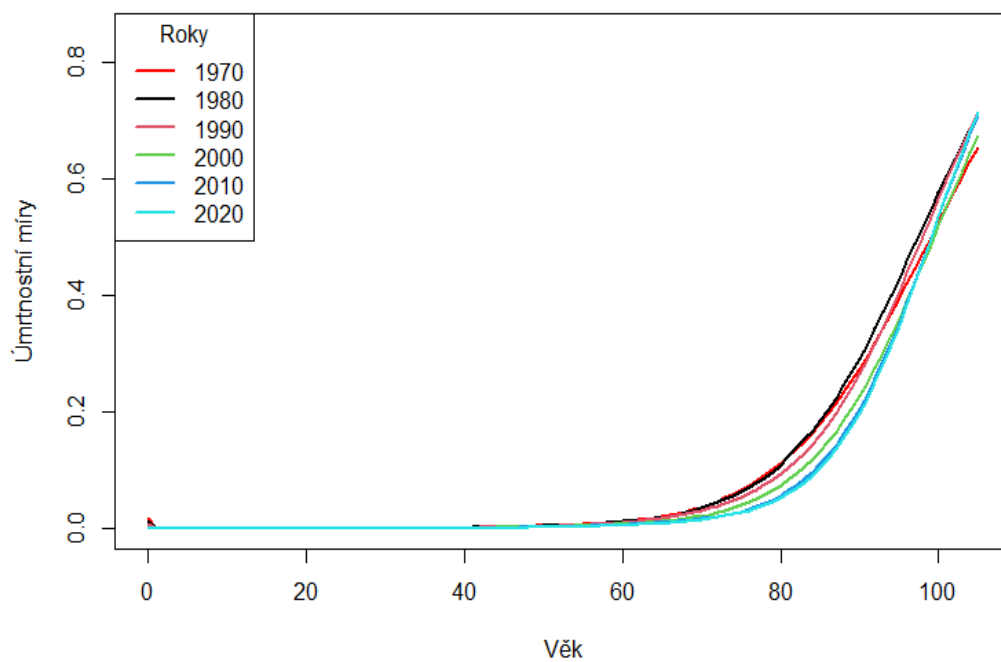
Obrázek 7.2: Porovnání tabulkových počtů zemřelých žen ve věku 25, 50, 75 a 100 let.

Muži

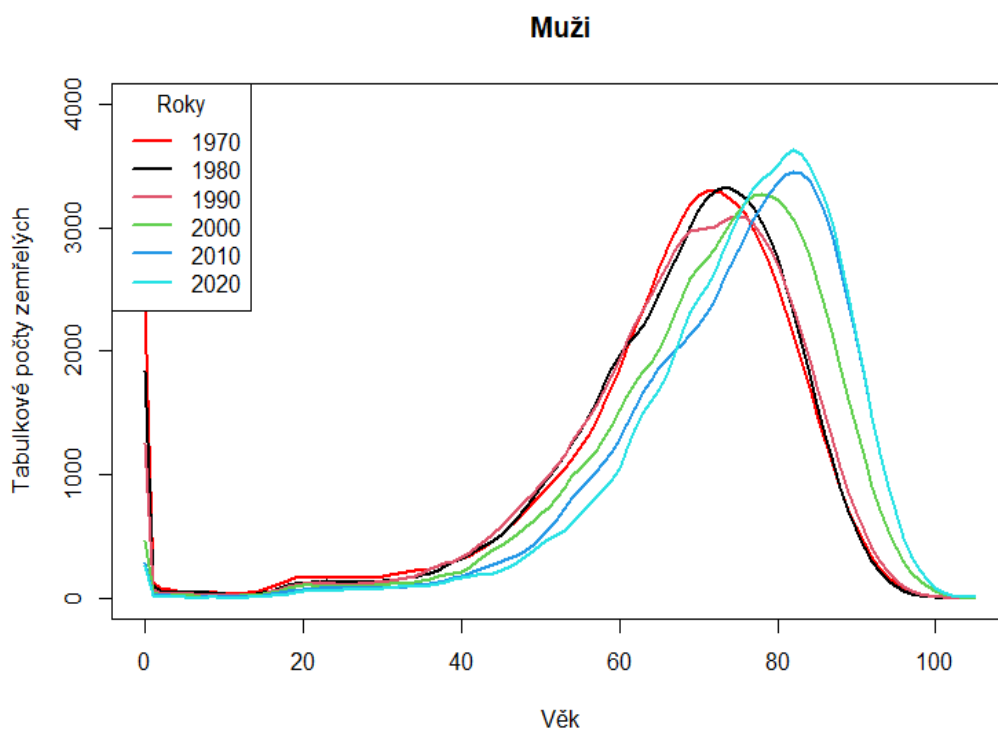


Obrázek 7.3: Porovnání úmrtnostních měr mužů z let 1970, 1980, 1990, 2000, 2010 a 2020.

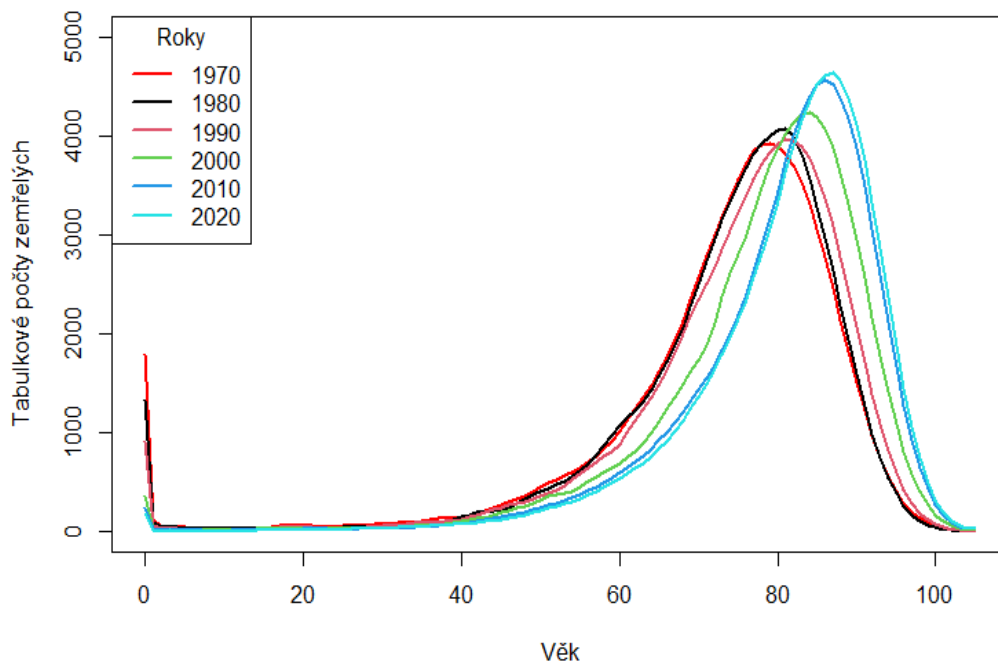
Ženy



Obrázek 7.4: Porovnání úmrtnostních měr žen z let 1970, 1980, 1990, 2000, 2010 a 2020.



Obrázek 7.5: Porovnání tabulkových počtů zemřelých mužů z let 1970, 1980, 1990, 2000, 2010 a 2020.



Obrázek 7.6: Porovnání tabulkových počtů zemřelých žen z let 1970, 1980, 1990, 2000, 2010 a 2020.

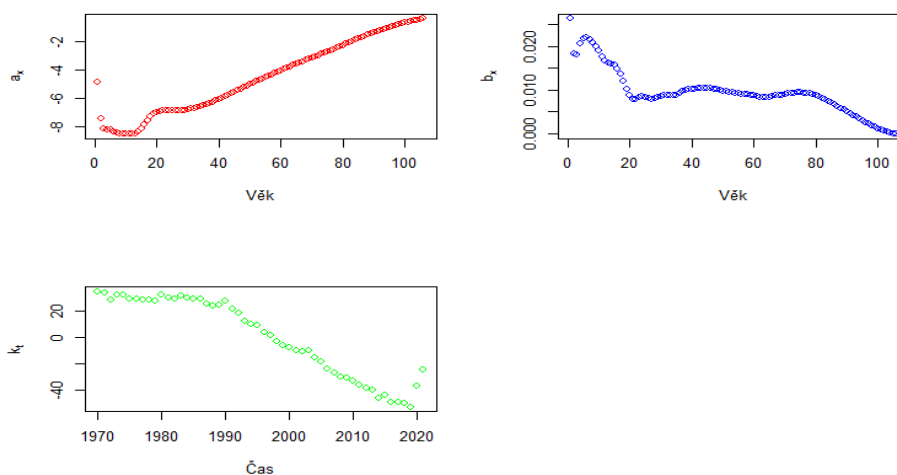
7.2 Odhady parametrů modelů

Pro odhady parametrů modelů využíváme data od roku 1970 do 2021 pro věky 0 až 105+. Platí tedy, že $n = 51$ a $m = 106$.

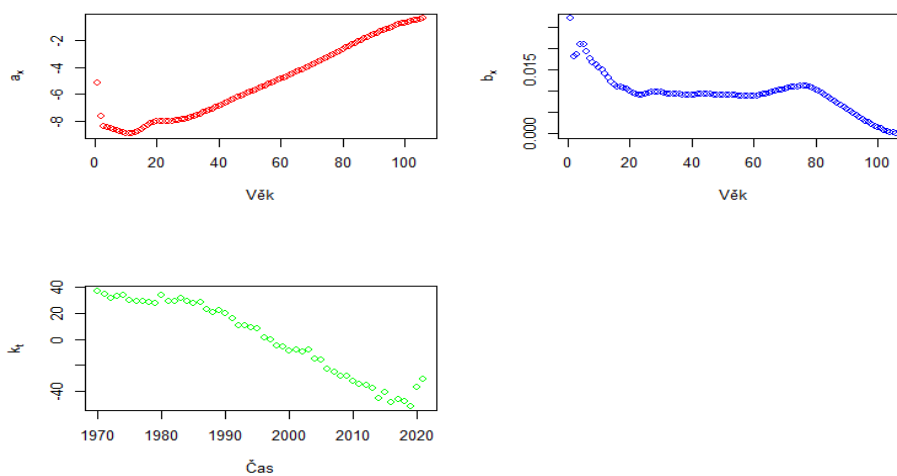
7.2.1 LC model

SVD metoda

Pro odhad parametrů jsme využili v R softwaru funkce `svd()`. Odhadnuté parametry jsme znormalizovali pomocí normalizačních podmínek 2.7, předodhadli jsme parametr k_t vyřešením rovnice 2.15 Newtonovým–Raphsonovým iteračním algoritmem a na závěr jsme odhady znovu znormalizovali. Výsledné odhady jsou k náhlednutí na obrázku 7.7 pro muže a na obrázku 7.8 pro ženy.

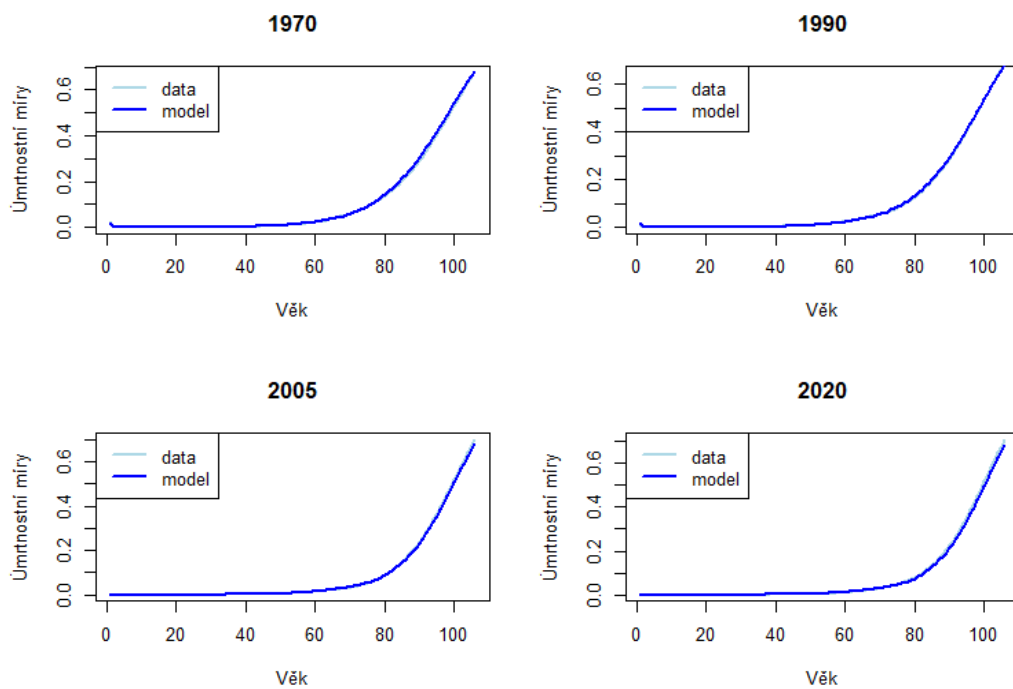


Obrázek 7.7: Finální odhady parametrů v LC-SVD modelu pro muže.

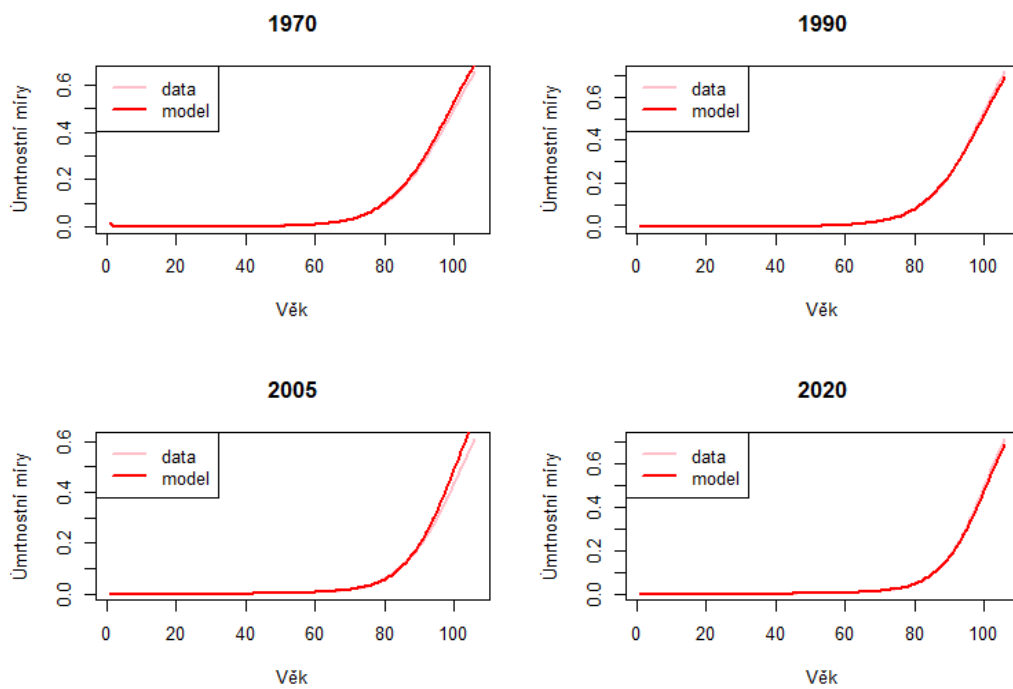


Obrázek 7.8: Finální odhady parametrů v LC-SVD modelu pro ženy.

Obrázky 7.9 a 7.10 ukazují, jak se liší odhadnuté úmrtnostní míry od úmrtnostních měř pozorovaných v letech 1970, 1990, 2005 a 2020 mužů a žen.



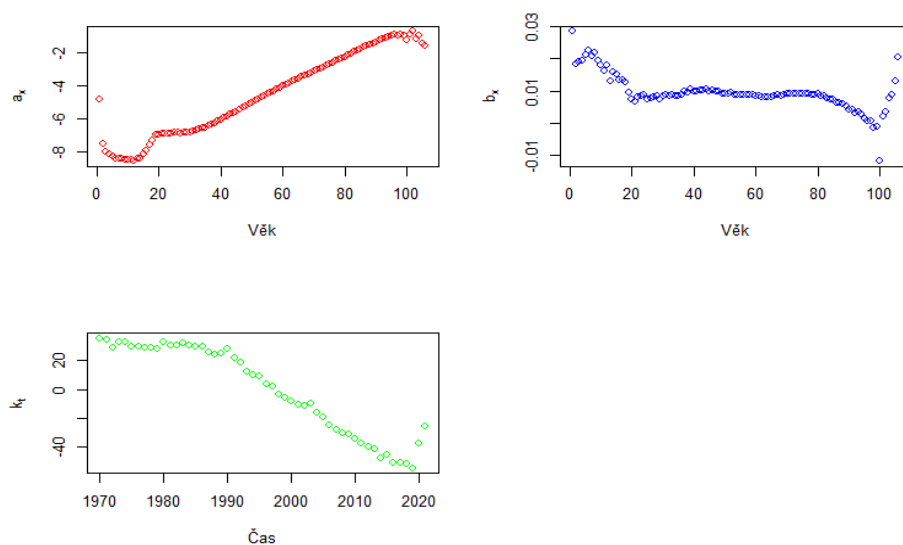
Obrázek 7.9: Porovnání pozorovaných a odhadnutých měř úmrtností pro muže LC-SVD modelem.



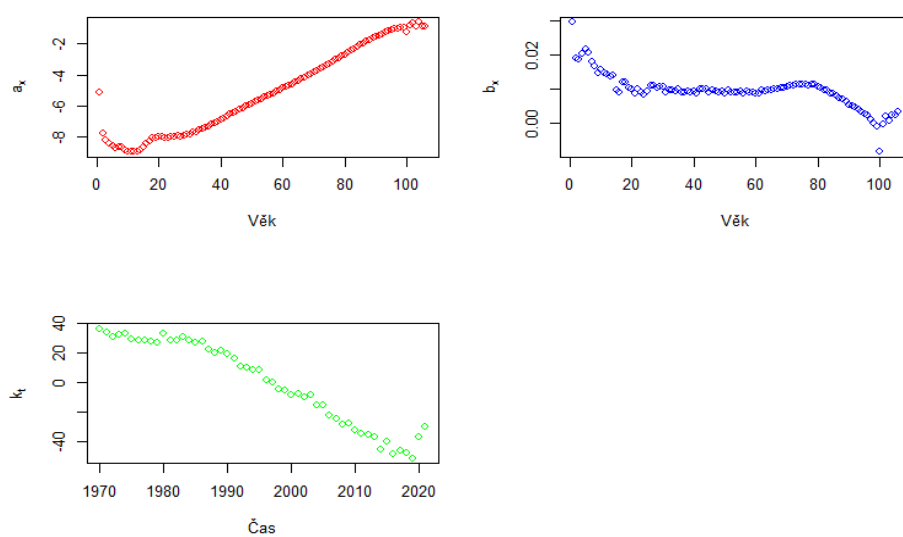
Obrázek 7.10: Porovnání pozorovaných a odhadnutých měř úmrtností pro ženy LC-SVD modelem.

Poissonova metoda

V Poissonově metodě jsme zvolili počáteční odhady parametrů na základě SVD rozkladu v R software opět pomocí funkce `svd()` a normalizačních podmínek 2.7. Dále jsme aktualizovali odhady parametrů pomocí iteračních vzorců 2.18, 2.19, 2.20, dokud úbytek hodnoty deviance $\text{Dev}(\mathbf{D}, \hat{\mathbf{D}})$ nebyl menší než 100. Následně jsme odhady parametrů opět znormalizovali a provedli předhadnutí parametru k_t zcela analogicky jako v sekci 7.2.1. Výsledné odhady jsou k nahlédnutí na obrázku 7.11 pro muže a na obrázku 7.12 pro ženy.



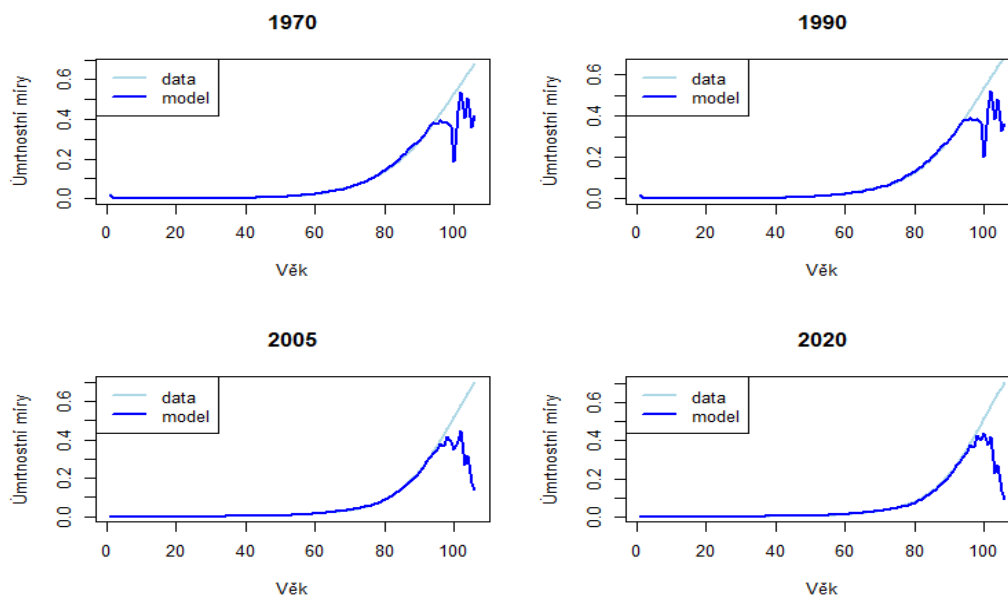
Obrázek 7.11: Finální odhady parametrů v LC-Poisson modelu pro muže.



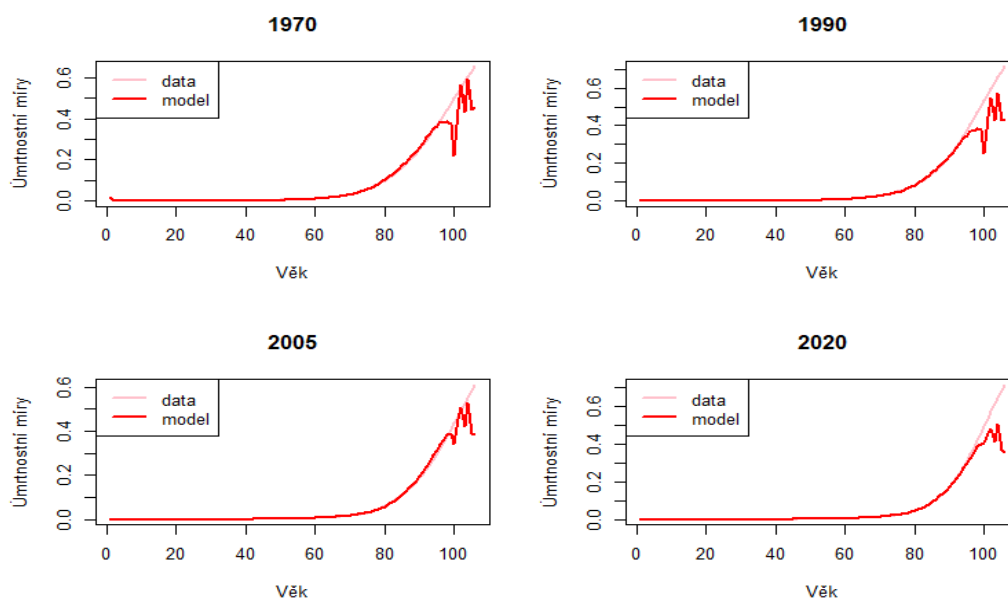
Obrázek 7.12: Finální odhady parametrů v LC-Poisson modelu pro ženy.

Obrázky 7.13 a 7.14 opět ukazují, jak se liší odhadnuté úmrtnostní míry od úmrtnostních měř pozorovaných v letech 1970, 1990, 2005 a 2020 mužů a žen.

Oproti modelu s SVD metodou si všímáme značných nepřesností pro nejvyšší věky. To je způsobené nedostupností dat pro věkové skupiny 100-105+ až do roku 2006. Poissonova metoda využívá k odhadování parametrů počty úmrtí D_x a centrální expozici N_x a právě nedostupnost těchto dat pro období 1970-2006 způsobuje zmíněné nepřesnosti.



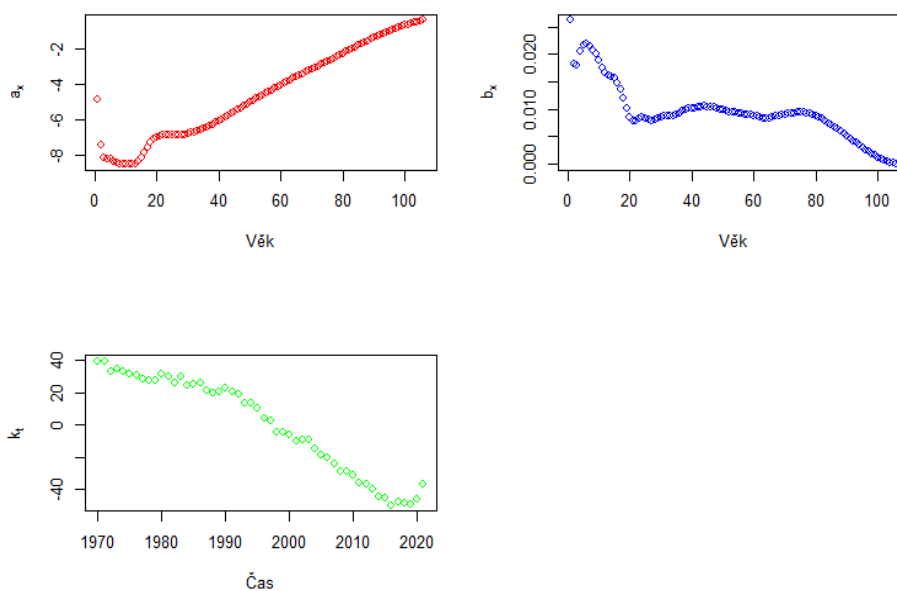
Obrázek 7.13: Porovnání pozorovaných a odhadnutých měř úmrtností pro muže LC-Poisson modelem.



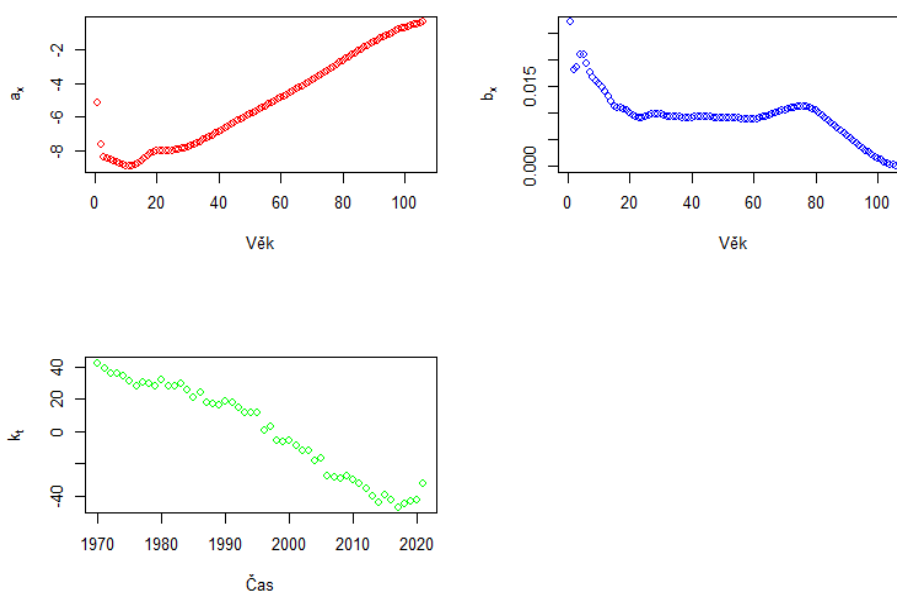
Obrázek 7.14: Porovnání pozorovaných a odhadnutých měř úmrtností pro ženy LC-Poisson modelem.

Alternativní metoda

Alternativní regresní metoda byla popsána v sekci 2.2.3. Výpočet v R-Studiu probíhal už pouze na základě vzorců 2.8, 2.21 a 2.24. Výsledné odhady jsou graficky zobrazeny na obrázku 7.15 pro muže a na obrázku 7.16 pro ženy.

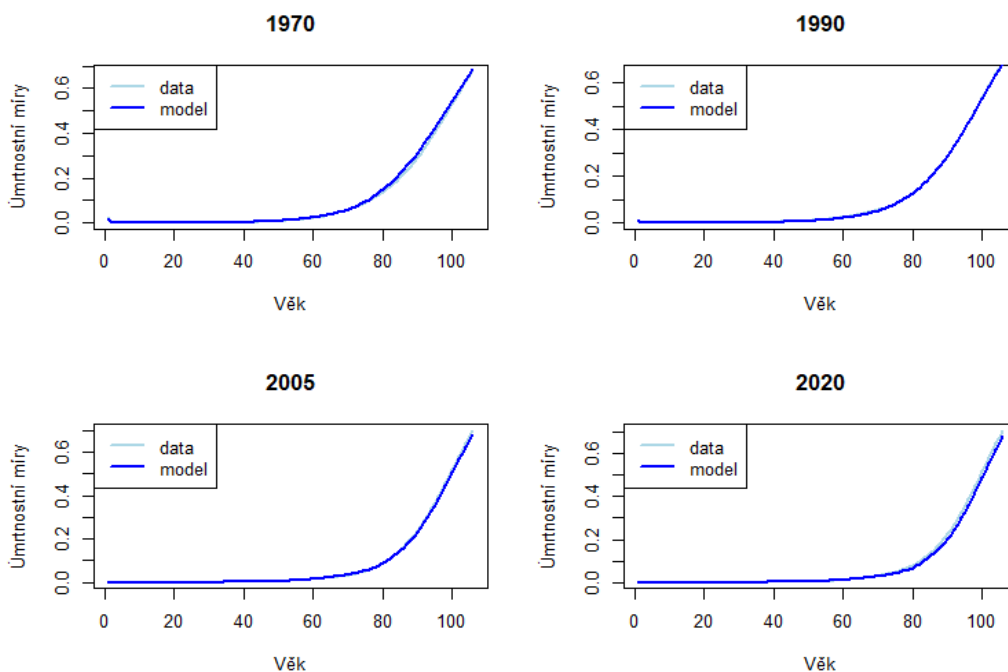


Obrázek 7.15: Finální odhady parametrů v LC-alternativním modelu pro muže.

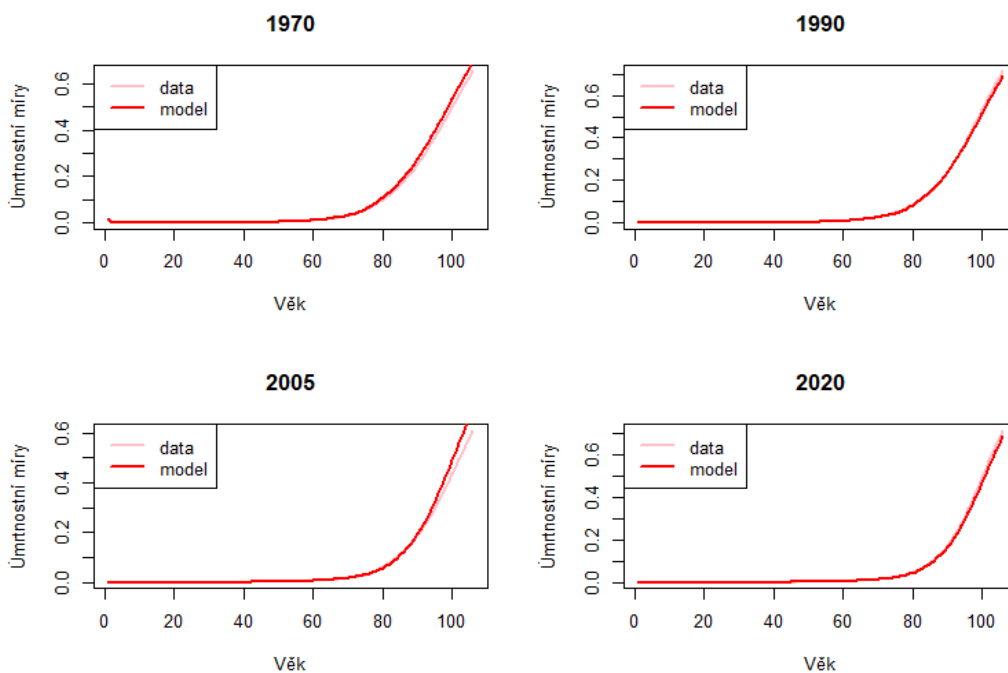


Obrázek 7.16: Finální odhady parametrů v LC-alternativním modelu pro ženy.

Obrázky 7.17 a 7.18 opět ukazují, jak se liší odhadnuté úmrtnostní míry od úmrtnostních měř pozorovaných v letech 1970, 1990, 2005 a 2020 mužů a žen.



Obrázek 7.17: Porovnání pozorovaných a odhadnutých měř úmrtností pro muže LC-alternativním modelem.

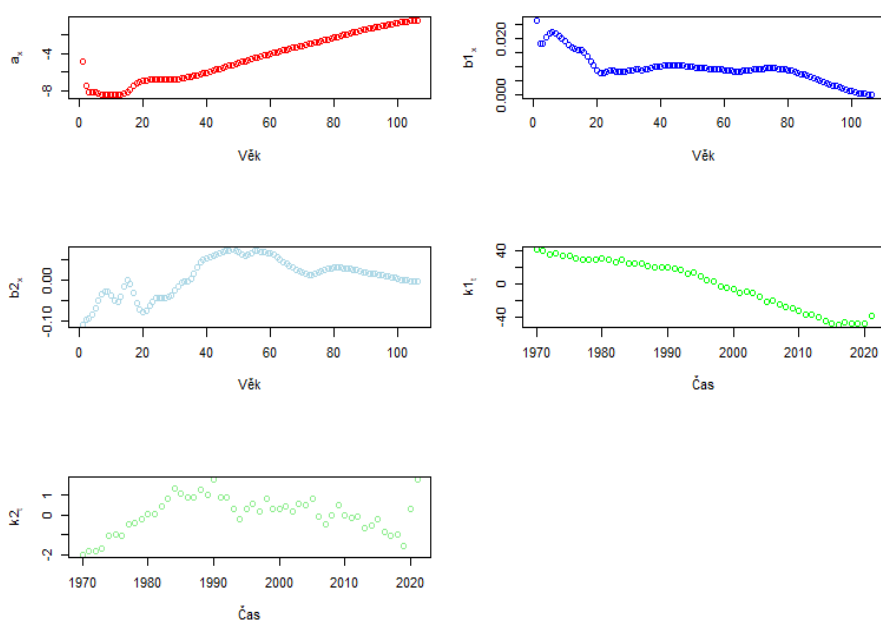


Obrázek 7.18: Porovnání pozorovaných a odhadnutých měř úmrtností pro ženy LC-alternativním modelem.

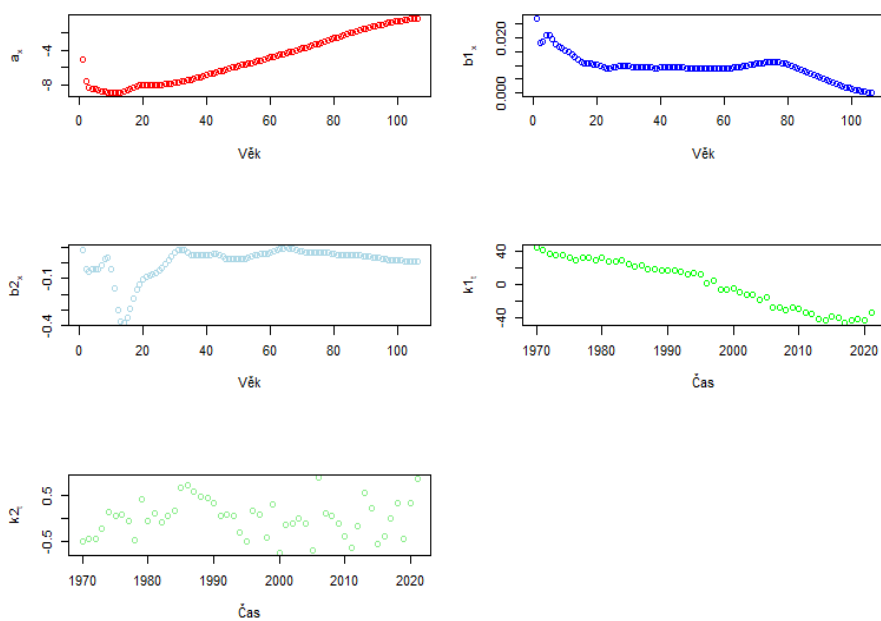
7.2.2 RH model

SVD metoda

Způsob odhadů parametrů v R-Studiu je zde analogický jako v sekci 7.2.1. Výsledné odhady parametrů \mathbf{a} , $\mathbf{b}^{(1)}$, $\mathbf{b}^{(2)}$, $\mathbf{k}^{(1)}$, $\mathbf{k}^{(2)}$ lze nahlédnout na obrázku 7.19 pro muže a na obrázku 7.20 pro ženy.

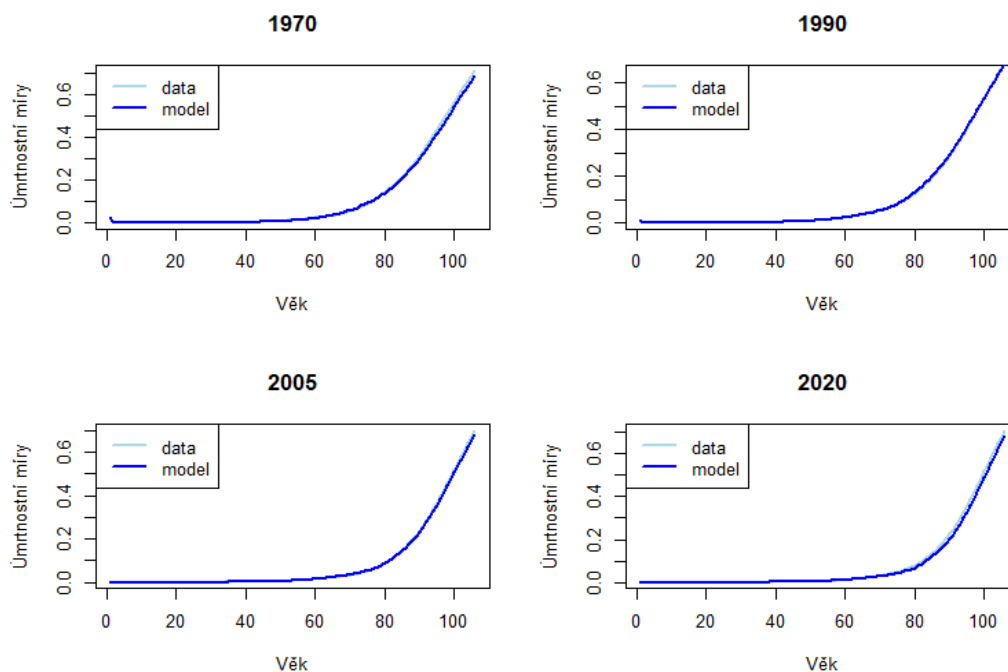


Obrázek 7.19: Finální odhady parametrů v RH-SVD modelu pro muže.

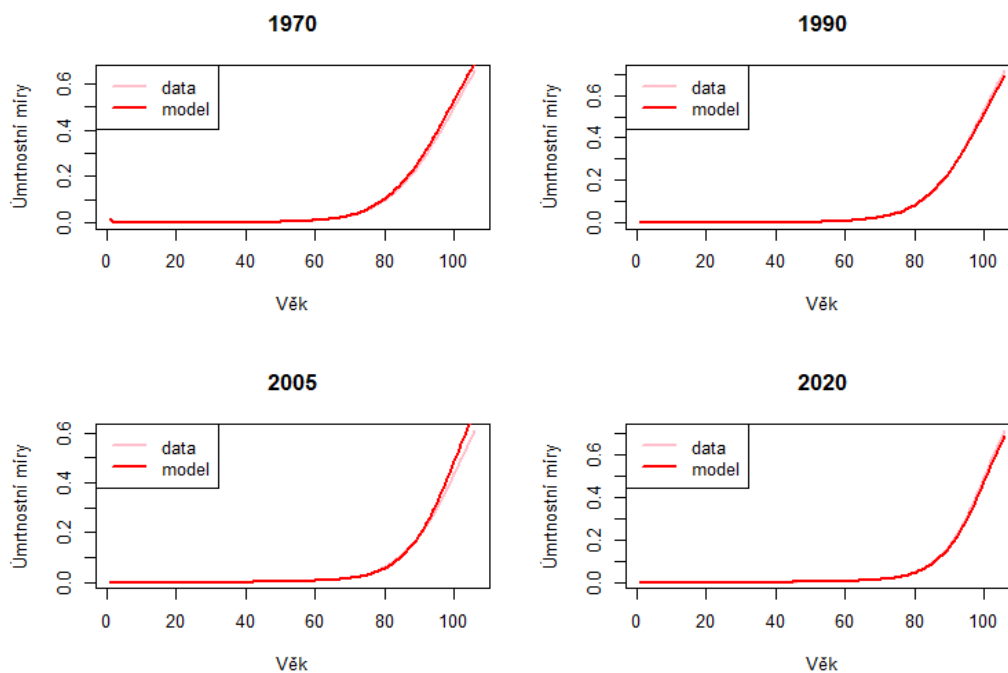


Obrázek 7.20: Finální odhady parametrů v RH-SVD modelu pro ženy.

Obrázky 7.21 a 7.22 opět ukazují, jak se liší odhadnuté úmrtnostní míry od úmrtnostních měř pozorovaných v letech 1970, 1990, 2005 a 2020 mužů a žen.



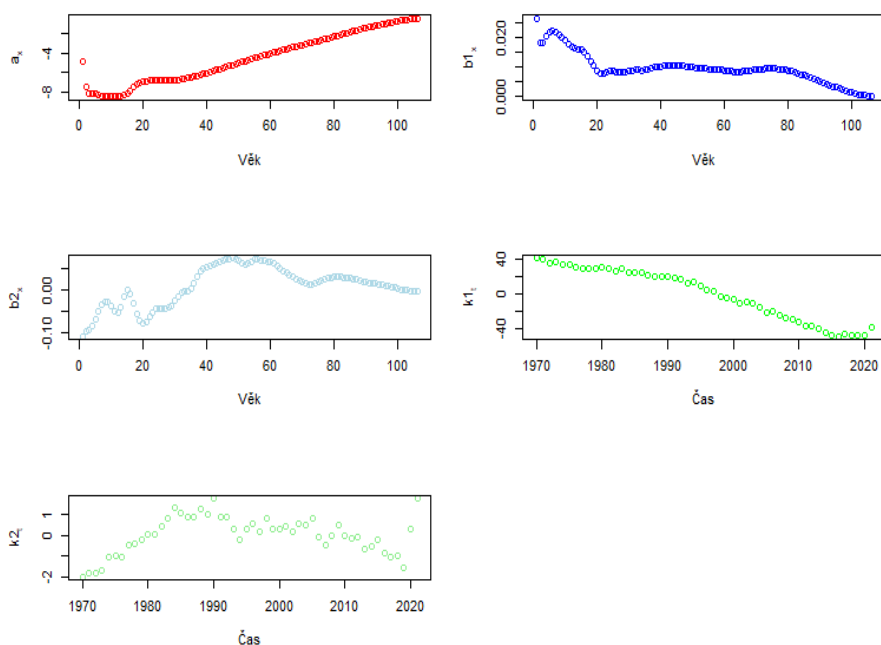
Obrázek 7.21: Porovnání pozorovaných a odhadnutých měř úmrtností pro muže RH-SVD modelem.



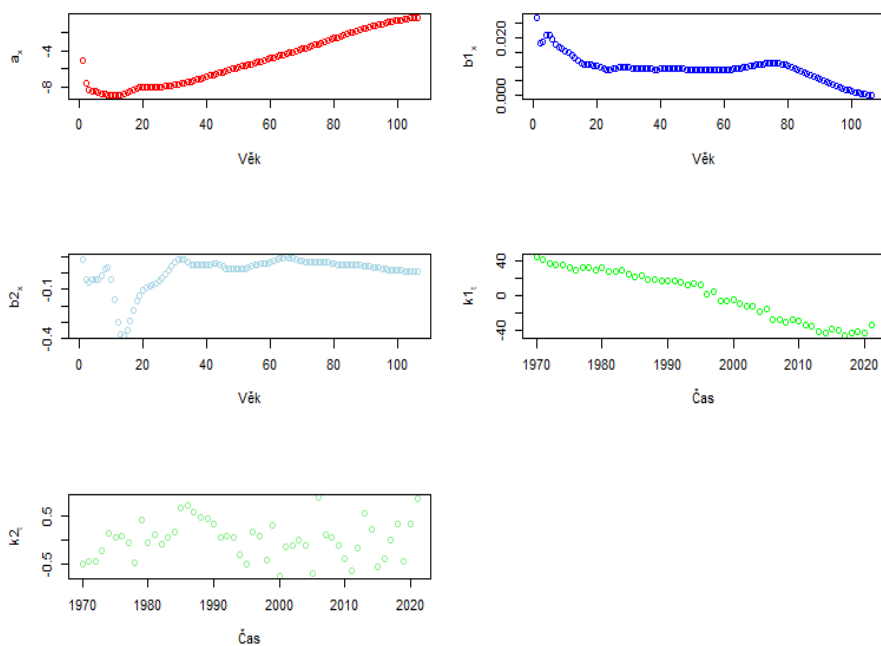
Obrázek 7.22: Porovnání pozorovaných a odhadnutých měř úmrtností pro ženy RH-SVD modelem.

Poissonova metoda

I zde je způsob odhadů parametrů v R-Studiu analogický odhadům Poissonovou metodou v LC modelu (7.2.1). Výsledné odhady parametrů jsou k dispozici na obrázcích 7.23 pro muže a na obrázku 7.24 pro ženy.



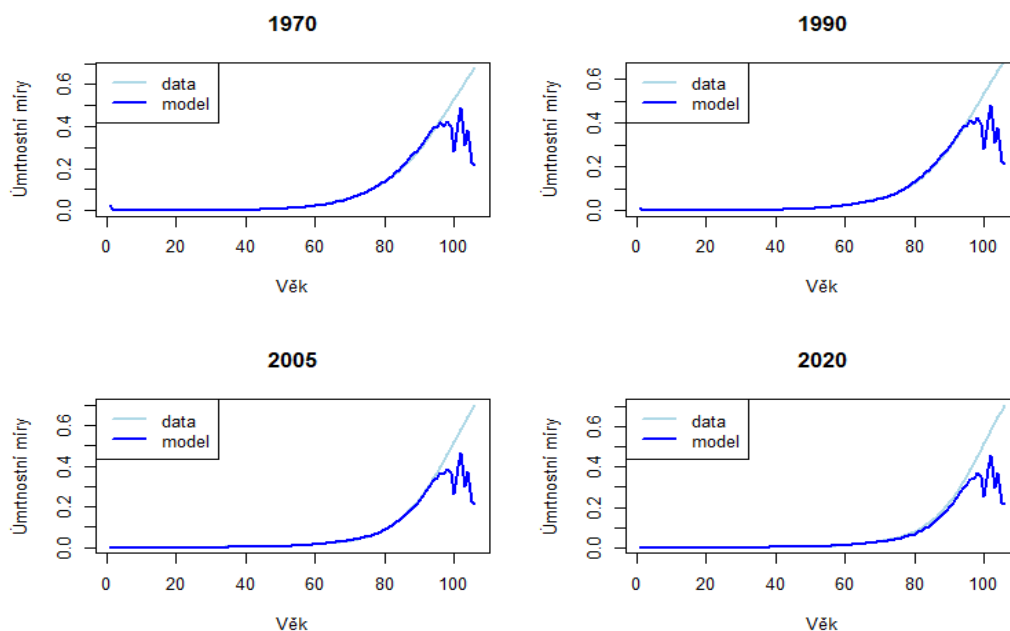
Obrázek 7.23: Finální odhady parametrů v RH-Poisson modelu pro muže.



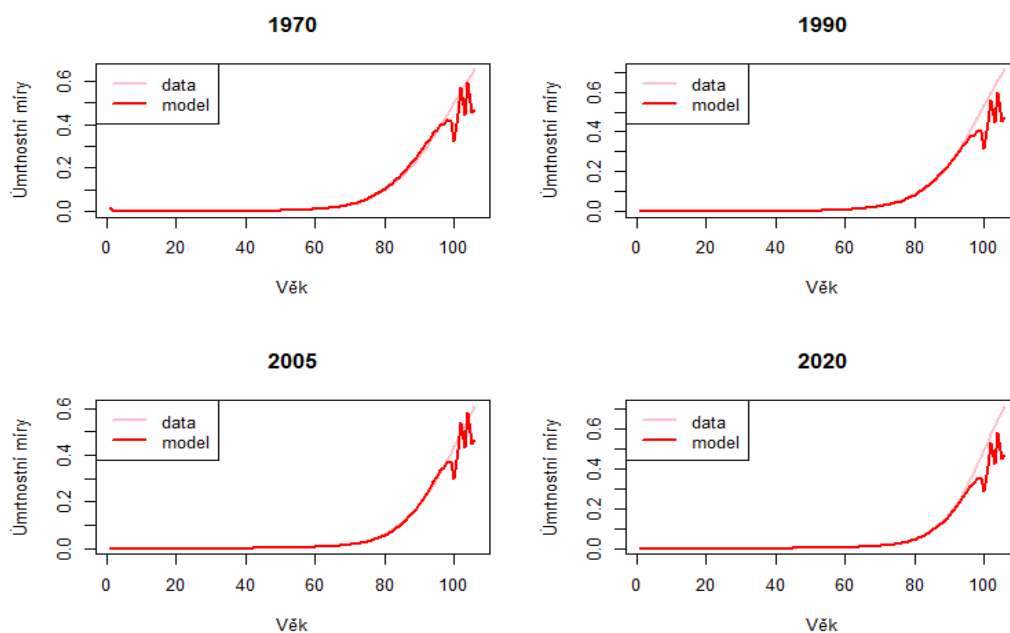
Obrázek 7.24: Finální odhady parametrů v RH-Poisson modelu pro ženy.

Obrázky 7.25 a 7.26 opět ukazují, jak se liší odhadnuté úmrtnostní míry od úmrtnostních měř pozorovaných v letech 1970, 1990, 2005 a 2020 mužů a žen.

Stejně jako v LC modelu i zde zaznamenáváme značné nepřesnosti v modelu pro nejvyšší věky. Důvod je opět nedostupnost dat D_x a N_x z let 1970-2006, které Poissonův algoritmus využívá pro odhad parametrů.



Obrázek 7.25: Porovnání pozorovaných a odhadnutých měř úmrtností pro muže RH-Poisson modelem.



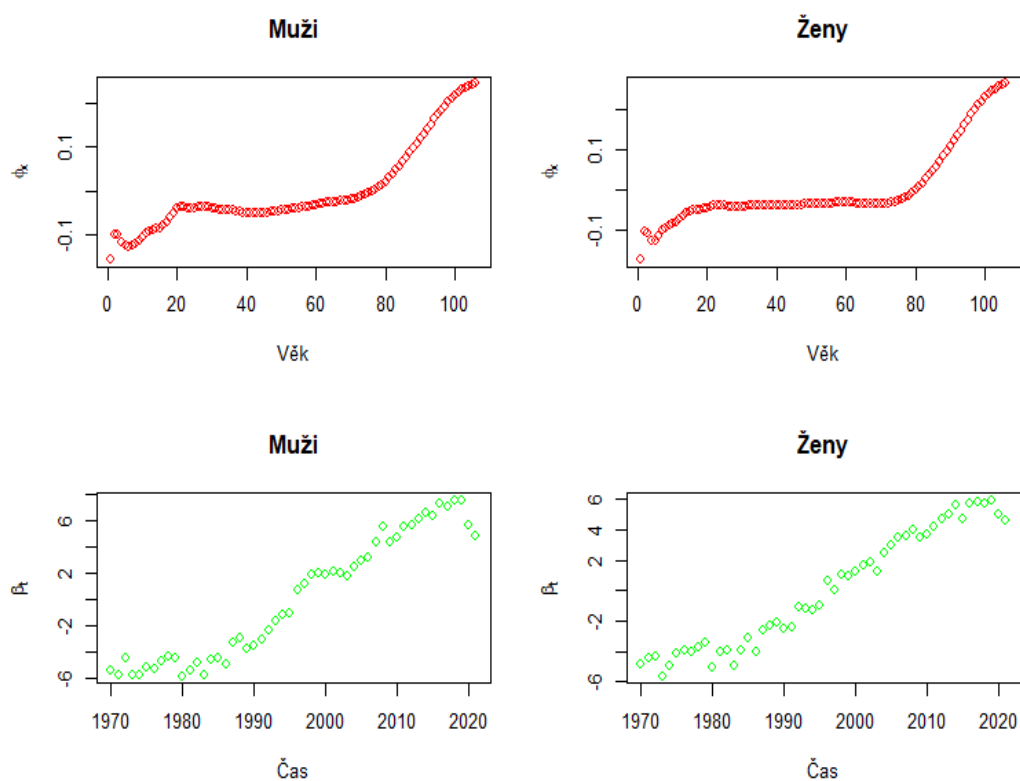
Obrázek 7.26: Porovnání pozorovaných a odhadnutých měř úmrtností pro ženy RH-Poisson modelem.

7.2.3 PCA model

V PCA modelu jsme nejprve přetransformovali data $d_{x,t}$ pomocí vzorců 4.2, 4.3, 4.4 a 4.5. Na základě transformovaných dat $z_{x,t}$ jsme podle rovnice 4.8 určili počet hlavních komponent, které budeme odhadovat. V obou případech, jak pro muže, tak pro ženy, nám vyšlo $L = 1$, neboť λ_1 zachytila 96.03 % variability v datech u mužů, resp. 94.48 % u žen. Následně jsme odhadli parametry $\hat{\beta}_{1,t}$ a $\hat{\phi}_{x,1}$ pomocí funkce `svd()` a vzorců 4.7.

Zajímavé by rovněž mohlo být uvažovat více hlavních komponent a faktorových skóre 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 mužů a 97.99 % u žen.

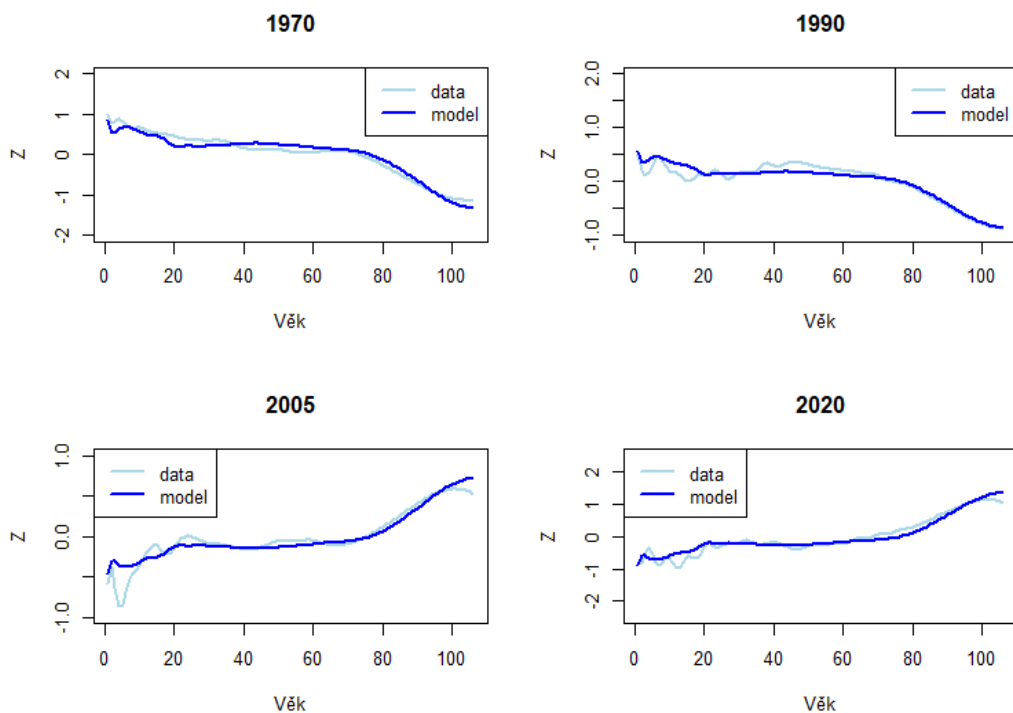
Obrázek 7.27 poskytuje grafický náhled na odhadnuté parametry $\hat{\beta}_{1,t}$ a $\hat{\phi}_{x,1}$ pro případ $L = 1$.



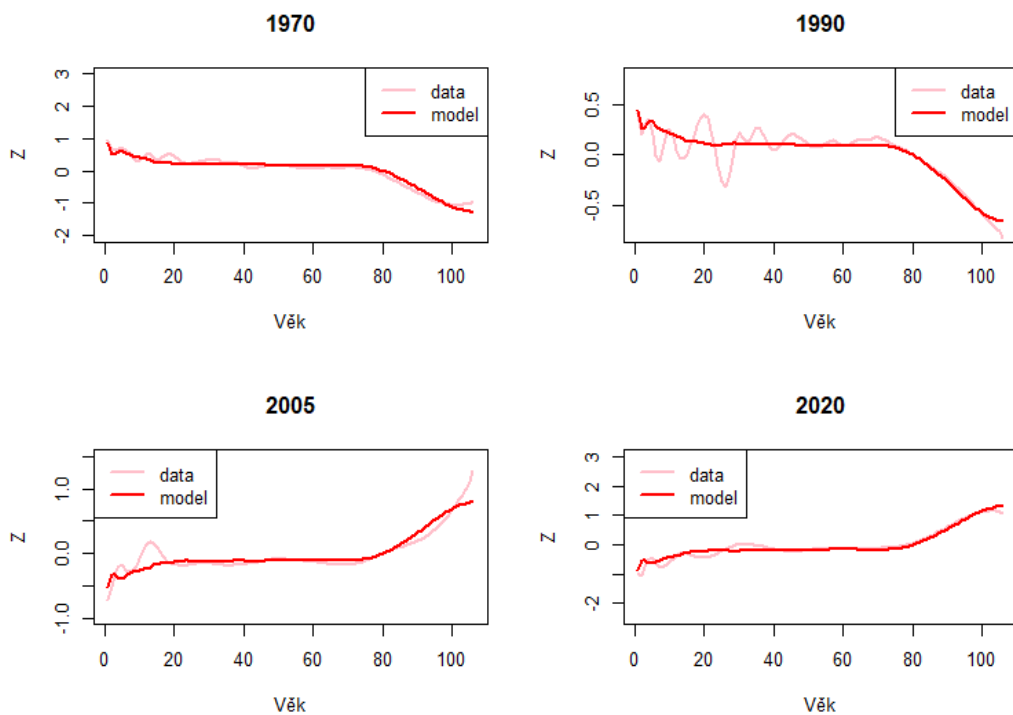
Obrázek 7.27: Finální odhady parametrů v PCA modelu s jednou hlavní komponentou pro muže a ženy.

Obrázky 7.28 a 7.29 poskytují porovnání odhadnutých parametrů $\hat{z}_{x,t}$ na základě PCA modelu s 1 hlavní komponentou od pozorovaných transformovaných dat $z_{x,t}$.

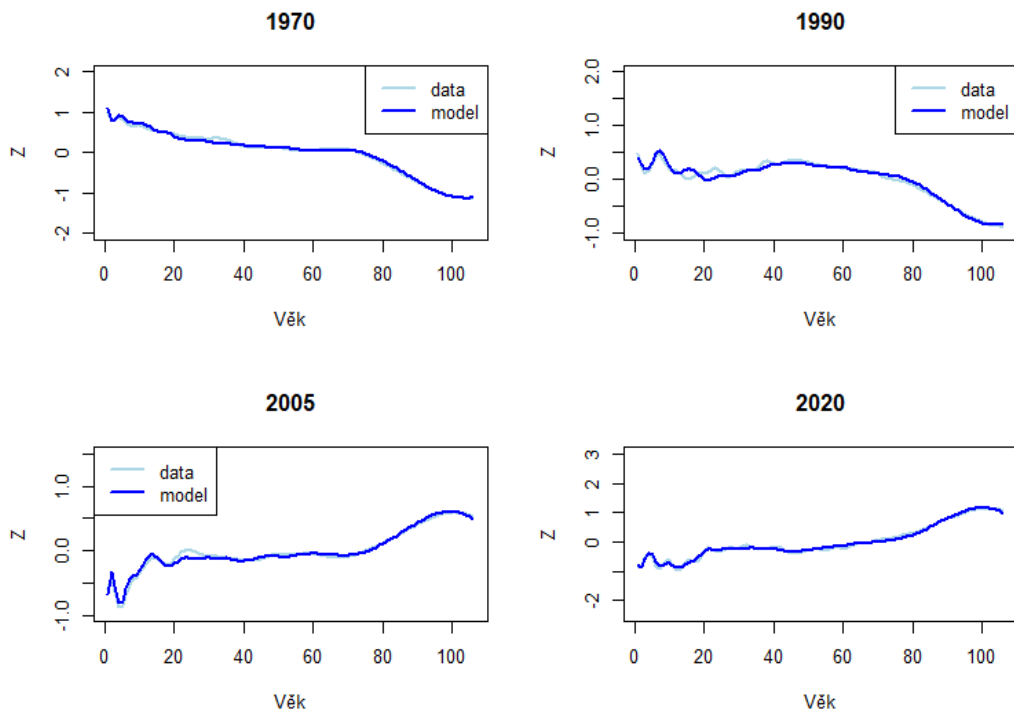
Obrázky 7.30 a 7.31 předkládají porovnání odhadnutých parametrů $\hat{z}_{x,t}$ na základě PCA modelu s 5 hlavními komponentami od pozorovaných transformovaných dat $z_{x,t}$. Pokud tyto grafy porovnáme s obrázky 7.28 a 7.29, všimneme si výrazného vylepšení.



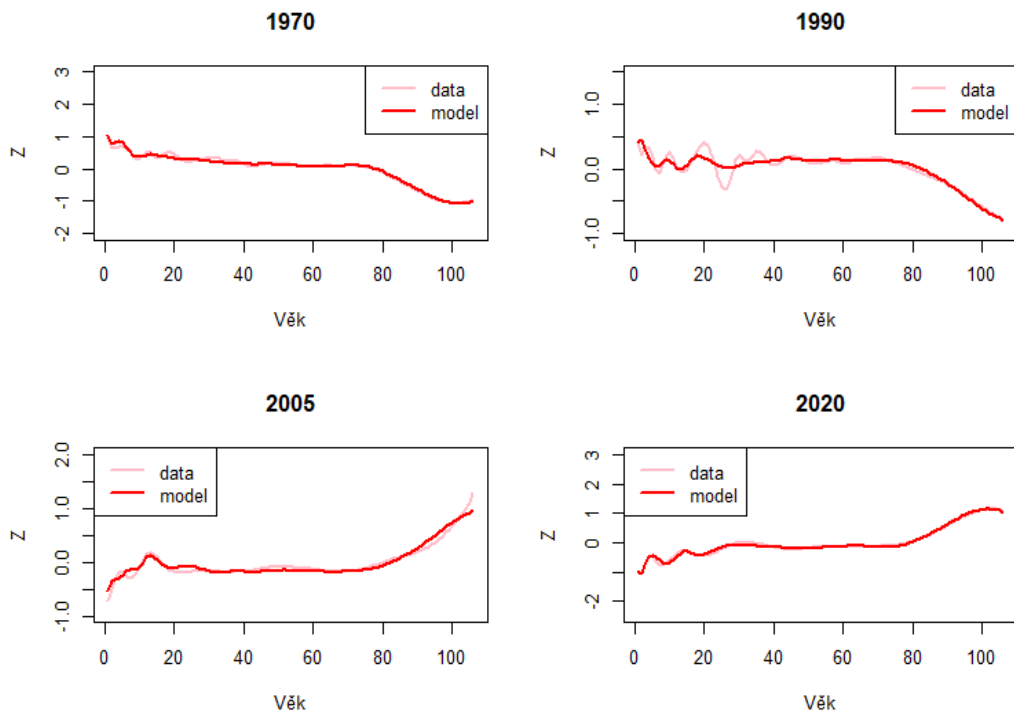
Obrázek 7.28: Porovnání pozorovaných transformovaných dat $z_{x,t}$ a odhadnutých parametrů $\hat{z}_{x,t}$ pro muže PCA modelem s 1 hlavní komponentou.



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

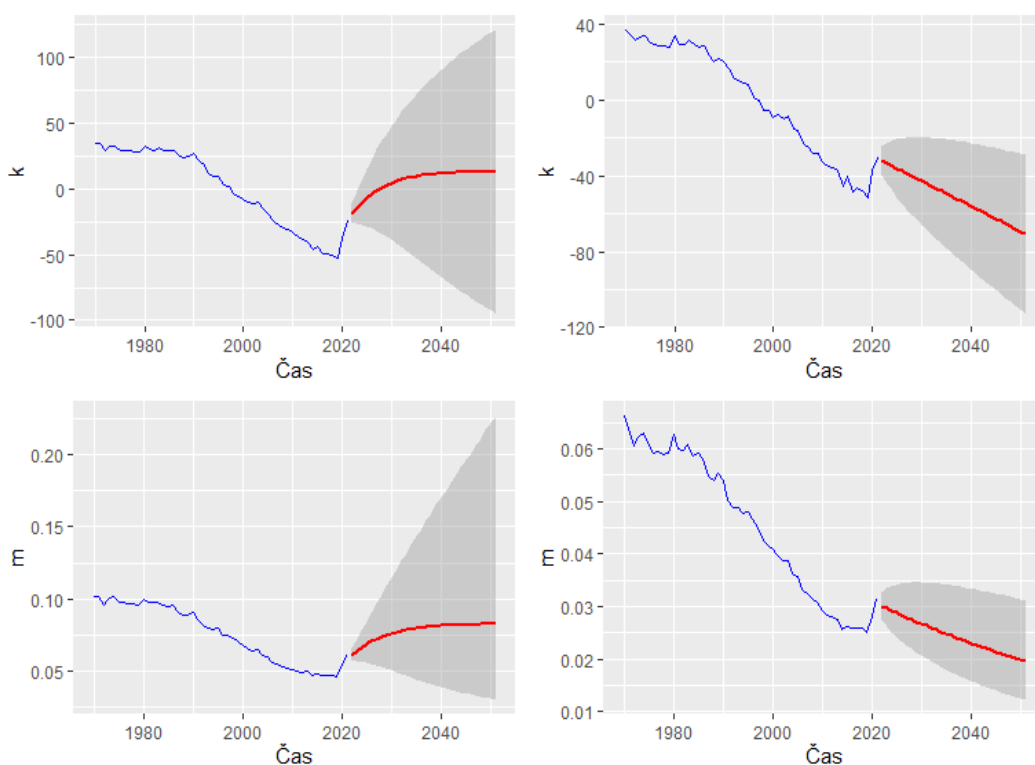
Na základě odhadnutých modelů budeme predikovat budoucí hodnoty tabulkových počtů úmrtí $d_{x,t}$, pro roky $t = 2022, \dots, 2051$. Horizont predikce je $H = 30$.

7.3.1 LC model

V LC modelu je na čase závislý pouze jeden parametr a tím je \mathbf{k} , na nějž budeme nahlížet jako na časovou řadu. Tu budeme analyzovat pomocí ARIMA modelů. V R Studiu jsme pro nalezení nejlepšího ARIMA modelu využili funkci `auto.arima()`. Predikce parametru \mathbf{k} byla získána funkcí `forecast()` a dopočtení budoucích hodnot $m_{x,t_{n+h}}$ a $d_{x,t_{n+h}}$ je pak přímočaré za předpokladu, že uvažujeme konstantní intenzitu úmrtnosti.

SVD metoda

Nejvhodnějším modelem pro LC model v kombinaci s SVD metodou pro odhad parametrů je ARIMA(1,1,1) bez driftu pro data mužů a ARIMA(0,1,0) s driftem pro data žen. Obrázek 7.32 poskytuje grafický pohled na současné a budoucí hodnoty parametru k včetně 95% intervalových odhadů, dále pak pro ilustraci také pozorované a odhadnuté budoucí hodnoty úmrtnostních měř včetně 95% intervalů spolehlivosti pro vybranou věkovou skupinu (75 let) mužů a žen.



Obrázek 7.32: Odhadnuté a budoucí hodnoty parametru \mathbf{k} včetně 95% intervalu spolehlivosti a pozorované a odhadnuté budoucí hodnoty úmrtnostních měř včetně 95% intervalu spolehlivosti - LC model SVD metoda. Vlevo grafy pro muže, vpravo pro ženy.

Poissonova metoda

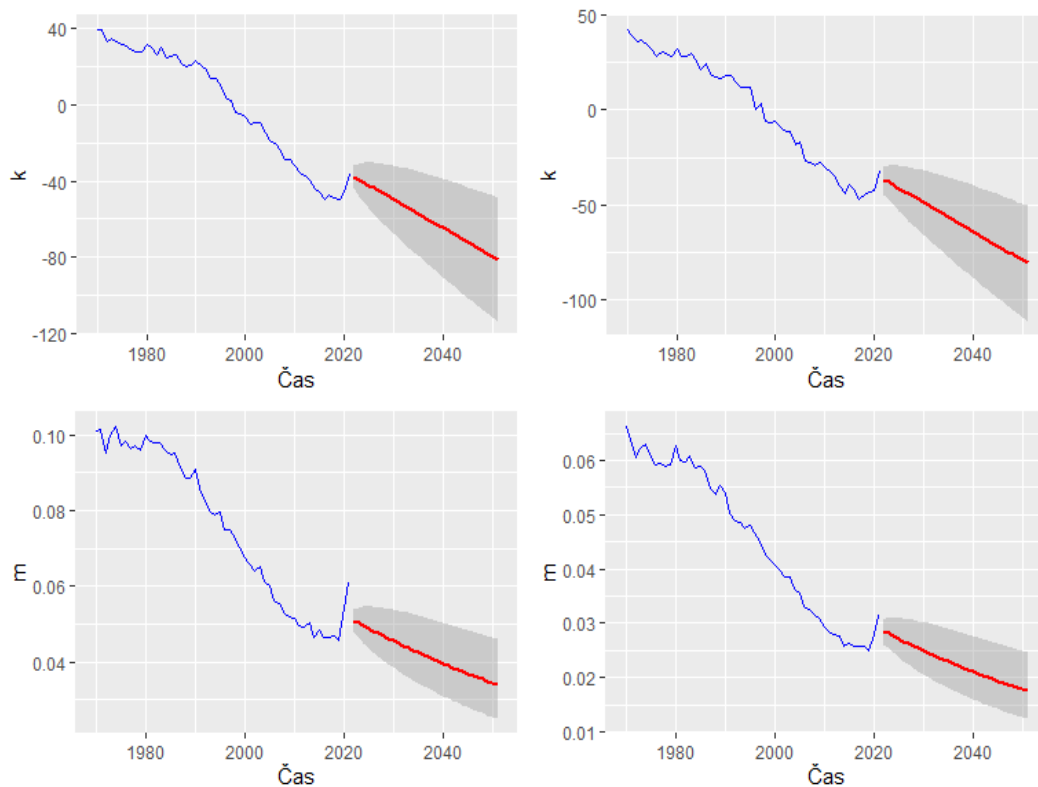
Stejně jako při použití SVD metody je i při použití Poissonovy metody nejlepším modelem pro mužská data model ARIMA(1,1,1) bez driftu a pro ženská data model ARIMA(0,1,0) s driftem. Na obrázku 7.33 jsou vykresleny současné a budoucí hodnoty parametru k včetně 95% intervalových odhadů, dále pak pro ilustraci také pozorované a odhadnuté budoucí hodnoty úmrtnostních měr včetně 95% intervalů spolehlivosti pro vybranou věkovou skupinu (75 let) mužů a žen.



Obrázek 7.33: Odhadnuté a budoucí hodnoty parametru k včetně 95% intervalu spolehlivosti a pozorované a odhadnuté budoucí hodnoty úmrtnostních měr včetně 95% intervalu spolehlivosti - LC model Poissonova metoda. Vlevo grafy pro muže, vpravo pro ženy.

Alternativní metoda

V LC modelu s využitím alternativní metody pro odhad parametrů vyšly nejlépe modely ARIMA(0,1,0) s driftem pro muže a ARIMA(1,1,0) s driftem pro ženy. Na obrázku 7.34 jsou vykresleny současné a budoucí hodnoty parametru k včetně 95% intervalových odhadů, dále pak pro ilustraci také pozorované a odhadnuté budoucí hodnoty úmrtnostních měr včetně 95% intervalů spolehlivosti pro vybranou věkovou skupinu (75 let) mužů a žen.



Obrázek 7.34: Odhadnuté a budoucí hodnoty parametru \mathbf{k} včetně 95% intervalu spolehlivosti a pozorované a odhadnuté budoucí hodnoty úmrtnostních měř včetně 95% intervalu spolehlivosti - LC model alternativní metoda. Vlevo grafy pro muže, vpravo pro ženy.

7.3.2 RH model

V RH modelu na čase závisí parametry $\mathbf{k}^{(1)}$ a $\mathbf{k}^{(2)}$. Tyto parametry jsme modelovali jako nezávislé časové řady pomocí ARIMA modelů. Stejně tak jako v LC modelu jsme v R Studiu pro výběr nejlepšího ARIMA modelu zvolili funkci `auto.arima()` a pro predikci parametrů $\mathbf{k}^{(1)}$ a $\mathbf{k}^{(2)}$ funkci `forecast()`. Dopočtení budoucích hodnot $m_{x,t_{n+h}}$ a $d_{x,t_{n+h}}$ je pak přímočaré za předpokladu, že uvažujeme konstantní intenzitu úmrtnosti.

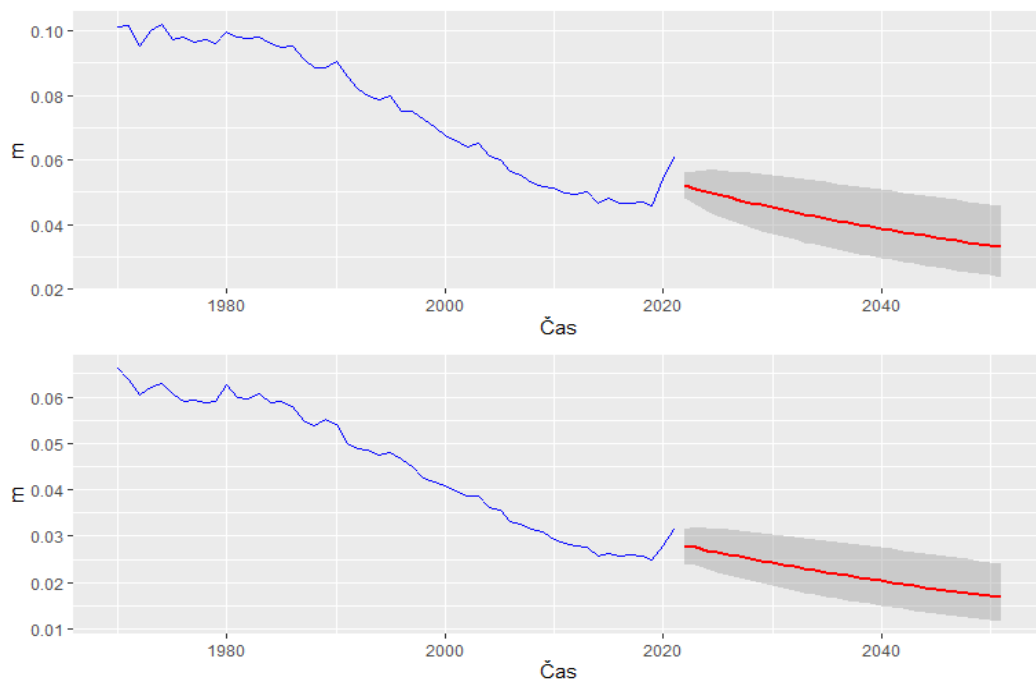
SVD metoda

V RH modelu s využitím SVD metody pro odhad parametrů jsme vybrali modely ARIMA(0,1,0) s driftem pro časovou řadu $\mathbf{k}^{(1)}$ pro muže, ARIMA(1,0,0) s nulovým průměrem pro $\mathbf{k}^{(2)}$ pro muže, ARIMA(1,1,0) s driftem pro časovou řadu $\mathbf{k}^{(1)}$ pro ženy a ARIMA(0,0,0) s nulovým průměrem pro $\mathbf{k}^{(2)}$ pro ženy. Současné a budoucí hodnoty parametrů $\mathbf{k}^{(1)}$ a $\mathbf{k}^{(2)}$ jsou k dispozici na obrázku 7.35.

Pro ilustraci opět přidáváme obrázek 7.36 s pozorovanými a odhadnutými budoucími hodnotami úmrtnostních měř včetně 95% intervalů spolehlivosti pro vybranou věkovou skupinu (75 let) mužů a žen.



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

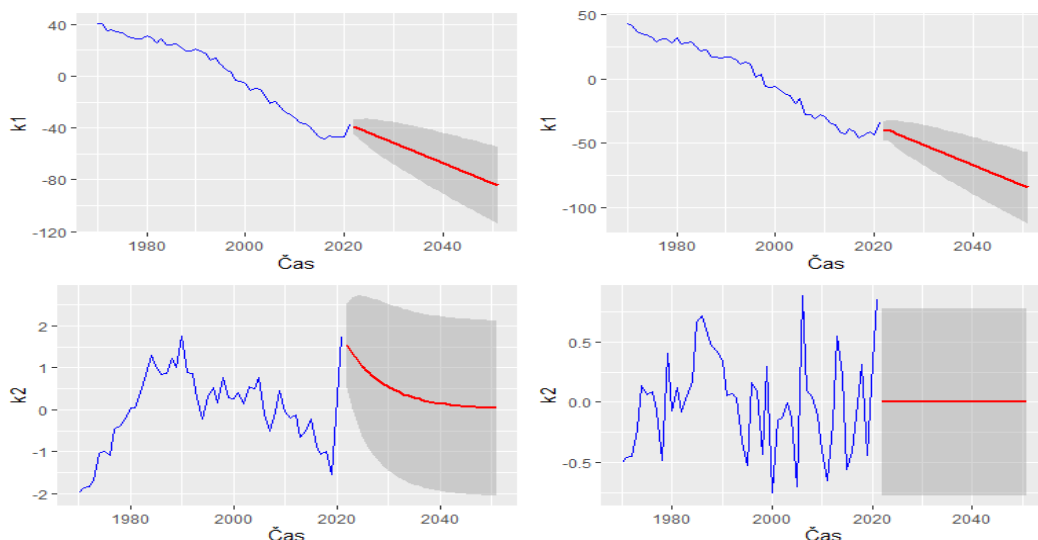


Obrázek 7.36: Pozorované a odhadnuté budoucí hodnoty úmrtnostních měř \mathbf{m} včetně 95% intervalu spolehlivosti - RH model, SVD metoda. První graf pro muže 75 let, druhý pro ženy 75 let.

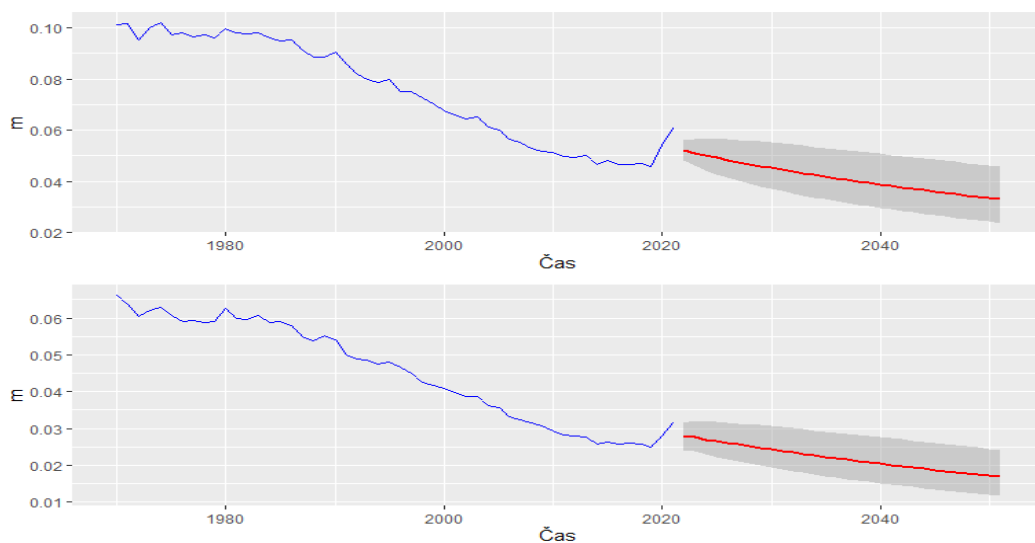
Poissonova metoda

V RH modelu s využitím Poissonovy metody pro odhad parametrů jsme vybrali modely ARIMA(0,1,0) s driftem pro $\mathbf{k}^{(1)}$ pro muže, ARIMA(1,0,0) s nulovým průměrem pro $\mathbf{k}^{(2)}$ pro muže, ARIMA(1,1,0) s driftem pro časovou řadu $\mathbf{k}^{(1)}$ pro ženy a ARIMA(0,0,0) s nulovým průměrem pro $\mathbf{k}^{(2)}$ pro ženy. Současné a budoucí hodnoty parametrů $\mathbf{k}^{(1)}$ a $\mathbf{k}^{(2)}$ jsou k dispozici na obrázku 7.37.

Pro ilustraci opět přidáváme obrázek 7.38 s pozorovanými a odhadnutými budoucími hodnotami úmrtnostních měř včetně 95% intervalů spolehlivosti pro vybranou věkovou skupinu (75 let) mužů a žen.



Obrázek 7.37: Odhadnuté a budoucí hodnoty parametru $\mathbf{k}^{(1)}$ a $\mathbf{k}^{(2)}$ včetně 95% intervalu spolehlivosti - RH model, Poissonova metoda. Vlevo grafy pro muže, vpravo pro ženy.



Obrázek 7.38: Pozorované a odhadnuté budoucí hodnoty úmrtnostních měř \mathbf{m} včetně 95% intervalu spolehlivosti - RH model, Poissonova metoda. První graf pro muže 75 let, druhý pro ženy 75 let.

7.3.3 PCA model

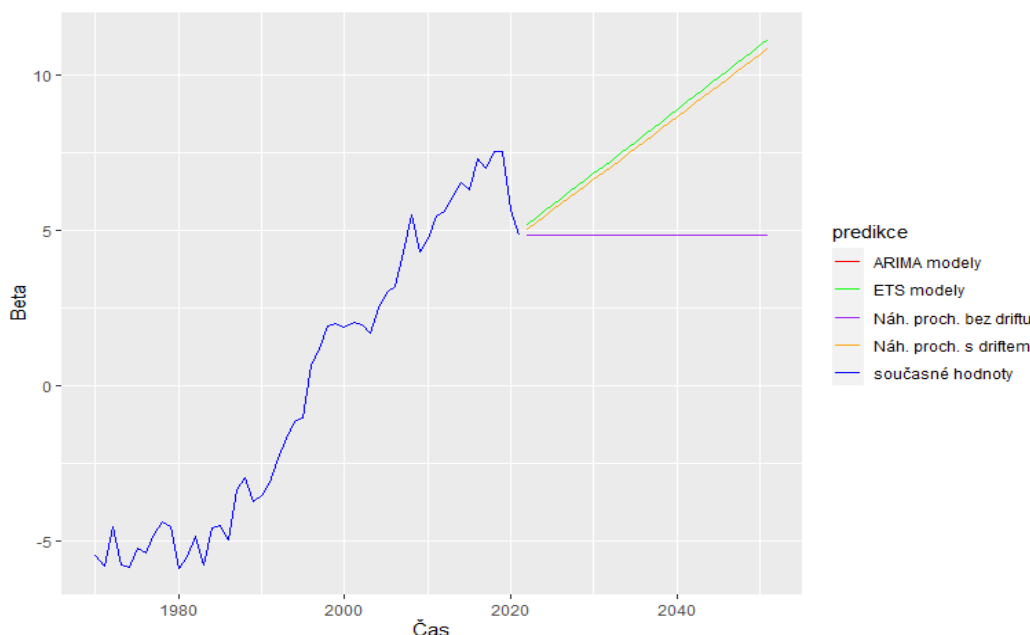
V PCA modelu závisí na čase pouze parametry $\beta_{l,t}$. Na tyto parametry opět budeme nahlížet jako na nezávislé časové řady. Máme čtyři různé způsoby, jak predikovat budoucí hodnoty. První možnost je opět nalézt vhodný ARIMA model, což provedeme v R Studiu opět pomocí funkce `auto.arima()`. Další možností je modelovat časovou řadu jako náhodnou procházku s driftem či bez driftu. V těchto případech budeme v R softwaru používat funkci `Arima()`. Poslední možnost je predikovat na základě ETS metody - v tom případě nalezneme vhodný ETS model pomocí funkce `ets()`¹.

Ve všech případech budoucí hodnoty spočítáme funkcí `forecast()`. Dopočtení budoucích hodnot $d_{x,t_{n+h}}$ je pak přímočaré.

1 hlavní komponenta

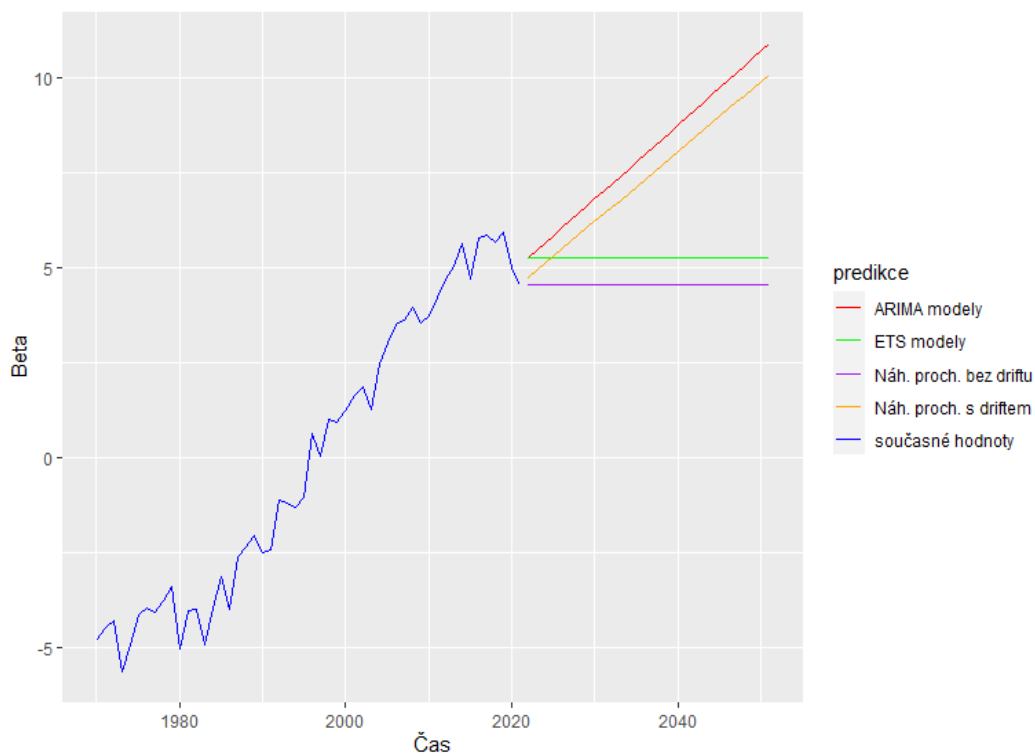
Nejprve se zaměříme na modely, kde jsme použili pro určení L metodu popsanou v 4.8. Zde máme pouze jeden parametr $\beta_{1,t}$ závislý na čase. Jako nejvhodnější ARIMA model pro muže se zdá být model ARIMA(0,1,0) s driftem a pro ženy ARIMA(0,1,1) s driftem. V případě ETS modelů jsme jako nejlepší model, jak pro muže, tak pro ženy, našli ETS(A,A,N), tedy model bez sezónnosti, který uvažuje aditivní trend a aditivní chybu.

Odhadnutý parametr $\beta_{1,t}$ včetně všech 4 způsobů predikce nalezneme na obrázku 7.39 pro mužská data a na obrázku 7.40 pro data žen. Pro ilustraci uvádíme i obrázky 7.41 a 7.42, kde si můžeme prohlédnout odhadnuté budoucí hodnoty $d_{x,t}$ pro vybrané věkové kategorie (25, 50, 75 a 100 let). Na obrázcích pro muže splývají predikce ARIMA modelů s náhodnou procházkou s driftem, neboť jsme v rámci ARIMA modelů vybrali právě náhodnou procházku s driftem.

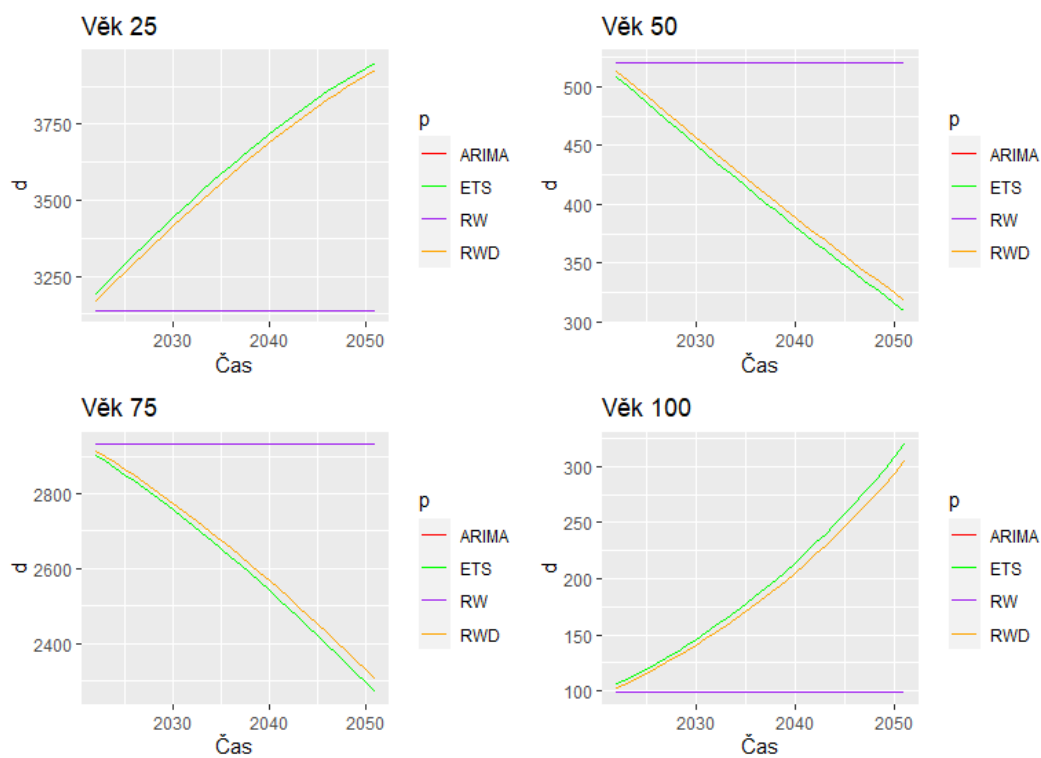


Obrázek 7.39: Budoucí hodnoty $\beta_{1,t_{n+h}}$ mužů na základě PCA modelu, v závislosti na predikční metodě.

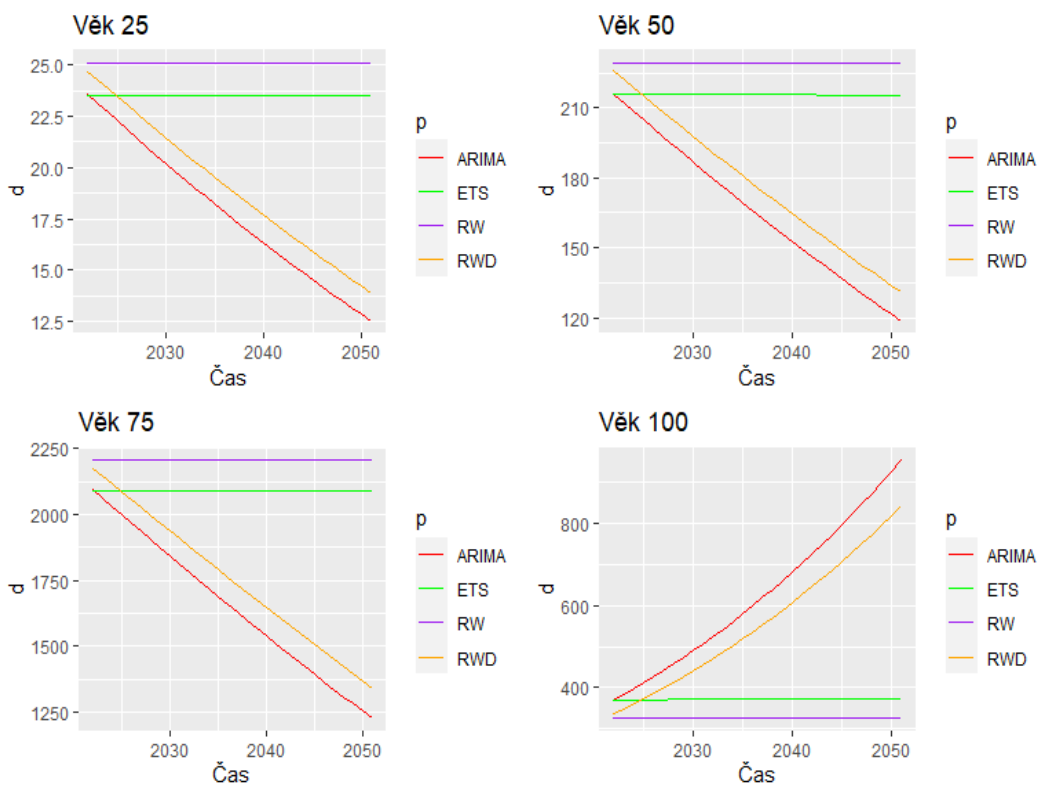
¹Funkce `ets()` v R Studiu v případě ročních dat neuvažuje modely se sezónností.



Obrázek 7.40: Budoucí hodnoty β_{1,t_n+h} žen na základě PCA modelu, v závislosti na predikční metodě



Obrázek 7.41: Budoucí hodnoty d_{x,t_n+h} mužů na základě PCA modelu v závislosti na predikční metodě pro vybrané věky 25, 50, 75 a 100 let.



Obrázek 7.42: Budoucí hodnoty $d_{x,t_{n+h}}$ žen na základě PCA modelu v závislosti na predikční metodě pro vybrané věky 25, 50, 75 a 100 let.

5 hlavních komponent

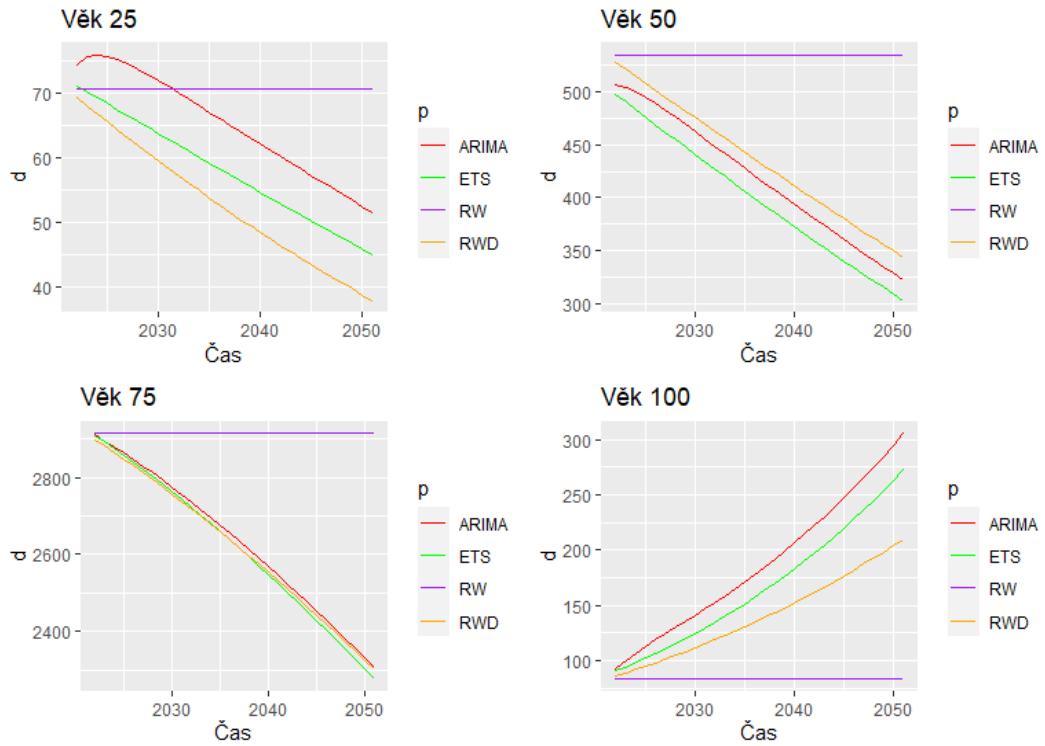
Pojďme se nyní podívat, jaké modely jsme vybrali pro jednotlivé časové řady $\beta_{l,t}$ v případě, kdy uvažujeme $L = 5$ pro muže i ženy. Nejvhodnější ARIMA a ETS modely pro zmíněné časové řady máme sepsány v tabulce² 7.1.

l	ARIMA muži	ARIMA ženy	ETS muži	ETS ženy
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)

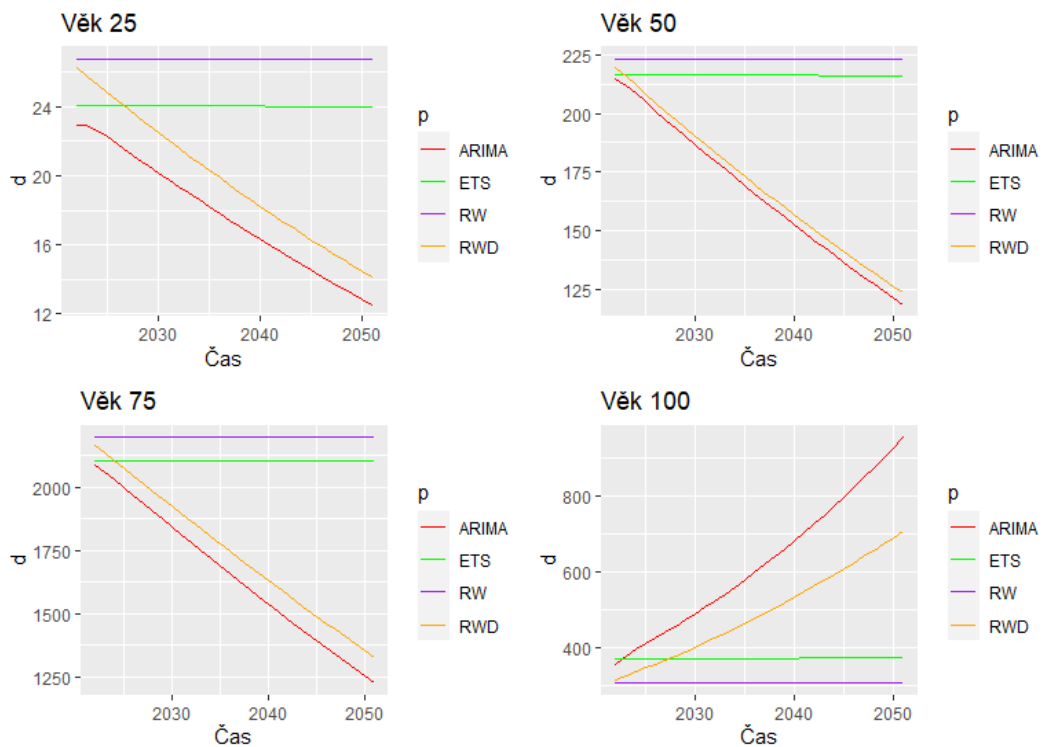
Tabulka 7.1: Nejlepší ARIMA a ETS modely pro PCA5 modely.

Obrázky 7.43 a 7.44 poskytují k nahlédnutí porovnání odhadnuté budoucí hodnoty $d_{x,t}$ na základě všech čtyř různých predikčních metod pro vybrané věkové skupiny (25, 50, 75 a 100 let).

²Symbol +d značí, že v modelu se uvažuje drift, v opačném případě se uvažuje model bez driftu.



Obrázek 7.43: Budoucí hodnoty $d_{x,t_{n+h}}$ mužů na základě PCA modelu s 5 hlavními komponentami v závislosti na predikční metodě pro vybrané věky 25, 50, 75 a 100 let.



Obrázek 7.44: Budoucí hodnoty $d_{x,t_{n+h}}$ žen na základě PCA modelu s 5 hlavními komponentami v závislosti na predikční metodě pro vybrané věky 25, 50, 75 a 100 let.

7.4 Porovnání modelů

Když máme modely odhadnuté a napredikovali jsme i budoucí hodnoty, přichází na řadu modely porovnat a určit, který model se pro naše data hodí nejvíce.

Naše práce, jak již název napovídá, se zabývá primárně počtem úmrtí, budeme tedy porovnávat, jak se liší od pozorovaných dat $d_{x,t}$ námi odhadnuté parametry $\hat{d}_{x,t}$.

7.4.1 Koeficient determinace

Nejprve se zaměříme na to, jak dobře modely kopírují pozorované hodnoty. K tomu použijeme koeficient determinace R^2 . V tabulce 7.2 máme hodnoty parametru R^2 pro všechny uvažované modely.

	LC-SVD	LC-PO	LC-ALT	RH-SVD	RH-PO	PCA1	PCA5
muži	0.9997	0.9997	0.9996	0.9997	0.9997	0.9996	0.9998
ženy	0.9999	0.9999	0.9998	0.9998	0.9998	0.9998	0.9999

Tabulka 7.2: Koeficient determinace pro jednotlivé modely mužů a žen.

Z tabulky 7.2 je patrné, že všechny modely popisují data velmi dobře, neboť jsou všechny hodnoty vyšší než 99.95 %. Nejvyšší hodnotu má PCA model s 5 hlavními komponentami pro mužská data a LC-SVD model³ pro data žen. Zajímavostí je, že LC model má vyšší hodnoty než RH model, tedy přidání druhé sady singulárních vektorů nevylepšílo kvalitu modelu v rámci R^2 .

7.4.2 MAPE a intervalové skóre

Dalšími kritérii pro určení nejvhodnějšího modelu jsou MAPE či intervalové skóre \bar{S}_γ . Zde odebíráme posledních 20 pozorování, rozměry dat tedy budou $n = 32$ a $m = 99$, neboť do roku 2006 máme dostupná data pouze do věku 99.

Z tabulky 7.3, kde máme zaznamenány hodnoty MAPE a $\bar{S}_{0.05}$ všech modelů, lze vyčíst, že nejlepším modelem pro muže je dle MAPE model PCA s 5 hlavními komponentami, kde jsme pro predikci volili náhodnou procházku s driftem. Pro ženy je dle MAPE nejlepší PCA model s jednou hlavní komponentou a náhodnou procházkou s driftem.

Pozorujeme, že obecně PCA modely poskytují nejpřesnější predikce, s výjimkou náhodné procházky bez driftu. Přidáním dalších komponent do modelu se predikce žen nezlepší. U mužů se budoucí hodnoty zpřesní kromě PCA-ARIMA modelu, největší rozdíl je u náhodné procházky s driftem.

LC model obecně zaostává za zbylými modely, nejhůře na tom je LC model s Poissonovou metodou. Nejlepší odhady z LC modelu má ten s alternativní metodou pro muže a s SVD metodou pro ženy. U RH modelu pozorujeme zpřesnění budoucích odhadů oproti LC modelu kromě SVD metody u žen.

Pokud budeme hodnotit pomocí intervalového skóre pro $\gamma = 0.05$ s tím, že intervalové odhady jsme sestrojili na základě bootstrap metod, získáme jako nejlepší variantu PCA model s 5 hlavními komponentami s náhodnou procházkou

³U žen jsou hodnoty pro LC-SVD = 0.999888, LC-Po = 0.999873, pro PCA5 = 0.999856.

Model	MAPE muži	MAPE ženy	$\bar{S}_{0.05}$ muži	$\bar{S}_{0.05}$ ženy
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

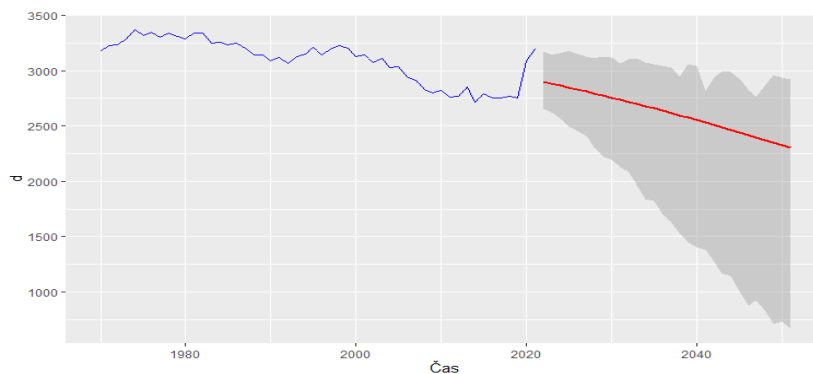
Tabulka 7.3: Porovnání modelu podle MAPE a intervalového skóre.

bez driftu pro muže a PCA model s 1 hlavní komponentou s náhodnou procházkou bez driftu pro ženy. Nejhuře dopadl LC model s SVD metodou. RH model má přesnější intervalové odhady v porovnání s LC modelem. PCA modely dopadly celkově nejlépe s tím, že nepozorujeme výrazné zlepšení přidáním dalších hlavních komponent.

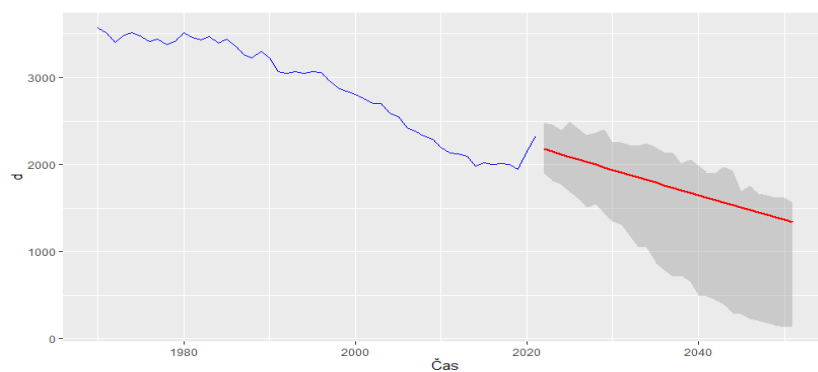
Vyšší hodnoty intervalového skóre jsou způsobené především tím, že někdy predikční interval nepokrývá skutečnou pozorovanou hodnotu. Nejvíce nepřesností pozorujeme v letech 2020 a 2021, kde počty zemřelých významně stouply i kvůli pandemii Covid-19.

Celkově volíme jako finální model pro muže PCA model s pěti hlavními komponentami a náhodnou procházkou s driftem, protože je nejlepší podle MAPE a hodnota intervalového skóre není příliš vysoká. U žen volíme PCA model s jednou hlavní komponentou a náhodnou procházkou s driftem ze stejného důvodu.

Na obrázcích 7.45 a 7.46 si můžeme prohlédnout pozorované hodnoty $d_{x,t}$ včetně jejich budoucích hodnot (získaných PCA-RWD modelem s 5 hlavními komponentami) a intervalového bootstrap odhadu pro vybranou věkovou skupinu (75 let) získaných finálními modely.

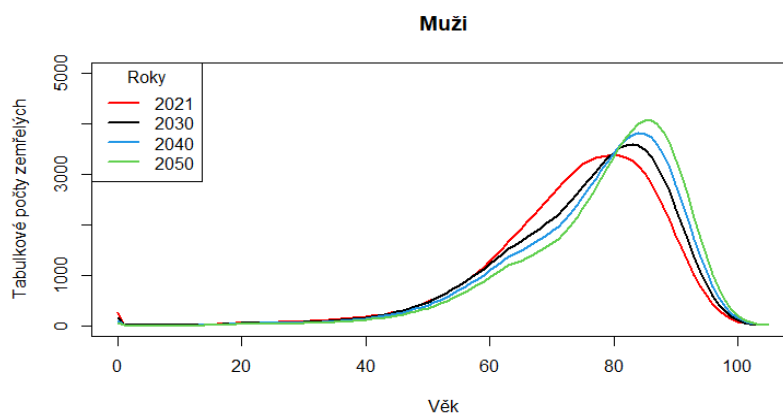


Obrázek 7.45: Pozorované hodnoty $d_{x,t}$ včetně jejich budoucích hodnot (získaných PCA-RWD modelem s 5 hlavními komponentami) a intervalového bootstrap odhadu pro vybranou věkovou skupinu 75 let mužů.

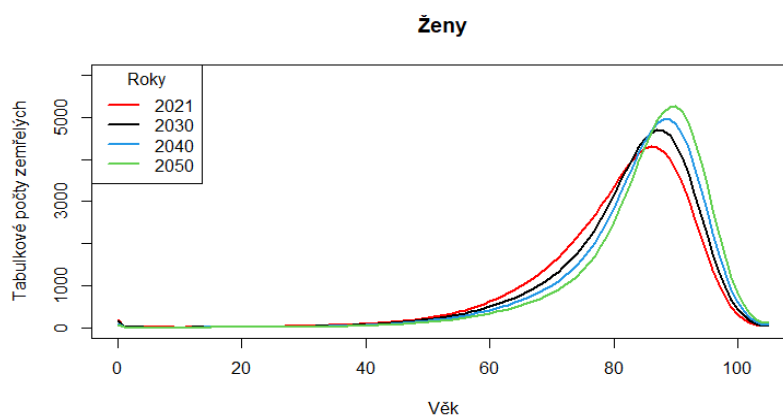


Obrázek 7.46: Pozorované hodnoty $d_{x,t}$ včetně jejich budoucích hodnot (získaných PCA-RWD modelem s 1 hlavní komponentou) a intervalového bootstrap odhadu pro vybranou věkovou skupinu 75 let žen.

Pro finální modely ještě uvedme rozdělení počtu úmrtí v roce 2021 a v predikovaných rocích 2030, 2040 a 2050 (Obrázky 7.47 a 7.48).



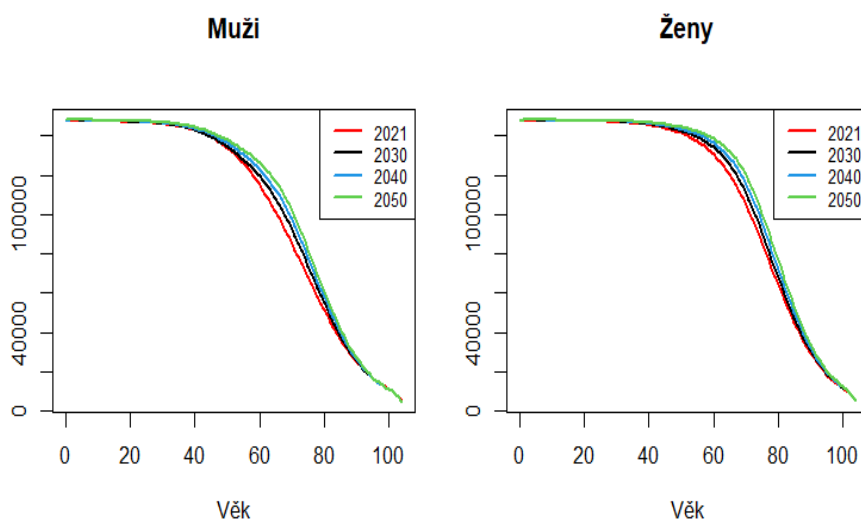
Obrázek 7.47: Rozdělení počtu úmrtí mužů v letech 2021, 2030, 2040 a 2050 pomocí PCA5-RWD modelu.



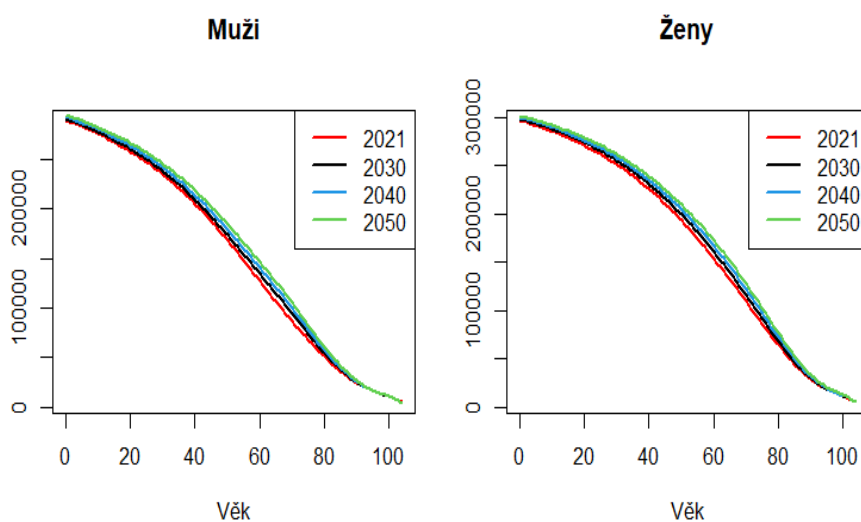
Obrázek 7.48: Rozdělení počtu úmrtí žen v letech 2021, 2030, 2040 a 2050 pomocí PCA1-RWD modelu.

7.5 Ocenění životních smluv

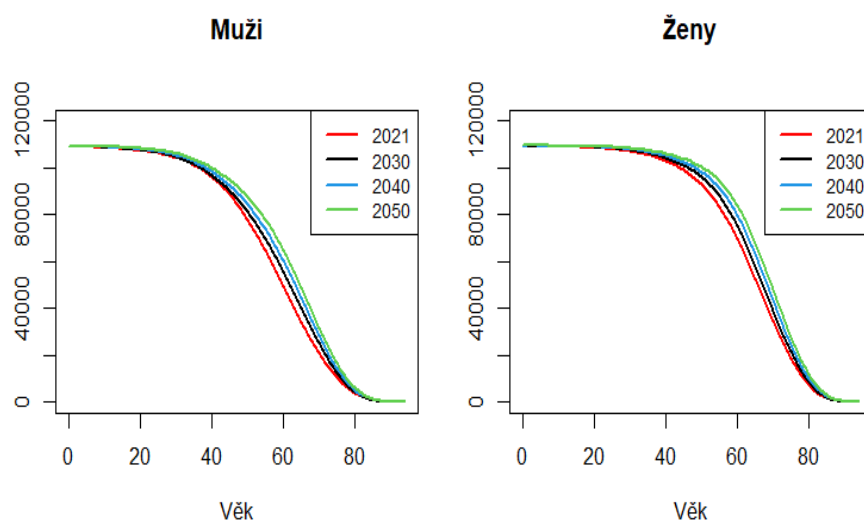
S vybranými finálními modely dopočteme pravděpodobnosti přežití p_x a ceny životních smluv představených v kapitole 6. U odložených smluv předpokládáme $m = 10$, u dočasných smluv bereme $n = 20$. Necht dále platí pro pojistné plnění $B = 10\,000$ a pro konstantní úrokovou míru $\mu = 0,03$. Na obrázcích 7.49, 7.50, 7.51 a 7.52 máme k nahlédnutí vývoj cen životních smluv v letech 2021, 2030, 2040 a 2050. Je patrné, že ceny produktů budou růst, což je způsobeno tím, že očekáváme, že se populace bude dožít vyšších věků.



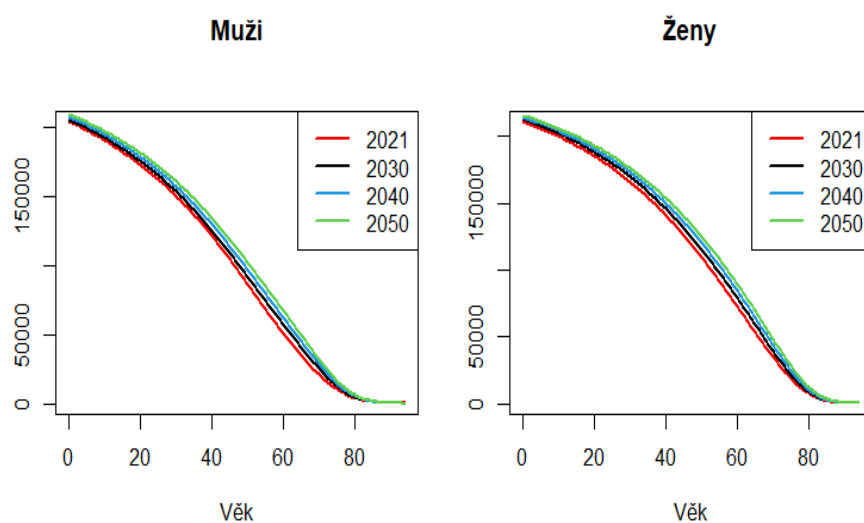
Obrázek 7.49: Dočasný okamžitý důchod v letech 2021, 2030, 2040 a 2050.



Obrázek 7.50: Doživotní okamžitý důchod v letech 2021, 2030, 2040 a 2050.



Obrázek 7.51: Dočasný odložený důchod v letech 2021, 2030, 2040 a 2050.



Obrázek 7.52: Doživotní odložený důchod v letech 2021, 2030, 2040 a 2050.

Mějme nyní portfolio o 10 000 smlouvách doživotního okamžitého důchodu, kde polovina jsou muži a polovina ženy, a jedinci jsou rovnoměrně ve věcích 50, 55, 60, 65 a 70 let. Abychom ukázali důležitost správné predikce počtu úmrtí, nastíníme si, jaký dopad to může mít. V tabulce 7.4 máme uvedeny ceny daného produktu pro vybrané věkové skupiny v závislosti na finálních modelech (PCA5-RWD pro muže a PCA1-RWD pro ženy) a na LC-SVD modelu v roce 2025. Celkový rozdíl na našem portfolio v roce 2025 by byl 63 179 277 Kč.

Věk	PCA5-RWD muži	LC-SVD muži	PCA1-RWD ženy	LC-SVD ženy
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 doživotního okamžitého důchodu pro vybrané věkové skupiny v závislosti na finálních modelech a LC-SVD modelu.

Závěr

V této práci jsme se zabývali modelováním úmrtnostních měř a počtu úmrtí. Nejprve jsme zavedli základní pojmy demografického modelu a úmrtnostních tabulek. Následně jsme se věnovali Leeovu–Carterovu modelu, kde jsme si model definovali, uvedli jsme tři různé metody pro odhad parametrů a popsali jsme, jak pomocí ARIMA modelů predikovat budoucí hodnoty. Obdobně jsme nabyli teoretické poznatky o Renshawovu-Habermanovu modelu, kde mnoho vlastností plyne z Leeova-Carterova modelu. Představili jsme si rovněž metodu od Shanga s Habermanem, jež využívá analýzu kompozičních dat. Zde jsme si popsali také ne-parametrické bootstrap metody pro intervalové odhady všech modelů.

V praktické části jsme na česká data úmrtí mužů a žen aplikovali všechny modely, odhadli jsme je a předpověděli budoucí hodnoty. Zjistili jsme, že všechny modely popisovaly data velmi dobře, neboť žádný z modelů neměl hodnotu koeficientu determinace nižší než 99.96 %. Největší nepřesnosti modelů byly pro věky 100-105, což bylo způsobené chybějícími daty pro tyto věky až do roku 2006. Největší rozdíly jsme pozorovali u modelů s Poissonovou metodou, neboť tento algoritmus pracoval s nedostupnými daty D_x a N_x . Z důvodu nedostupnosti některých dat a zároveň nedostatku dat pro tyto věkové skupiny bychom výsledky modelů těchto věkových skupin nebrali příliš vážně.

Na základě střední absolutní procentuální chyby jsme našli nejlepší modely - pro data mužů model PCA s pěti hlavními komponentami, kde pro predikci jsme využili náhodnou procházku s driftem. Pro ženy vyšel nejlépe PCA model s jednou hlavní komponentou a s metodou náhodné procházky s driftem pro predikování.

Zjistili jsme, že nový model využívající analýzy kompozičních dat poskytuje velmi dobré výsledky v porovnání s Leeovým–Carterovým a Renshawovým–Habermanovým modelem. Renshawův–Habermanův model vylepšuje predikce Leeova–Carterova modelu, vyjma modelu s Poissonovou metodou pro ženy. Přidáním dalších hlavních komponent do PCA modelu se predikce zlepšily pouze na datech pro muže kromě modelu s ARIMA modelem pro predikci, nejvýraznější zlepšení pozorujeme u náhodné procházky s driftem.

Nejllepší model podle intervalového skóre je pro muže trochu překvapivě PCA model s pěti hlavními komponentami s náhodnou procházkou bez driftu a pro ženy PCA model s jednou hlavní komponentou a náhodnou procházkou bez driftu. PCA modely s náhodnou procházkou bez driftu mají ale velmi nepřesné bodové odhady, a proto i přes nízké intervalové skóre nejsou vhodnými modely. Renshawův–Habermanův model je i zde obecně lepší než Leeův–Carterův model.

Jako finální modely jsme vybrali PCA model s pěti hlavními komponentami a náhodnou procházkou s driftem pro muže a PCA model s jednou hlavní komponentou a s náhodnou procházkou s driftem pro ženy. Na základě těchto modelů jsme určili cenu životních smluv - dočasný okamžitý důchod, doživotní okamžitý důchod, dočasný odložený důchod a doživotní odložený důchod v roce 2021 a v budoucnu. Pozorujeme, že cena těchto smluv poroste, což je způsobené postupným prodlužováním délky života populace.

Na závěr jsme na konkrétním portfoliu o 10 000 smlouvách ukázali důležitost přesnosti budoucích počtů úmrtí, když rozdíl mezi predikcemi na základě finálních modelů a LC-SVD modelu činil přes 63 milionů korun.

Seznam použité 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 TŮMA, 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 forecasting 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 demografie a pojištění*. 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/10256/742/0eppen?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-specific 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.
- ČESKÝ STATISTICKÝ ÚŘAD (2023). Český statistický úřad. Online. URL <https://www.czso.cz/>. Datum přístupu: 20. dubna 2023.

Seznam obrázků

7.1	Porovnání tabulkových počtů zemřelých mužů ve věku 25, 50, 75 a 100 let.	37
7.2	Porovnání tabulkových počtů zemřelých žen ve věku 25, 50, 75 a 100 let.	37
7.3	Porovnání úmrtnostních měř mužů z let 1970, 1980, 1990, 2000, 2010 a 2020.	38
7.4	Porovnání úmrtnostních měř žen z let 1970, 1980, 1990, 2000, 2010 a 2020.	38
7.5	Porovnání tabulkových počtů zemřelých mužů z let 1970, 1980, 1990, 2000, 2010 a 2020.	39
7.6	Porovnání tabulkových počtů zemřelých žen z let 1970, 1980, 1990, 2000, 2010 a 2020.	39
7.7	Finální odhady parametrů v LC-SVD modelu pro muže.	40
7.8	Finální odhady parametrů v LC-SVD modelu pro ženy.	40
7.9	Porovnání pozorovaných a odhadnutých měř úmrtností pro muže LC-SVD modelem.	41
7.10	Porovnání pozorovaných a odhadnutých měř úmrtností pro ženy LC-SVD modelem.	41
7.11	Finální odhady parametrů v LC-Poisson modelu pro muže.	42
7.12	Finální odhady parametrů v LC-Poisson modelu pro ženy.	42
7.13	Porovnání pozorovaných a odhadnutých měř úmrtností pro muže LC-Poisson modelem.	43
7.14	Porovnání pozorovaných a odhadnutých měř úmrtností pro ženy LC-Poisson modelem.	43
7.15	Finální odhady parametrů v LC-alternativním modelu pro muže.	44
7.16	Finální odhady parametrů v LC-alternativním modelu pro ženy.	44
7.17	Porovnání pozorovaných a odhadnutých měř úmrtností pro muže LC-alternativním modelem.	45
7.18	Porovnání pozorovaných a odhadnutých měř úmrtností pro ženy LC-alternativním modelem.	45
7.19	Finální odhady parametrů v RH-SVD modelu pro muže.	46
7.20	Finální odhady parametrů v RH-SVD modelu pro ženy.	46
7.21	Porovnání pozorovaných a odhadnutých měř úmrtností pro muže RH-SVD modelem.	47
7.22	Porovnání pozorovaných a odhadnutých měř úmrtností pro ženy RH-SVD modelem.	47
7.23	Finální odhady parametrů v RH-Poisson modelu pro muže.	48
7.24	Finální odhady parametrů v RH-Poisson modelu pro ženy.	48
7.25	Porovnání pozorovaných a odhadnutých měř úmrtností pro muže RH-Poisson modelem.	49
7.26	Porovnání pozorovaných a odhadnutých měř úmrtností pro ženy RH-Poisson modelem.	49
7.27	Finální odhady parametrů v PCA modelu s jednou hlavní komponentou pro muže a ženy.	50

7.28	Porovnání pozorovaných transformovaných dat $z_{x,t}$ a odhadnutých parametrů $\hat{z}_{x,t}$ pro muže PCA modelem s 1 hlavní komponentou. .	51
7.29	Porovnání pozorovaných transformovaných dat $z_{x,t}$ a odhadnutých parametrů $\hat{z}_{x,t}$ pro ženy PCA modelem s 1 hlavní komponentou. .	51
7.30	Porovnání pozorovaných transformovaných dat $z_{x,t}$ a odhadnutých parametrů $\hat{z}_{x,t}$ pro muže PCA modelem s 5 hlavními komponentami.	52
7.31	Porovnání pozorovaných transformovaných dat $z_{x,t}$ a odhadnutých parametrů $\hat{z}_{x,t}$ pro ženy PCA modelem s 5 hlavními komponentami.	52
7.32	Odhadnuté a budoucí hodnoty parametru \mathbf{k} včetně 95% intervalu spolehlivosti a pozorované a odhadnuté budoucí hodnoty úmrtnostních měř včetně 95% intervalu spolehlivosti - LC model SVD metoda. Vlevo grafy pro muže, vpravo pro ženy.	53
7.33	Odhadnuté a budoucí hodnoty parametru \mathbf{k} včetně 95% intervalu spolehlivosti a pozorované a odhadnuté budoucí hodnoty úmrtnostních měř včetně 95% intervalu spolehlivosti - LC model Poissonova metoda. Vlevo grafy pro muže, vpravo pro ženy.	54
7.34	Odhadnuté a budoucí hodnoty parametru \mathbf{k} včetně 95% intervalu spolehlivosti a pozorované a odhadnuté budoucí hodnoty úmrtnostních měř včetně 95% intervalu spolehlivosti - LC model alternativní metoda. Vlevo grafy pro muže, vpravo pro ženy.	55
7.35	Odhadnuté a budoucí hodnoty parametru $\mathbf{k}^{(1)}$ a $\mathbf{k}^{(2)}$ včetně 95% intervalu spolehlivosti - RH model, SVD metoda. Vlevo grafy pro muže, vpravo pro ženy.	56
7.36	Pozorované a odhadnuté budoucí hodnoty úmrtnostních měř \mathbf{m} včetně 95% intervalu spolehlivosti - RH model, SVD metoda. První graf pro muže 75 let, druhý pro ženy 75 let.	56
7.37	Odhadnuté a budoucí hodnoty parametru $\mathbf{k}^{(1)}$ a $\mathbf{k}^{(2)}$ včetně 95% intervalu spolehlivosti - RH model, Poissonova metoda. Vlevo grafy pro muže, vpravo pro ženy.	57
7.38	Pozorované a odhadnuté budoucí hodnoty úmrtnostních měř \mathbf{m} včetně 95% intervalu spolehlivosti - RH model, Poissonova metoda. První graf pro muže 75 let, druhý pro ženy 75 let.	57
7.39	Budoucí hodnoty $\beta_{1,t_{n+h}}$ mužů na základě PCA modelu, v závislosti na predikční metodě.	58
7.40	Budoucí hodnoty $\beta_{1,t_{n+h}}$ žen na základě PCA modelu, v závislosti na predikční metodě	59
7.41	Budoucí hodnoty $d_{x,t_{n+h}}$ mužů na základě PCA modelu v závislosti na predikční metodě pro vybrané věky 25, 50, 75 a 100 let.	59
7.42	Budoucí hodnoty $d_{x,t_{n+h}}$ žen na základě PCA modelu v závislosti na predikční metodě pro vybrané věky 25, 50, 75 a 100 let.	60
7.43	Budoucí hodnoty $d_{x,t_{n+h}}$ mužů na základě PCA modelu s 5 hlavními komponentami v závislosti na predikční metodě pro vybrané věky 25, 50, 75 a 100 let.	61
7.44	Budoucí hodnoty $d_{x,t_{n+h}}$ žen na základě PCA modelu s 5 hlavními komponentami v závislosti na predikční metodě pro vybrané věky 25, 50, 75 a 100 let.	61

7.45	Pozorované hodnoty $d_{x,t}$ včetně jejich budoucích hodnot (získaných PCA-RWD modelem s 5 hlavními komponentami) a intervalového bootstrap odhadu pro vybranou věkovou skupinu 75 let mužů. . . .	63
7.46	Pozorované hodnoty $d_{x,t}$ včetně jejich budoucích hodnot (získaných PCA-RWD modelem s 1 hlavní komponentou) a intervalového bootstrap odhadu pro vybranou věkovou skupinu 75 let žen.	64
7.47	Rozdělení počtu úmrtí mužů v letech 2021, 2030, 2040 a 2050 pomocí PCA5-RWD modelu.	64
7.48	Rozdělení počtu úmrtí žen v letech 2021, 2030, 2040 a 2050 pomocí PCA1-RWD modelu.	64
7.49	Dočasný okamžitý důchod v letech 2021, 2030, 2040 a 2050. . . .	65
7.50	Doživotní okamžitý důchod v letech 2021, 2030, 2040 a 2050. . . .	65
7.51	Dočasný odložený důchod v letech 2021, 2030, 2040 a 2050. . . .	66
7.52	Doživotní odložený důchod v letech 2021, 2030, 2040 a 2050. . . .	66