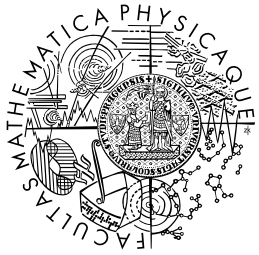


FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Bc. Boris Valter

Modelling mortality by causes of death

Department of Probability and Mathematical Statistics

Supervisor of the master thesis:	RNDr. Lucie Mazurová, Ph.D.
Study programme:	Mathematics
Study branch:	Financial and Insurance Mathematics

Prague 2020

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ..... signature of the author

Foremost, I would like to express my gratitude to my supervisor, RNDr. Lucie Mazurová, Ph.D., for the professional guidance and for the continuous support during the work on this thesis. I also wish to thank Doc. RNDr. Jan Hurt, CSc. for insightful suggestions and help with `Mathematica` software. I am particularly grateful for the assistance given by my brother E. Valter with medicine-related topics.

Title: Modelling mortality by causes of death

Author: Bc. Boris Valter

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Lucie Mazurová, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The aim of this thesis is to provide an overview of methods used in cause-of-death mortality analysis and to demonstrate the application on real data. In Chapter 1 we present the continuous model based on the force of mortality and present the approach using copula functions. In Chapter 2 we focus on the multinomial logit model formulated for cause-specific mortality data. In Chapter 3 we apply the multinomial logit model on the data from Czech Statistical Office. We identify the regression model, check its assumptions, present the outputs including the fitted life expectancy, and predicted mortality rates. Later in Chapter 3 we consider several stress scenarios in order to demonstrate the impact of shocked mortality rates on the life expectancy. In Chapter 4 we apply copula functions and compare the results for different approaches to cause-of-death mortality modelling.

Keywords: Cause-of-death mortality, force of mortality, copulas, multinomial logit, regression, stress scenarios

Contents

Introduction

Mortality rates' modelling has always been essential in various aspects of life and in actuarial science in particular. Distinguishing between causes of death within a model, is definitely an improvement over a model which attributes death to a single cause. Competing risks framework was developed to provide an additional insight into this topic.

The aim of this thesis is to provide an overview of methods used in cause-of-death mortality analysis and to demonstrate the application on real data. The basic statistical measures of the death risk are survival probability, death probability, and life expectancy.

In Chapter 1, we first introduce the traditional approach based on the force of mortality and the estimation method which uses the data about current population as an input. Another approach based on copula functions is briefly reviewed later in this chapter. The latter method allows to incorporate the complex dependence structures into the model.

The main focus of Chapter 2 is the multinomial logistic model which also provides a framework for analysing cause-specific mortality. We also discuss the application of shocks in the multinomial logistic model.

In Chapter 3 we shall focus on the practical application of multinomial logistic regression. We will work with the data from Czech Statistical Office to construct cause-specific life tables in order to use them as an input for the regression model. Next, we shall assess the model and present the outputs. In addition, several scenarios and their impacts on life expectancy will be discussed. These scenarios are meant to demonstrate the model's response to some catastrophic (under Solvency II) or just adverse events that might take place.

In Chapter 4 we apply copula functions in order to evaluate net survival functions. Later in this chapter, we consider cause-elimination scenario and compare the results when using copula functions with the outputs based on the model from Chapter 3.

1. Continuous model

In this chapter we shall focus on the concept of competing risks in mortality analysis. We will present the traditional approach introduced in Chiang [1968] along with the estimation method based on the data about current population. Later in this chapter we shall discuss the life expectancy and the application of copula functions.

1.1 Competing risks in mortality analysis

In this section we present the competing risks framework based on Chiang [1968]. Let us assume a group of lives where every individual may die from one of n competing risks (causes of death). Let $(X_1, \ldots, X_n)^\top$ be a vector of potential lifetimes, where X_i denotes a lifetime of an individual provided that he would die from cause $i = 1, \ldots, n$. The actual lifetime Y of an individual is then given by

$$Y = \min\,(X_1, \ldots, X_n).$$

The absolutely continuous joint distribution function of n-dimensional random vector of potential lifetimes $\mathbf{X} = (X_1, \ldots, X_n)^\top$ is

$$F_{\mathbf{X}}(\mathbf{x}) = F_{\mathbf{X}}(x_1, \ldots, x_n).$$

The distribution function of the potential lifetime of an individual if death occurs from i-th cause is the (continuous) marginal distribution function of $\mathbf{X}$:

$$F_{X_i}(x) = \mathbb{P}(X_i \leq x).$$

The force of mortality of i-th risk is

$$\mu(x; i) = \frac{d\, F_{X_i}(x)/dx}{1 - F_{X_i}(x)}.$$

We further assume the independence of competing risks, i.e. competing risks of death are independent of one another in the sense that the force of mortality of each risk remains unchanged after one or more risks are eliminated or adjusted in a certain way. It is also possible to incorporate more complex dependence structures. The approach using copula functions will be briefly described in section 1.4.

It is essential to present three general types of probabilities of death with respect to a specific risk i in the age interval (x_j, x_{j+1}):

- The crude probability ... the probability of death from a specific cause in the presence of all other competing risks (Q_{ij});
- The net probability ... the probability of death if a specific risk is the only one risk in effect in the population (q_{ij}) or in absence of all other competing risks $(q_{i,j})$;
- The partial crude probability ... the probability of death from a specific cause i when another risk or risks are eliminated from the population.

Probabilities of death and survival in the interval (x_j, x_{j+1}) of an individual at age x_j are denoted as q_j and p_j, respectively, with $q_j + p_j = 1$.

In the presence of n causes of death, $i = 1, \ldots, n$, for each risk i there is a corresponding force of mortality $\mu(x;i)$ (cause-specific), such that $\mu(x;i)\,dx$ is the probability that an individual alive at age x will die from cause i in the infinitesimal time interval $(x, x+dx)$ for $i = 1, \ldots, n$. The **total** force of mortality $\mu(x)$ is such that $\mu(x)\,dx$ is equal to the probability that an individual alive at age x will die in $(x, x+dx)$. Under the assumption of independence between causes of death, the total force of mortality can be written as a sum of cause-specific forces of mortality:

$$\mu(x) = \sum_{i=1}^{n} \mu(x;i).$$

Probabilities of death and survival in the interval (x_j, x_{j+1}) of an individual at age x_j can be expressed by means of the total force of mortality:

$$q_j = 1 - \mathsf{exp}\left\{ -\int_{x_j}^{x_{j+1}} \mu(x)\,dx \right\},$$

$$p_j = \mathsf{exp}\left\{ -\int_{x_j}^{x_{j+1}} \mu(x)\,dx \right\}.$$

Another assumption is required for the theory of competing risks, namely the proportionality assumption. Under this assumption, in the interval (x_j, x_{j+1}), the following ratio

$$\frac{\mu(x;i)}{\mu(x)} = c_{ij} \tag{1.1}$$

is independent of x, but at the same time depends on the age interval and cause of death R_i.

If R_i is the only risk in effect in the population, the net probability of death is equal to

$$q_{ij} = 1 - \mathsf{exp}\left\{ -\int_{x_j}^{x_{j+1}} \mu(x;i)\,dx \right\}.$$

The crude probability of death discussed earlier, for $x_j \le x \le x_{j+1}$, is given by

$$Q_{ij} = \int_{x_j}^{x_{j+1}} \mathsf{exp}\left\{ -\int_{x_j}^{x} \mu(s)\,ds \right\} \mu(x;i)\,dx. \tag{1.2}$$

Using the assumption (1.1), the expression (1.2) can be rewritten as

$$\begin{aligned} Q_{ij} &= \frac{\mu(x;i)}{\mu(x)} \int_{x_j}^{x_{j+1}} \mathsf{exp}\left\{ -\int_{x_j}^{x} \mu(s)\,ds \right\} \mu(x)\,dx \\ &= \frac{\mu(x;i)}{\mu(x)} \left[1 - \mathsf{exp}\left\{ -\int_{x_j}^{x_{j+1}} \mu(x)\,dx \right\} \right] = \frac{\mu(x;i)}{\mu(x)} q_j. \end{aligned}$$

And thus

$$\frac{\mu(x;i)}{\mu(x)} = \frac{Q_{ij}}{q_j}. \tag{1.3}$$

1.2 Current population mortality analysis

In this section we shall focus on the current population mortality analysis introduced in Chiang [1968]. This method is used to estimate the probabilities described earlier in this chapter.

For the age interval(x_j, x_{j+1}), let $n_j = x_{j+1} - x_j$ be the length of the interval, E_j^c the midyear population state (also referred to as central exposure), D_j the total number of deaths , a_j is the average fraction of the age interval that each of the individuals survive before dying and E_j the unobserved population state at x_j (initial exposure). The age specific mortality rate is

$$m_j = \frac{D_j}{E_j^c}. \tag{1.4}$$

The estimator of the probability of death in the interval is given by

$$\widehat{q}_j = \frac{D_j}{E_j}, \tag{1.5}$$

where the initial exposure E_j can be estimated from

$$\widehat{E}_j = (E_j^c + (1 - a_j)n_j D_j)/n_j.$$

Thus, (1.5) transforms into

$$\widehat{q}_j = \frac{n_j m_j}{1 + (1 - a_j)n_j m_j}. \tag{1.6}$$

The corresponding survival probability is therefore

$$\widehat{p}_j = \frac{1 - a_j n_j m_j}{1 + (1 - a_j)n_j m_j}. \tag{1.7}$$

Switching over to the cause-specific probabilities, we use the fact that the total number of deaths is equal to the sum of cause-specific numbers of deaths:

$$D_j = \sum_{i=1}^{n} D_{ij}.$$

The cause-specific mortality rate is then given by

$$m_{ij} = \frac{D_{ij}}{E_j^c}.$$

Similarly, the crude probability of death is estimated from

$$\widehat{Q}_{ij} = \frac{D_{ij}}{E_j}, \tag{1.8}$$

which can be also expressed as

$$\widehat{Q}_{ij} = \frac{n_j m_{ij}}{1 + (1 - a_j)n_j m_j}. \tag{1.9}$$

Solving the equations (1.6) and (1.9) with respect to death rates m_j and m_{ij}, we get

$$m_j = \frac{\widehat{q}_j}{\widehat{q}_j a_j n_j + (1 - \widehat{q}_j) n_j},$$
$$m_{ij} = \frac{\widehat{Q}_{ij}}{\widehat{q}_j a_j n_j + (1 - \widehat{q}_j) n_j},$$

which implies that

$$\frac{m_{ij}}{m_j} = \frac{\widehat{Q}_{ij}}{\widehat{q}_j}.$$

Thus, the expression above is an analogy of the formula 1.3.

The expression 1.8 can be also obtained as maximum likelihood estimator of Q_{ij}, if we assume that for a given calendar year t, cause-specific number of deaths D_{ij} follows a binomial distribution $\mathrm{Bi}(E_j, Q_{ij})$ with probability mass function

$$\mathbb{P}(D_{ij} = d_{ij}) = \binom{E_j}{d_{ij}} Q_{ij}^{d_{ij}} (1 - Q_{ij})^{E_j - d_{ij}},$$

where E_j is the measure of initial exposure. The likelihood function is given by

$$\mathcal{L}(\mathbf{Q} \mid \mathbf{d}) = \prod_{i=1}^{n} \binom{E_j}{d_{ij}} Q_{ij}^{d_{ij}} (1 - Q_{ij})^{E_j - d_{ij}}.$$

The log-likelihood function is then of the form

$$\ell(\mathbf{Q} \mid \mathbf{d}) = \sum_{i=1}^{n} \left[\mathsf{ln} \binom{E_j}{d_{ij}} + d_{ij} \, \mathsf{ln}(Q_{ij}) + (E_j - d_{ij}) \mathsf{ln}(1 - Q_{ij}) \right].$$

The maximum likelihood estimate of $\mathbf{Q}$ can be obtained from

$$\hat{Q}_{ij} \equiv \arg\max_{Q_{ij}} \ell(\mathbf{Q} \mid \mathbf{d}).$$

Taking the derivative of log-likelihood with respect to Q_ij and setting it to zero we get

$$\frac{\partial \ell(\mathbf{Q} \mid \mathbf{d})}{\partial Q_{ij}} = 0 = \frac{d_{ij}}{Q_{ij}} - \frac{E_j - d_{ij}}{1 - Q_{ij}},$$

which leads to

$$\hat{Q}_{ij} = \frac{d_{ij}}{E_j}.$$

The above expression then corresponds to (1.8). It is essential to note that the likelihood function is concave and hence log-likelihood is indeed maximized at $\hat{Q}_{ij}$. Therefore, it is reasonable to consider the estimator given by (1.8). It can be also shown that MLE estimate of Q_{ij} is unbiased:

$$\mathsf{E}\, \hat{Q}_{ij} = \frac{\mathsf{E} D_{ij}}{E_j} = \frac{E_j \, Q_{ij}}{E_j} = Q_{ij}.$$

Consistency then follows from Chebyshev's inequality:

$$\begin{aligned}\mathbb{P}(|\hat{Q}_{ij} - Q_{ij}| \geq \varepsilon) &\leq \frac{\mathsf{var}\,\hat{Q}_{ij}}{\varepsilon^2} \\ &= \frac{\mathsf{var}\,D_{ij}}{E_j^2\,\varepsilon^2} \\ &= \frac{Q_{ij}(1-Q_{ij})}{E_j\,\varepsilon^2} \xrightarrow[E_j\to\infty]{} 0.\end{aligned}$$

1.3 Life expectancy

Another important statistical measure that we will focus on is life expectancy. In this section we shall use the actuarial notation and introduce the necessary theory based on Gerber [1997]. We denote by T_x the remaining lifetime at age x. T_x is a continuous random variable and expresses the exact future lifetime. The distribution function of T_x is

$$F_x(t) = \mathbb{P}(T \leq t), \;\; t \geq 0.$$

We assume that F_x is continuous and has a probability density function $f_x(t) = F_x'(t)$. The probability that an individual aged x will die within t years is denoted by ${}_tq_x$ and the respective probability of survival by ${}_tp_x$. The following relations then hold:

$$\begin{aligned}{}_tq_x &= F_x(t) \\ {}_tp_x &= 1 - {}_tq_x.\end{aligned}$$

The complete expectation of life is given by

$$\begin{aligned}e_x^0 = \mathsf{E}T_x &= \int_0^\infty t \cdot f_x(t)\,dt \\ &\overset{\text{PP}}{=} \left[-t\cdot(1-F_x(t))\right]_0^\infty + \int_0^\infty \left[1-F_x(t)\right]\,dt \\ &= \int_0^\infty \left[1-F_x(t)\right]\,dt \\ &= \int_0^\infty {}_tp_x\,dt.\end{aligned}$$

The curtate remaining lifetime is defined by $K_x = \lfloor T_x \rfloor$ and expresses the number of future years completed prior to death or, in other words, the greatest integer of T_x. Its probability mass function is

$$\begin{aligned}\mathbb{P}(K_x = k) &= \mathbb{P}(k \leq T_x < k+1) \\ &= \mathbb{P}(k < T_x \leq k+1) \quad \text{[by continuity of } T_x\text{]} \\ &= {}_kp_x \cdot q_{x+k}.\end{aligned}$$

Another alternative expression can be obtained in order to derive the expectation of K_x:

$$\begin{aligned}\mathbb{P}(K_x = k) &= \mathbb{P}(k < T_x \leq k+1) \\ &= \mathbb{P}(T_x > k) - \mathbb{P}(T_x > k+1) \\ &= {}_kp_x - {}_{k+1}p_x.\end{aligned}$$

The curtate expectation of life can be then computed as

$$\begin{aligned}e_x = \mathsf{E}K_x &= \sum_{k=0}^{\infty} k \cdot \mathbb{P}(K_x = k) \\ &= \sum_{k=0}^{\infty} k \cdot {}_kp_x - \sum_{k=0}^{\infty} k \cdot {}_{k+1}p_x \\ &= \sum_{k=1}^{\infty} k \cdot {}_kp_x - \sum_{k=1}^{\infty} (k-1) \cdot {}_kp_x,\end{aligned}$$

which finally reduces to

$$e_x = \sum_{k=1}^{\infty} \mathbb{P}(K_x \geq k) = \sum_{k=1}^{\infty} {}_kp_x.$$

Obviously, by definition of K_x, it holds

$$K_x \leq T_x \leq K_x + 1$$

and hence also

$$e_x \leq e_x^0 \leq e_x + 1.$$

Even though there is no explicit relationship between the complete and curtate expectation of life, the reasonable approximation can be achieved. Under the assumption of linearity of ${}_uq_x$, i.e. ${}_uq_x = uq_x$ for $u \in [0,1)$:

$$\begin{aligned}{}_{k+u}p_x = {}_kp_x \cdot {}_up_{x+k} &= (1 - {}_kq_x)(1 - {}_uq_{x+k}) \\ &= 1 - u \cdot q_{x+k} - {}_kq_x + {}_kq_x \cdot u \cdot q_{x+k} \\ &= {}_kp_x - u \cdot q_{x+k}(1 - {}_kq_x) \\ &= {}_kp_x - u \cdot (1 - p_{x+k}) \cdot {}_kp_x \\ &= (1-u)_kp_x + u \cdot p_{x+k} \cdot {}_kp_x \\ &= (1-u)_kp_x + u \cdot {}_{k+1}p_x, \quad u \in [0,1).\end{aligned}$$

The complete expectation of life can be then approximated by using the relation above:

$$\begin{aligned}e_x^0 = \int_0^{\infty} {}_tp_x \, dt &= \sum_{k=0}^{\infty} \int_k^{k+1} {}_tp_x \, dt \\ &= \sum_{k=0}^{\infty} \int_0^1 {}_{k+u}p_x \, du \\ &= \sum_{k=0}^{\infty} \left({}_kp_x \int_0^1 (1-u) \, du + {}_{k+1}p_x \int_0^1 u \, du \right),\end{aligned}$$

which can be further simplified to obtain

$$\begin{aligned} e_x^0 &= \frac{1}{2}\sum_{k=0}^{\infty} {}_kp_x + \frac{1}{2}\sum_{k=0}^{\infty} {}_{k+1}p_x \\ &= \frac{1}{2}\left(1 + \sum_{k=1}^{\infty} {}_kp_x\right) + \frac{1}{2}\sum_{k=1}^{\infty} {}_kp_x \\ &= \frac{1}{2} + \sum_{k=1}^{\infty} {}_kp_x = e_x + \frac{1}{2}. \end{aligned}$$

1.4 Competing risks and copula functions

As pointed out earlier in this chapter, competing risks do not necessarily act independently. In order to capture the dependence structure, one may use copula functions. In this section, we shall focus on several basic properties of copula functions, measures of association, Archimedean copulas and provide a brief overview of the method which will be applied in Chapter 4.

1.4.1 Basic properties

The following basic properties, definitions and theorems are based on McNeil et al. [2005].

Definition (copula). *A d-dimensional copula is a distribution function of a d-dimensional vector, for which all univariate distributions are uniform on $[0,1]$.*

Equivalently, copula C is a mapping of the form $C : [0,1]^d \to [0,1]$, satisfying the following conditions:

- *$C(u_1, \ldots, u_d)$ is increasing in each component u_i.*
- *$C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$ for all $i = 1, \ldots, d$, $u_i \in [0,1]$.*
- *For all $\left(u_1^{(1)}, \ldots, u_d^{(1)}\right)$, $\left(u_1^{(2)}, \ldots, u_d^{(2)}\right)$ in $[0,1]^d$ such that $u_i^{(1)} \leq u_i^{(2)}$ for all $i = 1, \ldots, d$, it holds*

$$\sum_{i_1=1}^{2} \cdots \sum_{i_d}^{2} (-1)^{i_1+\cdots+i_d} C\left(u_1^{(i_1)}, \ldots, u_d^{(i_d)}\right) \geq 0.$$

The following theorem is fundamental in copula theory.

Theorem (Sklar's theorem). *Let F be a joint d.f. with marginal distribution functions $F_1, \ldots F_d$. Then there exists a copula $C : [0,1]^d \to [0,1]$ such that for all $x_1, \ldots x_d \in [-\infty, +\infty]$,*

$$F(x_1, \ldots x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \tag{1.10}$$

If the marginal distributions are continuous, then C is unique. Otherwise, C is uniquely determined in $\mathsf{Ran}(F_1) \times \cdots \times \mathsf{Ran}(F_d)$, where $\mathsf{Ran}(F_i)$ denotes the range of F_i.

Conversely, if C is a copula and $F_1, \ldots, F_d$ are univariate distribution functions, then the function F defined in (1.10) is a joint distribution function with marginals $F_1, \ldots, F_d$.

The following theorem sets the bounds for any copula function.

Theorem (The Fréchet-Hoeffding Bounds). *For every copula $C(u_1, \ldots, u_d)$ we have the bounds*

$$\max\left\{1 - d + \sum_{i=1}^{d} u_i,\, 0\right\} \leq C(\mathbf{u}) \leq \min\{u_1, \ldots, u_d\}.$$

The Fréchet upper bound is called the comonotonicity copula and the lower bound in the bivariate case ($d = 2$) is often referred to as countermonotonicity copula.

1.4.2 Archimedean copulas

In this section we shall provide a brief introduction to Archimedean copulas based on McNeil et al. [2005] and describe Ali-Mikhail-Haq (AMH) copula based on Pranesh [2010].

Definition (Pseudo-inverse). *Suppose that $\phi : [0,1] \to [0,\infty]$ is continuous and strictly decreasing with $\phi(1) = 0$ and $\phi \leq \infty$. A pseudo-inverse of ϕ with domain $[0,\infty]$ is defined by*

$$\phi^{[-1]}(t) = \begin{cases} \phi^{-1}(t), & 0 \leq t \leq \phi(0) \\ 0, & \phi(0) \leq t \leq \infty. \end{cases}$$

Theorem (Bivariate Archimedean copula). *Let $\phi : [0,1] \to [0,\infty]$ be continuous and strictly decreasing with $\phi(1) = 0$ and pseudo-inverse $\phi^{[-1]}(t)$. Then*

$$C(u_1, u_2) = \phi^{[-1]}(\phi(u_1) + \phi(u_2)) \tag{1.11}$$

is a copula if and only if ϕ is convex.

Copulas which are constructed according to 1.11 are called Archimedean copulas and the function ϕ from the previous theorem is called Archimedean copula generator.

Clayton copula is an Archimedean copula defined by

$$C^{CL}(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}, \quad 0 < \theta < \infty.$$

For $\theta \to 0+$ and $\theta \to \infty$ we obtain the independence and the comonotonicity copulas, respectively. The Fréchet lower bound cannot be reached in the case of the Clayton copula and, therefore, only positive dependence can be modelled with this copula. Clayton copula generator is given by

$$\phi(x) = \frac{1}{\theta}(x^{-\theta} - 1).$$

Ali-Mikhail-Haq copula is another example of an Archimedean copula defined by

$$C^{AMH}(u_1, u_2) = \frac{u_1 u_2}{1 - \theta(1 - u_1)(1 - u_2)}, \quad \theta \in [-1, 1].$$

AMH copula allows for modelling both positive and negative dependence. AMH copula generator is given by

$$\phi(x) = \frac{1}{x}\mathsf{ln}[1 - \theta(1 - x)].$$

1.4.3 Rank correlations

For the purposes of this section we consider a bivariate random vector with continuous and strictly increasing marginals and copula C. We shall focus on the most commonly used rank correlation measures, namely Kendall's tau and Spearman's rho. The definitions are based on McNeil et al. [2005].

Definition (Kendall's tau) *For random variables X_1 and X_2 Kendall's tau is defined as*

$$\tau(X_1, X_2) = \mathsf{E}\left[\mathsf{sign}\left(X_1 - \tilde{X}_1\right)\left(X_2 - \tilde{X}_2\right)\right],$$

where $\left(\tilde{X}_1, \tilde{X}_2\right)^\top$ is an independent copy of $(X_1, X_2)^\top$

Given n independent observations $(x_1, y_1), \ldots, (x_n, y_n)$ of the random vector (X, Y), Kendall's tau can be expressed as follows:

$$\widehat{\tau} = \frac{2}{n(n-1)} \sum_{i<j} \mathsf{sgn}(x_i - x_j)\mathsf{sgn}(y_i - y_j),$$

The following relationship holds between the Kendall's tau and the generator of and Archimedean copula (Pranesh [2010]):

$$\tau = 1 + 4 \int_0^1 \frac{\phi(x)}{\phi'(x)}\, dx. \tag{1.12}$$

Using the formula 1.12, one can derive the relations between the Kendall's tau and parameters of Archimedean copulas:

- For the Clayton copula $\tau = \frac{\theta}{\theta+2}$,
- For the AMH copula $\tau = \frac{3\theta-2}{3\theta} - \frac{2(1-\theta)^2\mathsf{ln}(1-\theta)}{3\theta^2}$.

However, the last relationship holds for $\tau \in \left[\frac{5-8\,\mathsf{ln}2}{3}, \frac{1}{3}\right] \approx [-0.1817, 0.3333]$ and thus AMH copula allows to model only a weak dependence in terms of Kendall's tau.

Definition (Spearman's rho) *For random variables X_1 and X_2 with distribution functions F_1 and F_2, Spearman's rho is defined as*

$$\rho_S(X_1, X_2) = \rho\left(F_1(X_1), F_2(X_2)\right),$$

where ρ denotes Pearson (linear) correlation coefficient.

The relationship between the Spearman's rho and parameters of Archimedean copulas is often relatively complicated (e.g. for Clayton or AMH copulas).

1.4.4 Methodology

In this section we shall provide a brief overview of the method based on Kaishev et al. [2007]. We recall that competing risks framework operates with a vector of potential lifetimes $\mathbf{X} = (X_1, \ldots, X_n)^\top$ assigned to an individual with respect to causes of death $i = 1, \ldots, n$. In practise, however, we observe only actual lifetime, which is equal to min $(X_1, \ldots, X_n)$. The joint distribution function of $\mathbf{X}$ is given by

$$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

and the joint survival function is

$$S(x_1, \ldots, x_n) = \mathbb{P}(X_1 > x_1, \ldots, X_n > x_n).$$

The crude survival function is a cause-specific survival function in the presence of all other competing risks in the population:

$$S^{(i)}(x) = \mathbb{P}(\min\,(X_1, \ldots X_n) > x, \min\,(X_1, \ldots X_n) = X_i)$$

and it obviously holds that

$$S(x, \ldots, x) = S^{(1)}(x) + \cdots + S^{(n)}(x).$$

The net survival function is a survival function in the presence of only one risk in effect:

$$S'^{(i)}(x) = \mathbb{P}(X_i > x).$$

Under the assumption of independence, the joint survival function can be expressed as

$$S(x_1, \ldots, x_n) = S'^{(1)}(x_1) \times \cdots \times S'^{(n)}(x_n).$$

Nevertheless, random variables $X_1, \ldots X_n$ will be considered stochastically dependent and non-defective in the sense that $\mathbb{P}(X_i < \infty) = 1$.

By Sklar's theorem, there exists a unique n-dimensional copula C such that

$$F(x_1, \ldots x_n) = C(F_{X_1}(x_1), \ldots, F_{X_n}(x_n))$$

and the joint survival function of X_i is uniquely determined by

$$S(x_1, \ldots x_n) = \bar{C}(S'^{(1)}(x_1), \ldots, S'^{(n)}(x_n)), \tag{1.13}$$

where $\bar{C}$ is the survival copula with respect to copula C. In the bivariate case the following relationship holds

$$\bar{C}(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2).$$

Therefore, the dependence structure can be incorporated by choosing a suitable copula function and estimating its parameters.

Given the copula function $\bar{C}(u_1, \ldots, u_n)$ and the net survival functions $S'^{(i)}(x_i)$, $i = 1, \ldots, n$, the joint survival function 1.13 can be evaluated. The following lemma formulated in Carriere [1994], provides an important representation of the crude survival function.

Lemma. *If $S(x_1, \ldots, x_n)$ is differentiable with respect to $x_i > 0$ for all $i = 1, \ldots, n$, then*

$$S^{(i)}(x) = \int_x^\infty -S_i(r, \ldots, r)\, dr,$$

where

$$S_i(r, \ldots, r) = \frac{\partial}{\partial x_i} S(x_1, \ldots, x_n)|_{x_k = r, \forall k}.$$

Using the above lemma and applying the chain rule for 1.13, the following theorem is obtained in Carriere [1994].

Theorem. *If $\bar{C}(u_1, \ldots, u_n)$ is differentiable with respect to $u_i \in (0, 1)$ and $S'^{(i)}(x_i)$ is differentiable with respect to $x_i > 0$ for all $i = 1, \ldots, n$, then*

$$\begin{aligned}
\frac{d}{dx} S^{(1)}(x) &= \bar{C}_1(S'^{(1)}(x), \ldots, S'^{(n)}(x)) \times \frac{d}{dx} S'^{(1)}(x) \\
\frac{d}{dx} S^{(2)}(x) &= \bar{C}_2(S'^{(1)}(x), \ldots, S'^{(n)}(x)) \times \frac{d}{dx} S'^{(2)}(x) \\
&\vdots \\
\frac{d}{dx} S^{(n)}(x) &= \bar{C}_n(S'^{(1)}(x), \ldots, S'^{(n)}(x)) \times \frac{d}{dx} S'^{(n)}(x),
\end{aligned} \tag{1.14}$$

where

$$\bar{C}_i(u_1, \ldots, u_n) = \frac{\partial}{\partial u_i} \bar{C}(u_1, \ldots, u_n).$$

The system of non-linear differential equations given by 1.14 can be then solved numerically with respect to the net survival functions $S'^{(i)}(x)$, given the selected copula $\bar{C}(u_1, \ldots, u_n)$ and the estimates of $S^{(i)}(x)$ in a functional form, e.g. splines or regression curve. The estimates of $S^{(i)}(x)$ can be then substituted into 1.14 to evaluate left-hand sides of the system. Therefore, the joint survival function, as well as the overall survival function $S(x, \ldots, x)$, can be evaluated by substituting the net survival functions into 1.13.

In order to show the partial and the complete cause elimination effect, it is essential to add an age subscript for the net and crude survival functions. We further assume that survival functions will be taken over integral years $x \equiv k = 1, 2, \ldots, 110$. The net survival functions at birth can be then expressed as

$$S_0'^{(i)}(k) = S_0'^{(i)}(1) \times S_1'^{(i)}(1) \times \cdots \times S_{k-1}'^{(i)}(1),$$

which can be rewritten using actuarial symbols (omitting index i) as

$$S_0'^{(i)}(k) = p'_0 \times p'_1 \times \cdots \times p'_{k-1} = (1 - q'_0) \times (1 - q'_1) \times \cdots \times (1 - q'_{k-1}).$$

Thus, the partial cause elimination, or, generally speaking, modification, impact can be captured by setting $q''_l = \rho_l \cdot q'_l$, $l = 0, 1, 2, \ldots, k-1$, where values of $\rho_l \geq 0$ greater than one correspond to increased probabilities of death. The modified net survival function is then given by

$$S_0''^{(i)}(k) = (1 - q''_0) \times (1 - q''_1) \times \cdots \times (1 - q''_{k-1}).$$

The overall survival function is then of the form

$$S(k,\ldots,k) = \bar{C}(S_0^{'(1)}(k),\ldots,S_0^{'(i-1)}(k),S_0^{''(i)}(k),S_0^{'(i+1)}(k),\ldots,S_0^{'(n)}(k)).$$

The complete elimination of the i-th cause of death ($\rho_l = 0$) corresponds to the following expression for the overall survival function:

$$S(k,\ldots,k) = \bar{C}(S_0^{'(1)}(k),\ldots,S_0^{'(i-1)}(k),1,S_0^{'(i+1)}(k),\ldots,S_0^{'(n)}(k)).$$

2. Multinomial regression

In this chapter we are going to introduce the multinomial logistic model based on Alai et al. [2015]. The model extends logistic regression framework to multiclass problems and allows to predict the probabilities of more than two possible outcomes of the dependent variable for a given set of covariates, which can be either numerical or categorical ones.

In order to formulate the problem of cause-of-death mortality in terms of multinomial logistic regression (MLR), we shall use the following notation:

- $D_i(x,t)$ ... cause-specific deaths at age x and at time t;
- $L(x,t)$ ... underlying survivors.

The data for n causes of death can be then represented by

$$Y(x,t) = (D_1(x,t), D_2(x,t), \ldots, D_n(x,t), L(x,t))^\top.$$

We assume that $Y(x,t)$ follows a multinomial distribution with probability mass function for a given x and t

$$\mathbb{P}(D_1 = d_1, \ldots, D_n = d_n, L = l) = \frac{E!}{d_1! \cdots d_n! \cdot l!} q_1^{d_1} \cdots q_n^{d_n} p^l,$$

where

$$\sum_{i=1}^{n} q_i(x,t) + p(x,t) = 1 \tag{2.1}$$

and $q_i(x,t)$ denotes cause-specific probabilities of death, $p(x,t)$ stands for probability of survival and

$$E(x,t) = l(x,t) + \sum_{i=1}^{n} d_i(x,t),$$

where $d_i(x,t)$ are observed cause-specific numbers of deaths and $l(x,t)$ are respective numbers of survivors. Setting probability of survival as a reference category, the problem can be then formulated in terms of the multinomial logistic regression as follows:

$$\mathsf{ln}\frac{q_i(x,t)}{p(x,t)} = \mathbf{X}(x,t)\beta_i, \quad i = 1, \ldots, n \tag{2.2}$$

where $\mathbf{X}(x,t)$ is the corresponding row of the model matrix and β_i is cause-specific vector of regression coefficients. The expression above is often referred to as linear predictor, i.e. the covariates are linearly related to the log-odds of the response, and we shall call it, for simplicity, $\mathsf{logit}(mortality)$. To obtain predicted probabilities we exponentiate and rewrite (2.2) in terms of the sequence of binary models:

$$\begin{aligned} q_1(x,t) &= p(x,t)e^{\mathbf{X}(x,t)\beta_1} \\ q_2(x,t) &= p(x,t)e^{\mathbf{X}(x,t)\beta_2} \\ &\vdots \\ q_n(x,t) &= p(x,t)e^{\mathbf{X}(x,t)\beta_n}. \end{aligned}$$

Using 2.1 we derive:

$$p(x,t) = 1 - \sum_{i=1}^{n} q_i(x,t)$$
$$p(x,t) = 1 - p(x,t) \sum_{i=1}^{n} e^{\mathbf{X}(x,t)\beta_i}$$
$$p(x,t) = \frac{1}{1 + \sum_{i=1}^{n} e^{\mathbf{X}(x,t)\beta_i}}. \tag{2.3}$$

The expression for cause-specific probability of death can be then obtained from 2.2:

$$q_i(x,t) = \frac{e^{\mathbf{X}(x,t)\beta_i}}{1 + \sum_{i=1}^{n} e^{\mathbf{X}(x,t)\beta_i}}, \quad i = 1, \ldots, n \tag{2.4}$$

It is also essential to show the cause-elimination (or alteration) impact. This adjustment on the underlying probabilities of death and survival will have a major impact on the life expectancy.

Let us introduce a shock factor $\rho_{i,x} \geq 0$ which is applied to cause i and age interval x, where shock values greater than one correspond to an increase in mortality rates. Setting a shock factor to zero then corresponds to cause-elimination. Thus, the underlying probabilities of death and survival (2.3 and 2.4) are adjusted as follows:

$$q_i(x,t) = \frac{\rho_{i,x} \cdot e^{\mathbf{X}(x,t)\beta_i}}{1 + \sum_{k=i}^{n} \rho_{i,x} \cdot e^{\mathbf{X}(x,t)\beta_i}} \tag{2.5}$$
$$p(x,t) = \frac{1}{1 + \sum_{i=1}^{n} \rho_{i,x} \cdot e^{\mathbf{X}(x,t)\beta_i}}. \tag{2.6}$$

It can also be assumed that a shock factor is the same for all age intervals, therefore, we can rewrite (2.5) and (2.6) as

$$q_i(x,t) = \frac{\rho_i \cdot e^{\mathbf{X}(x,t)\beta_i}}{1 + \sum_{k=i}^{n} \rho_i \cdot e^{\mathbf{X}(x,t)\beta_i}} \tag{2.7}$$
$$p(x,t) = \frac{1}{1 + \sum_{i=1}^{n} \rho_i \cdot e^{\mathbf{X}(x,t)\beta_i}}. \tag{2.8}$$

From a practical point of view, the direct application of the multinomial regression framework poses a serious problem in terms of computational efficiency. In order to address this problem in practical part, our approach will be to calculate the log-odds of cause-specific probabilities of death from the data, rather than estimating the latter using MLE.

In fact, if it had been the case, the direct application of MLR would have required some serious computational power of the underlying software as well as the hardware. Moreover, even the structure of the inputs would have been quite different and lacking compactness.

3. Application of MLR

3.1 Data

We obtained the data for Czech Republic from Czech Statistical Office for years 2003 to 2017 (2171 observations). The data contain cause-specific numbers of deaths along with central exposures by five years age intervals. There is however an exception for age groups from 0 to 1, from 1 to 4 and the final age group 95+ is open-ended. In Table 3.1 we provide a detailed overview of various causes of death according to the International Classification of Diseases (ICD) which are present in the data as well as categorization (third column) used in the regression model.

Table 3.1: Classification of Diseases according to ICD (1993).

ICD	Name CZ	Category
I	Některé infekční a parazitární nemoci (A00-B99)	Other
II	Novotvary (C00-D48)	Neoplasms
III	Nemoci krve, krvetvorných orgánů a některé poruchy týkající se mechanismu imunity (D50-D89)	Other
IV	Nemoci endokrinní, výživa přeměny látek (E00-E90)	Other
V	Poruchy duševní a poruchy chování (F00-F99)	Other
VI	Nemoci nervové soustavy (G00-G99)	Nervous system
VII	Nemoci oka a očních adnex (H00-H59)	Other
VIII	Nemoci ucha a bradavkového výběžku (H60-H95)	Other
IX	Nemoci oběhové soustavy (I00-I99)	Circulatory system
X	Nemoci dýchací soustavy (J00-J99)	Respiratory system
XI	Nemoci trávicí soustavy (K00-K93)	Digestive system
XII	Nemoci kůže a podkožního vaziva (L00-L99)	Other
XIII	Nemoci svalové a kosterní soustavy a pojivové tkáně (M00-M99)	Other
XIV	Nemoci močové a pohlavní soustavy (N00-N99)	Other
XV	Těhotenství, porod a šestinedělí (O00-O99)	Other
XVI	Některé stavy vzniklé v perinatálním období (P00-P96)	Other
XVII	Vrozené vady, deformace a chromosomální abnormality (Q00-Q99)	Other
XVIII	Příznaky, znaky a abnormální klinické a laboratorní nálezy nezařazené jinde (R00-R99)	Other
XX	Vnější příčiny poranění a otrav (V01-Y98)	External causes

We note that most of the causes of death were classified into category Other, since numbers of deaths by every particular cause are relatively small compared to the other groups. Nevertheless, aggregated numbers of deaths in category Other actually form a significant proportion of the data. The numerical notation in Table 3.2 will be used for causes of death categories. Circulatory system will be selected as a reference category for the purposes of the regression model.

Table 3.2: Coding of causes of death.

Category	Code
Circulatory system	0
Digestive system	1
External causes	2
Neoplasms	3
Nervous system	4
Other	5
Respiratory system	6

Age groups, as we mentioned earlier, are for the most part represented with 5 year intervals. Age group from 0 to 1 will enter the regression model as a reference category. In Table 3.3 we can see the corresponding coding for age groups.

Table 3.3: Coding of age groups.

Age groups	Code
0 to 1	0
1 to 4	1
5 to 9	2
10 to 14	3
15 to 19	4
20 to 24	5
25 to 29	6
30 to 34	7
35 to 39	8
40 to 44	9
45 to 49	10
50 to 54	11
55 to 59	12
60 to 64	13
65 to 69	14
70 to 74	15
75 to 79	16
80 to 84	17
85 to 89	18
90 to 94	19
95+	20

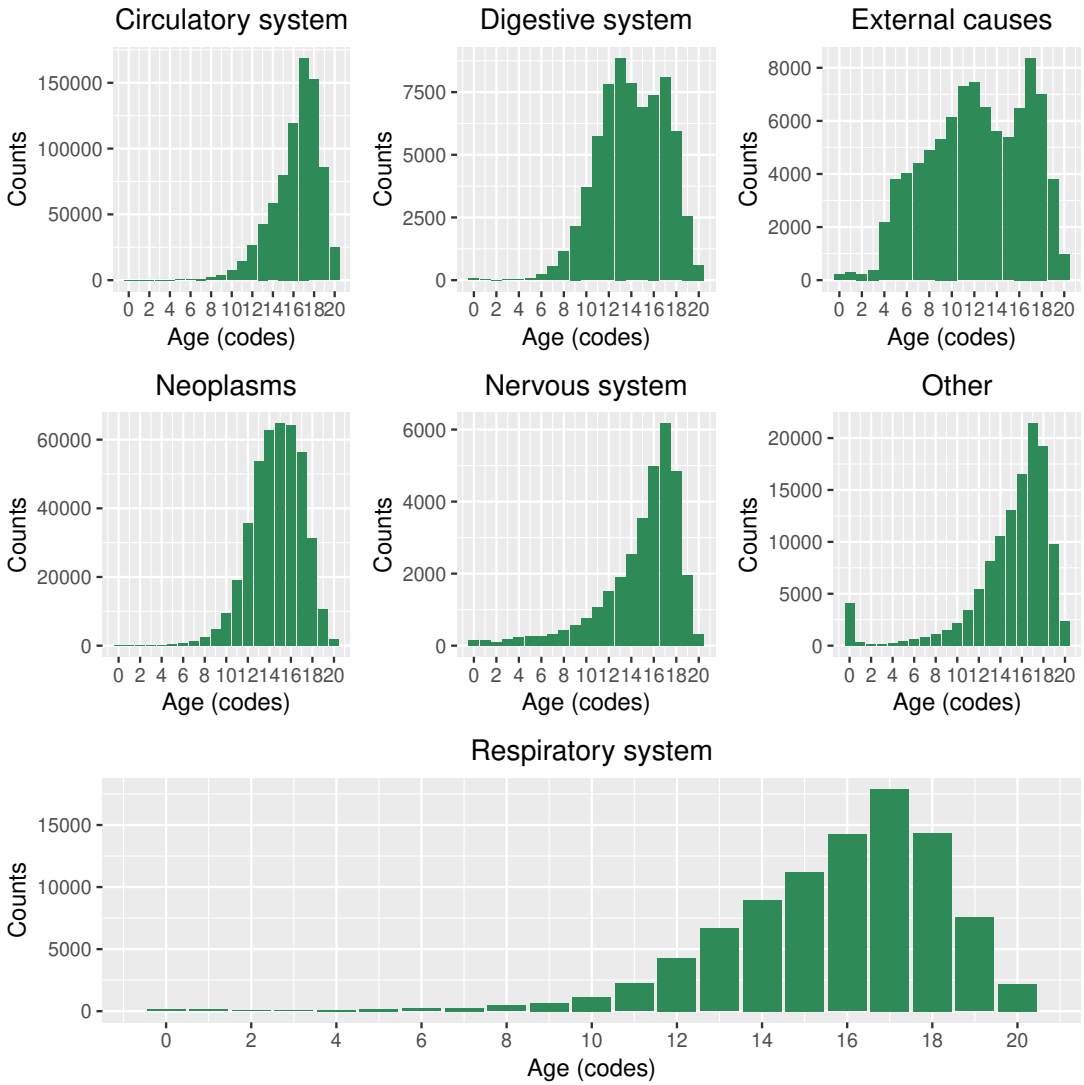


Figure 3.1: Histograms of the numbers of deaths.

Figure 3.1 presents histograms of the numbers of deaths for 15 years. Judging by the values on y-axis, we can conclude that circulatory system failure and neoplasms appear to be the most lethal and represent roughly 75% of the total number of deaths. While circulatory system failures tend to be more frequent in the last age cohorts, neoplasms mostly affect the population at retirement age. In general, deaths' distribution for circulatory, nervous, respiratory systems and other causes seems to be of the same form in terms of the negative skewness. Digestive system failures and deaths from external causes, for the most part, occur in middle age.

3.2 Regression model

We note that considering the nature of the data it makes sense to take into account the interaction between covariates age and time. The regression model can be then schematically written as follows:

$$\mathsf{logit}(mortality) = cause + t + age + t * age, \tag{3.1}$$

which can be rewritten more formally for k-th observation as

$$\begin{aligned}\mathsf{logit}(mortality_k) =& \beta_0 + \sum_{j=1}^{6} \beta_j \cdot \mathbb{1}[cause = j] + \beta_7 \cdot t \\ &+ \sum_{h=8}^{27} \beta_h \cdot \mathbb{1}[age = h - 7] + \sum_{m=28}^{47} \beta_m \cdot t \cdot \mathbb{1}[age = m - 27]\end{aligned}$$

for $k = 1, \ldots, 2171$. We recall that the expression $\mathsf{logit}(mortality)$ refers to 2.2, i.e. the logarithm of the cause-specific probability of death over the probability of survival, and $t = 1, \ldots, 15$ is a numerical time covariate. While fitting the model, we realized that many regression coefficients were insignificant, which leads to a conclusion that the initial model is overparameterized. Our aim now is to check whether the initial model can be reduced to a more simple one by conducting F-test on a submodel (see e.g. Fox [2016]). We shall consider removing the interaction term from the initial model and test whether the initial model given by 3.1 is significantly better than the smaller one (null hypothesis) or not (alternative hypothesis). Test statistic is given by

$$F = \frac{\frac{SS_e^0 - SS_e}{r - r_0}}{\frac{SS_e}{n - r}},$$

where SS_e^0 and SS_e denote the residual sums of squares corresponding to the smaller and to the initial model respectively, $r - r_0$ is the difference between the ranks of the corresponding model matrices and $n - r$ is equal to residual degrees of freedom of the initial model. Under the null hypothesis the test statistic F has $\mathcal{F}$ distribution with $r - r_0$ and $n - r$ degrees of freedom, hence we reject the null hypothesis on a significance level α, if $F \geq \mathcal{F}_{r-r_0,n-r}(1 - \alpha)$, i.e. for large values of the test statistic. Given the realized value f_0 of the test statistic, p-value is equal to

$$p = 1 - \mathsf{CDF}_{\mathcal{F},r-r_0,n-r}(f_0).$$

In our case $F = 0.5466$ and $p = 0.9475$, hence we do not reject the null hypothesis on a significance level of $\alpha = 0.5$ and, for further analysis, we shall stick with the smaller model without the interaction term. In other words, taking into account age-period interactions seems to be excessive at least in the case of the Czech Republic from 2003 to 2017.

We have also tried out several transformations of the time covariate in order to capture the behaviour of the response and figured out that the logarithmic transformation is probably the most appropriate in our case. Henceforth, our study will be based on the model

$$\mathsf{logit}(mortality) = cause + \mathsf{log}(t) + age,$$

where logit mortality rates will be calculated according to 1.9 and the corresponding estimate for the probability of survival. In Table 3.4 we show the output from `R` software which includes the estimates of the regression coefficients, standard errors, values of test statistic and p-values of individual t-tests. We can see that the majority of regression coefficients are statistically significant, i.e. most of the individual t-tests lead to rejecting the null hypothesis that the given coefficient

Table 3.4: Characteristics of the regression coefficients.

	Estimate	Std. Error	t-value	p-value
(Intercept)	-8.1445	0.1088	-74.84	**0.0000**
Digestive system	-1.4798	0.0709	-20.86	**0.0000**
External causes	-0.3450	0.0699	-4.94	**0.0000**
Neoplasms	-0.0351	0.0700	-0.50	0.6157
Nervous system	-1.7560	0.0699	-25.12	**0.0000**
Other	-0.6160	0.0699	-8.81	**0.0000**
Respiratory system	-1.2851	0.0699	-18.38	**0.0000**
log(t)	-0.1268	0.0246	-5.16	**0.0000**
1 to 4	-0.2669	0.1229	-2.17	0.0300
5 to 9	-0.5816	0.1239	-4.69	**0.0000**
10 to 14	-0.4435	0.1243	-3.57	0.0004
15 to 19	0.1241	0.1223	1.01	0.3104
20 to 24	0.4835	0.1212	3.99	0.0001
25 to 29	0.7410	0.1212	6.12	**0.0000**
30 to 34	1.1149	0.1212	9.20	**0.0000**
35 to 39	1.6046	0.1212	13.24	**0.0000**
40 to 44	2.1307	0.1212	17.59	**0.0000**
45 to 49	2.6564	0.1212	21.93	**0.0000**
50 to 54	3.1380	0.1212	25.90	**0.0000**
55 to 59	3.5712	0.1212	29.48	**0.0000**
60 to 64	3.9499	0.1212	32.60	**0.0000**
65 to 69	4.3366	0.1212	35.79	**0.0000**
70 to 74	4.7867	0.1212	39.51	**0.0000**
75 to 79	5.3735	0.1212	44.35	**0.0000**
80 to 84	6.1217	0.1212	50.53	**0.0000**
85 to 89	7.1227	0.1212	58.79	**0.0000**
90 to 94	9.0250	0.1212	74.49	**0.0000**
95+	9.1250	0.1212	75.32	**0.0000**

can be set to zero. In Table 3.5 we provide values of R^2 which appear to be quite high for our model.

We shall now discuss whether the assumptions of a normal linear model are satisfied. In order to do that, we shall investigate the diagnostic plots from Figure 3.2. In the first graph, we do not expect to see any clear trend. The LOWESS curve should be roughly $y = 0$. It means that the expected value of residuals is close to zero. However, we can see a slight quadratic trend, as well as few distant points. Nevertheless, the LOWESS curve is very close to zero and thus we consider the assumption of the conditional expectation of the residuals being equal to zero, i.e. $\mathsf{E}(\epsilon_i|\mathbf{X}_i = \mathbf{x}) = 0$, to be satisfied.

The second graph is normal QQ-plot, which compares quantiles of standardised residuals with theoretical ones. If green points form the line $y = x$, the residuals are normally distributed, i.e. $\epsilon_i|\mathbf{X}_i = \mathbf{x} \sim \mathcal{N}(0, \sigma^2)$, and the assumption is met. It is obviously not the case as tails' behaviour of standardized residuals' distribution is different.

Lastly we shall discuss the homoscedasticity (third graph), i.e. $\mathsf{var}(\epsilon_i|\mathbf{X}_i =$

Table 3.5: Coefficients of determination.

Multiple R^2	Adjusted R^2
0.9246	0.9237

$\mathbf{x}) = \sigma^2$ for some constant σ^2. In the perfect case, we again do not expect to see any patterns and the LOWESS curve to be close to $y = 1$. Here we can actually observe the same patterns as in the first graph and this time the quadratic trend seems to be even more apparent. Though, taking into account the scale on the y-axis, we conclude that the deviation from the curve $y = 1$ is rather minor, therefore we consider the homoscedasticity assumption to be satisfied.

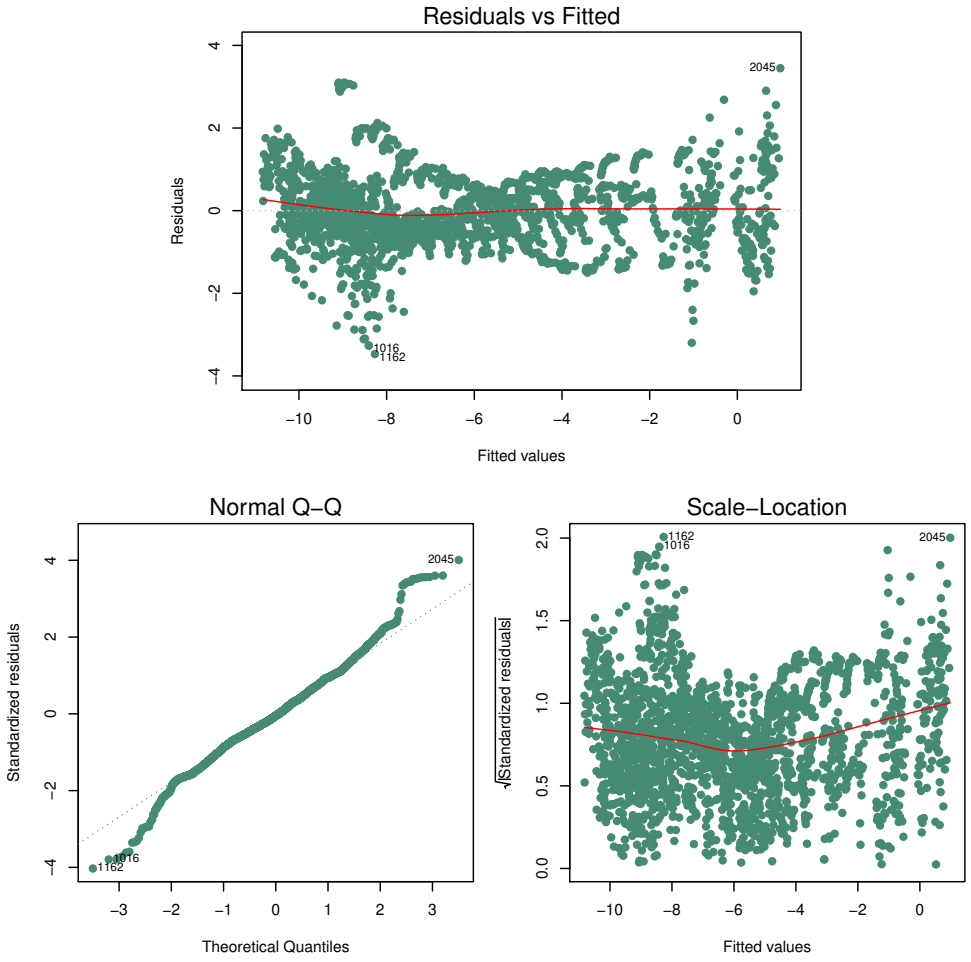


Figure 3.2: Diagnostic plots.

3.3 Outputs

In Figure 3.3 we compare the observed and fitted logit mortality rates across all age-groups in 2017. The red line corresponds to $y = x$. In general, we can say that fitted logit mortality rates are close the observed ones except for the external causes, which, however, represent only 5% of the deaths.

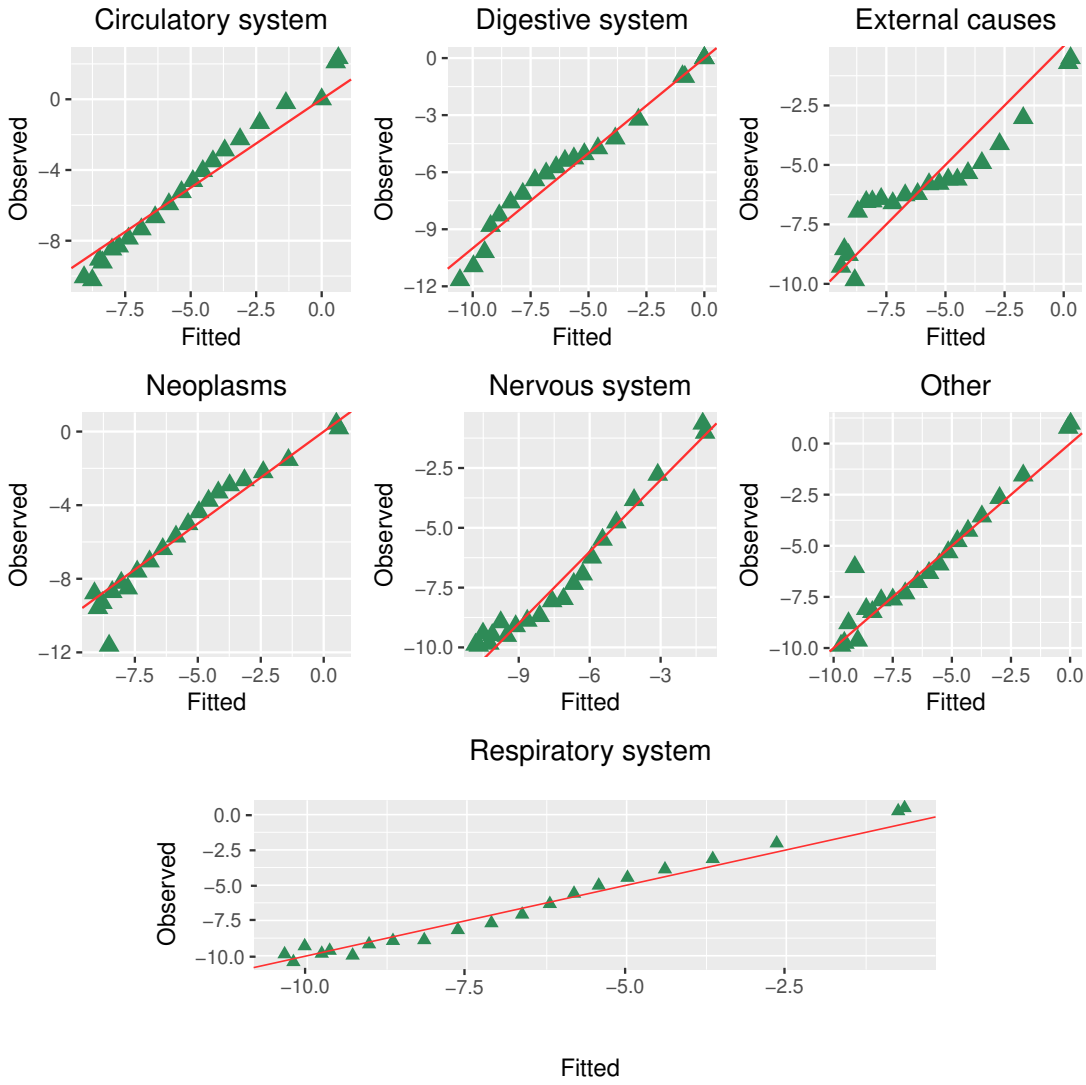


Figure 3.3: Observed vs fitted logit mortality rates in 2017.

Life expectancy will be calculated according to the methodology of Czech Statistical office. Given the probabilities of death in calendar year t in the interval (x_i, x_{i+1}), we have

$$l(x_{i+1}, t) = l(x_i, t)(1 - q(x_i, t)).$$

Number of deaths in the interval is given by

$$d(x_i, t) = l(x_{i+1}, t) - l(x_i, t).$$

Number of person-years lived in the interval is

$$L(x_i, t) = l(x_i, t) - (1 - a_i)d(x_i, t).$$

Number of person-years lived beyond the start of interval is

$$T(x_i, t) = \sum_i L(x_i, t).$$

Life expectancy at age x_i is then given by

$$e(x_i, t) = \frac{T(x_i, t)}{L(x_i, t)}.$$

Table 3.6 presents the observed and fitted life expectancy at birth and at the retirement age of 65. In both cases it is clear that the model does not provide quite an accurate fit, which might be explained by the lack of history available for the study. Last 15 years taken into account seem to be insufficient to fully capture the impact of time on the mortality rates. Nevertheless, the model still reflects the fact that life expectancy tends to increase over time, which makes sense generally and for the Czech Republic in particular. Moreover, it was empirically confirmed (based on the data) that the shorter is the history, the lower is the impact (significance) of time on the logit mortality rates.

Table 3.6: Life expectancy.

At birth			At retirement		
Year	Observed	Fitted	Year	Observed	Fitted
2017	78.81	79.91	2017	17.88	19.02
2016	78.83	79.88	2016	17.92	19.00
2015	78.44	79.85	2015	17.57	18.98
2014	78.61	79.81	2014	17.79	18.96
2013	78.06	79.77	2013	17.40	18.94
2012	77.88	79.73	2012	17.35	18.91
2011	77.69	79.68	2011	17.29	18.88
2010	77.45	79.63	2010	17.13	18.85
2009	77.18	79.57	2009	16.94	18.82
2008	77.09	79.50	2008	16.99	18.77
2007	76.82	79.42	2007	16.81	18.73
2006	76.62	79.32	2006	16.66	18.67
2005	76.07	79.19	2005	16.25	18.59
2004	75.85	79.00	2004	16.11	18.48
2003	75.31	78.69	2003	15.73	18.29

In Figure 3.4 we show the fitted mortality rates at age 40 with five-year outlook based on the considered regression model. It is clear that mortality rates tend to slowly diminish over time which might be explained by the overall progress in medicine along with improved quality of life.

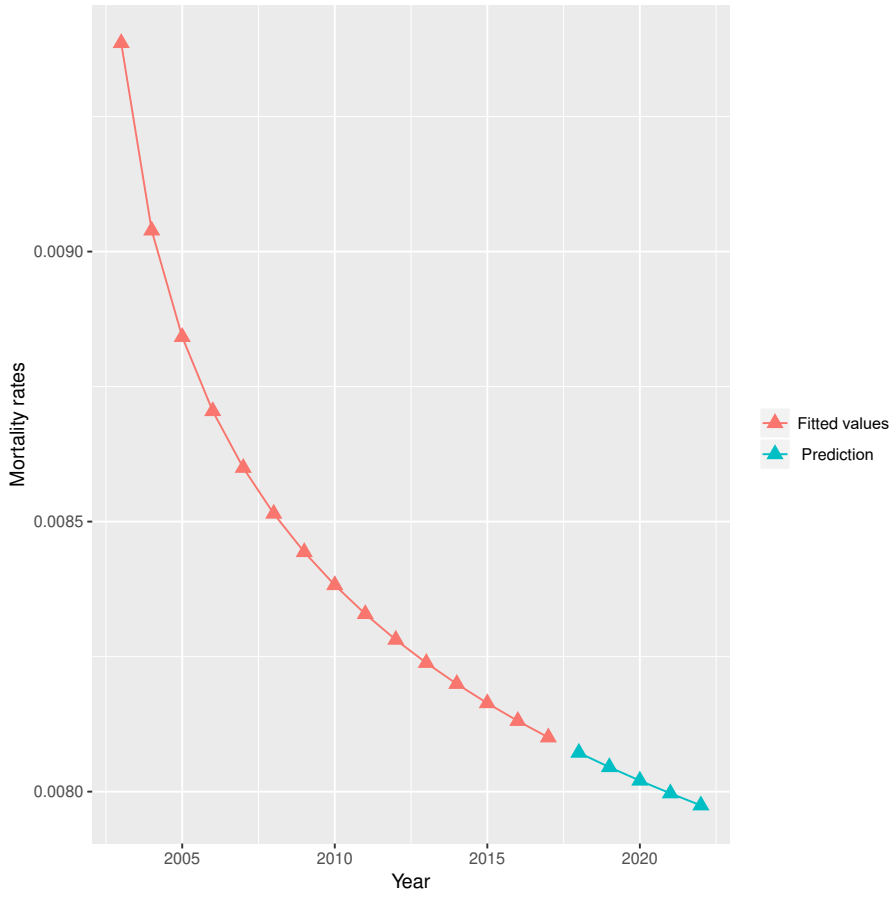


Figure 3.4: Fitted mortality rates with five-year outlook.

3.4 Stress scenarios

In this section we shall first demonstrate the impact of several scenarios on life expectancy by simulating life underwriting shocks assumed under the Solvency II regulatory regime. We shall focus on the following sub-modules: Mortality risk, Longevity risk and Life CAT risk. Our aim is to specify shock factors for each scenario based on the standard formula approach. We note that our main objective in this section is not to calculate (or estimate) the SCR (solvency capital requirement), however, the approach will most likely correspond, or will be at least similar to how these shocks are implemented in practise. Some assumptions will be made in order to address the problem of using age intervals. Scenario descriptions and the underlying assumptions will be fully based on EIOPA [2014], Technical Specification for the Preparatory Phase (Part I) published by European Insurance and Occupational Pensions Authority (EIOPA).

Later in this section we shall also consider several scenarios which might take place worldwide and in the Czech Republic in particular under some **theoretical** adverse circumstances. Here we emphasize that the latter scenarios will be considered due to their realism and complexity with no prior knowledge about the shock factors. Calibration methods are not the focus of this work. In the following scenarios we shall assume the independence of competing risks.

3.4.1 Life mortality risk

According to paragraph SCR.7.9 of Technical Specifications, Mortality risk is the risk of loss, or of adverse change in the value of insurance liabilities, resulting from changes in the level, trend, or volatility of mortality rates, where an increase in the mortality rate leads to an increase in the value of insurance liabilities.

The scenario definition, in general, is provided in paragraph SCR.7.11 and assumes that the SCR should be equal to the loss in basic own funds (BOF) of insurance and reinsurance undertakings that would result from an instantaneous[1] permanent increase in the mortality rates used for the calculation of technical provisions (BEL[2]+RM[3]).

In the calculation part of this sub-module it is assumed that mortality shock will result in instantaneous and permanent increase of mortality rates by 15%. Taking into account the methodology introduced earlier in this section, the underlying probabilities of death and survival should be adjusted as follows:

$$q_i(x,t) = \frac{1.15 \cdot e^{\mathbf{X}(x,t)\beta_i}}{1 + \sum_{i=1}^{n} 1.15 \cdot e^{\mathbf{X}(x,t)\beta_i}}$$
$$p(x,t) = \frac{1}{1 + \sum_{i=1}^{n} 1.15 \cdot e^{\mathbf{X}(x,t)\beta_i}}.$$

In the above expressions we used the shock factor $\rho_{i,x} = 1.15$, which is assumed to be applied uniformly for all age intervals and for all causes of death.

3.4.2 Life longevity risk

As outlined in paragraph SCR.7.20, Longevity risk is associated with the risk of loss, or of adverse change in the value of insurance liabilities, resulting from changes in the level, trend, or volatility of mortality rates, where a decrease in the mortality rate leads to an increase in the value of insurance liabilities.

The SCR should be then equal to the loss in BOF of insurance and reinsurance undertakings that would result from an instantaneous permanent decrease in the mortality rates used for the calculation of technical provisions (paragraph SCR.7.21).

Longevity scenario is applied by considering an instantaneous and permanent decrease of mortality rates by 20%. As a result of this change, probabilities of death and survival transform into

$$q_i(x,t) = \frac{0.8 \cdot e^{\mathbf{X}(x,t)\beta_i}}{1 + \sum_{i=1}^{n} 0.8 \cdot e^{\mathbf{X}(x,t)\beta_i}}$$
$$p(x,t) = \frac{1}{1 + \sum_{i=1}^{n} 0.8 \cdot e^{\mathbf{X}(x,t)\beta_i}}.$$

The shock factor is then equal to $\rho_{i,x} = 0.8$ for all age groups and for all causes of death.

[1] Applied at the projection start date of insurer's liabilities
[2] Best estimate of liabilities
[3] Risk margin

3.4.3 Life CAT risk

Paragraph SCR.7.75 states that Catastrophe risk stems from extreme or irregular events whose effects are not sufficiently captured in the other life underwriting risk sub-modules. Examples could be a **pandemic event** or a **nuclear explosion**.

Life CAT risk is assumed to result in an instantaneous increase in mortality rates by 0.15 percentage points in the following 12 months. Here we recall that in our data age is a categorical variable, hence it is not possible to fully reflect the shock duration for any of the age groups except for the first category from 0 to 1. For the purposes of this study, we shall assume that Life CAT scenario affects exactly one age interval regardless of its length. That being said, in a hypothetical liabilities projection model, given an individual at age x and in age group $[x,\, x+5]$, cause-specific mortality rates will be adjusted for this particular interval, for the next age group a model will read (generate) central scenario mortality rates.

Technical specification provides two examples of possible CAT scenarios, and in this work we are going to simulate both of them separately. Apparently, it does not quite make sense to neither assume both pandemic and nuclear explosion to happen at the same time nor to treat this situation as one CAT event, at least not from Solvency II perspective as it would be highly improbable. Hypothetically speaking, an insurance company could calculate the SCR by taking the one, which leads to a greater loss in BOF, i.e. the most adverse one. Nevertheless, such approach seems to be somewhat beyond the scope of the standard formula.

We assume that **pandemic scenario** will result in a large number of claims due to circulatory system failure, e.g. caused by Ebola hemorrhagic fever. Therefore, such event leads to an increase in mortality rates on a single cause of death. We further assume that **nuclear explosion scenario** will lead to mass external causes claims by the devastating impact of the initial blast along with neoplasms (cancer) claims by radioactive contamination. In both scenarios the shock factor is calculated as

$$\rho_{i,x} = \frac{q_i(x,t) + 0.0015}{q_i(x,t)}.$$

From the above expression it is also clear that the shock factor is greater for younger age groups. Thus, Life CAT exposure of an insurance company, whose portfolio consists of younger clients, is higher.

3.4.4 Global climate change

Nowadays global climate change, in particular global warming, is a topic widely discussed. Shifted weather patterns, changes in the global sea level, overall temperature increase and other potentially dangerous environmental changes may sooner or later lead to various adverse events. For the purposes of this work, we shall focus on the global scenario that is assumed to be of a permanent duration and which will result in an increase of the number of disease vectors[4].

Firstly, we consider an increase in the population of insect vectors of human pathogens, namely the genus Anopheles of mosquito. Many species of this genus are widely known for transmitting human malaria which causes circulatory system

[4]An organism who carries and transmits a pathogen into another organism

failure. Malaria is widely spread in the tropical and subtropical regions, however, due to adverse climate change, the disease is assumed to spread to northern areas as well.

Secondly, an increase of vectors who carry the fungus Histoplasma capsulatum is considered. This fungus transmitted by bats, is known for causing histoplasmosis characterized by interstitial pneumonia which affects respiratory system.

The global scenario is then assumed to result in a permanent increase in mortality rates on circulatory and respiratory systems by 60% and 75%, respectively. As a result of this change, the following adjustments of probabilities will be considered for all age groups:

$$q_i(x,t) = \frac{1.6 \cdot e^{\mathbf{X}(x,t)\beta_i}}{1 + \sum_{i=1}^{n} 1.6 \cdot e^{\mathbf{X}(x,t)\beta_i}} \qquad q_i(x,t) = \frac{1.75 \cdot e^{\mathbf{X}(x,t)\beta_i}}{1 + \sum_{i=1}^{n} 1.75 \cdot e^{\mathbf{X}(x,t)\beta_i}}$$

$$p(x,t) = \frac{1}{1 + \sum_{i=1}^{n} 1.6 \cdot e^{\mathbf{X}(x,t)\beta_i}} \qquad p(x,t) = \frac{1}{1 + \sum_{i=1}^{n} 1.75 \cdot e^{\mathbf{X}(x,t)\beta_i}}.$$

3.4.5 Drug resistance

In recent years drug resistance has become a major concern in medicine. In particular, a misuse and overuse of antibiotics is nowadays considered as an increasing problem not only in human but also in veterinary medicine. As pointed out in Adámková [2015], the antibiotic therapy has to be carefully assessed and should be based on the knowledge of local epidemiology.

We consider an appearance of multi-drug resistant strain of bacteria that will result in increased mortality rates on several causes of death. In order to illustrate the adverse impact of the considered scenario, we assume the following changes: increase of mortality rates on respiratory system and other causes by 80%, circulatory and nervous systems by 50%, digestive system by 40%. We note that due to the nature of this scenario, neoplasms and external causes are out of scope.

In the worst case, the drug resistance scenario might be considered of a permanent duration, nevertheless, it is essential to take into account that the medical society will most likely implement certain strategies to deal with the problem. Thus, similarly to Life CAT scenario, we shall consider the duration equal to 30 years, i.e. roughly 7 age intervals.

3.4.6 Impacts on the life expectancy

The impact of stress scenarios will be illustrated by means of the life expectancy at age 40 in 2017, since the population is the most dense at this age. Also, it is probably safe to assume that the latter is at least close the average age in a hypothetical portfolio of an insurance company. The outputs from stress scenarios will be compared with the observed life expectancy. Henceforth, the scenario, when no shock factors are considered, will be referred to as **central scenario**.

In Table 3.7 we show the impacts of Solvency II scenarios on the life expectancy. Mortality risk appears to have the most adverse impact on the life expectancy, on the other hand, impacts of these scenarios ($\triangle$BOF) really depend on the structure of the underlying portfolio.

Table 3.7: Shocked vs central life expectancies (SII scenarios).

Central	Mortality risk	Longevity risk	CAT pandemic	CAT explosion
40.74	39.70	42.31	40.69	40.63

Table 3.8 presents the impacts of global climate change and drug resistance scenarios. It appears that drug resistance case is more adverse, hence the exposure to multiple risks might be potentially more dangerous, even though the limited duration was considered.

Table 3.8: Shocked vs central life expectancies (Other scenarios).

Central	Climate change	Drug resistance
40.74	39.31	38.98

4. Application of copula functions

In this chapter we are going to focus on building the dependence model between two competing risks using copula functions. In previous chapter we used the data from Czech Statistical Office which contained various causes of death. For the purposes of this chapter we shall take the data from the year 2017 and consider regrouping all causes of deaths into two categories: circulatory system failures and other causes.

4.1 Evaluating crude survival functions

As we mentioned earlier in Chapter 1, the system of differential equations 1.14 requires crude survival functions to be in a functional form. To begin with, we shall use the approach presented in Kaishev et al. [2007] to calculate crude survival functions from the data.

Let $l_0 = D^{(c)} + D^{(o)}$, where $D^{(c)}$ denotes the total number of deaths from circulatory system failures for all age intervals and $D^{(o)}$ is the complementing number of deaths from all other causes. Probability that a newborn will die from a circulatory system failure is equal to

$$ {}_{\infty}q^{(c)} = \frac{D^{(c)}}{l_0}. $$

The respective probability for other causes is equal to

$$ {}_{\infty}q^{(o)} = \frac{D^{(o)}}{l_0}. $$

Crude survival functions are then evaluated as follows

$$ \begin{aligned} S^{(c)}(k) &= {}_{\infty}q^{(c)} - \sum_{x<k} D_x^{(c)} \,/\, l_0 \\ S^{(o)}(k) &= {}_{\infty}q^{(o)} - \sum_{x<k} D_x^{(o)} \,/\, l_0, \end{aligned} $$

for $k = 1, \ldots, 110$. In the data described in Chapter 3 we had 21 age intervals in total with 95+ being the last one. In order to calculate crude survival functions for higher ages, we used the data (columns Dx) from males and females life tables from the year 2017 (ČSÚ [2017]) which contain deaths from all causes. Numbers of deaths (e.g. for other causes) for age intervals 95-99, 100-104 and 105-109 are calculated as follows:

$$ \begin{aligned} D^{(o)}_{95\text{-}99} &= D^{(o)}_{95+} \times \frac{D_{95\text{-}99}}{D_{95\text{-}99} + D_{100\text{-}104} + D_{105\text{-}109}} \\ D^{(o)}_{100\text{-}104} &= D^{(o)}_{95+} \times \frac{D_{100\text{-}104}}{D_{95\text{-}99} + D_{100\text{-}104} + D_{105\text{-}109}} \\ D^{(o)}_{105\text{-}109} &= D^{(o)}_{95+} \times \frac{D_{105\text{-}109}}{D_{95\text{-}99} + D_{100\text{-}104} + D_{105\text{-}109}}, \end{aligned} $$

where $D_{95\text{-}99}$, $D_{100\text{-}104}$ and $D_{105\text{-}109}$ are taken from the mentioned life tables. We assume that by the age of 110 there is no one alive in the population.

Kernel Regression

Now when crude survival functions are calculated from the data, we need to find an appropriate way to represent them in a functional form. The smoothing technique we shall focus on is the kernel regression covered e.g. in Campbell et al. [1997].

Suppose that we are interested in modelling the relation between two random variables Y_t and X_t which satisfy

$$Y_t = m(X_t) + \epsilon_t, \quad t = 1, \ldots, T,$$

where $m(\cdot)$ is a nonlinear unknown function and ϵ_t are i.i.d random variables with zero mean. We are looking for an estimator of $m(x)$ in a form

$$\widehat{m}(x) = \frac{1}{T} \sum_{t=1}^{T} \omega_{t,T}(x) Y_t, \tag{4.1}$$

where $\omega_{t,T}(x)$ plays the role of a weighting function. The kernel regression framework considers the construction of $\omega_{t,T}(x)$ from a symmetric probability density function $K : \mathbb{R} \to \mathbb{R}$ which is called a kernel:

$$K(x) \geq 0, \qquad \int_{\mathbb{R}} K(x)\, dx = 1.$$

By rescaling the kernel for some $h > 0$ (so-called bandwidth), we get

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right), \qquad \int_{\mathbb{R}} K_h(x)\, dx = 1.$$

The weighting function $\omega_{t,T}(x)$ can be then defined as

$$\omega_{t,T}(x) = \frac{K_h(x - X_t)}{\frac{1}{T}\sum_{t=1}^{T} K_h(x - X_t)}. \tag{4.2}$$

Substituting 4.2 into 4.1 leads to the Nadaraya-Watson estimator of $m(x)$:

$$\widehat{m}(x) = \frac{1}{T} \sum_{t=1}^{T} \omega_{t,T}(x) Y_t = \frac{\sum_{t=1}^{T} K_h(x - X_t) Y_t}{\sum_{t=1}^{T} K_h(x - X_t)}.$$

We shall use the Gaussian kernel to obtain the estimators for crude survival functions:

$$K_h(x) = \frac{1}{h\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2h^2}\right\}.$$

The choice of the bandwidth h is essential in the kernel regression, however, from the practical point of view it is reasonable to assess the optimality of h graphically, which we did in `Wolfram Mathematica` software.

One of the data-driven techniques to select an optimal bandwidth is the cross-validation method. Let

$$\widehat{m}_{h,j}(X_j) = \frac{1}{T} \sum_{t \neq j} \omega_{t,T}(X_j) Y_t,$$

which is the so-called leave-one-out kernel estimator based on the observations $(X_1, Y_1), \ldots, (X_{j-1}, Y_{j-1}), \ldots, (X_{j+1}, Y_{j+1}), \ldots, (X_T, Y_T)$. The cross-validation function is given by

$$\mathsf{CV}(h) = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \widehat{m}_{h,t}(X_t))^2 w(X_t),$$

where $w(X_t)$ is a non-negative weight function used to reduce the variability of $\mathsf{CV}(h)$. Finally, the optimal bandwidth is such h that minimizes the cross-validation function, i.e.

$$\widehat{h}_{CV} := \arg\min_{h>0} \mathsf{CV}(h).$$

In Figure 4.1 we show interpolated crude survival functions for circulatory system failures (Scrude1) and other causes (Scrude2).

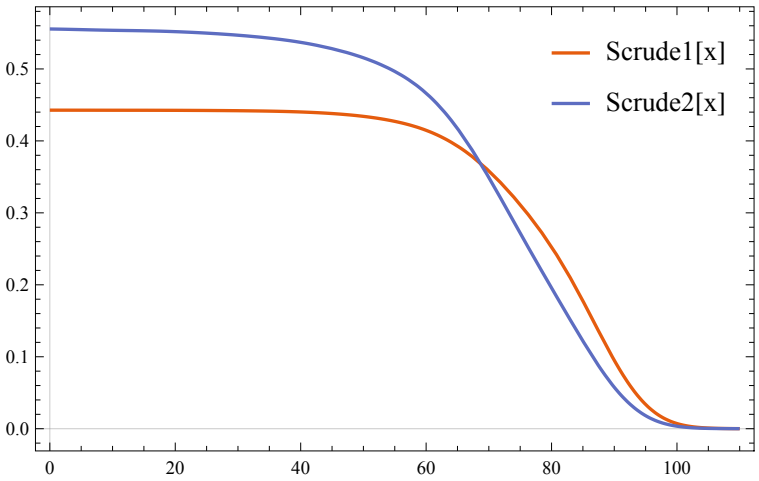


Figure 4.1: Interpolated crude survival functions.

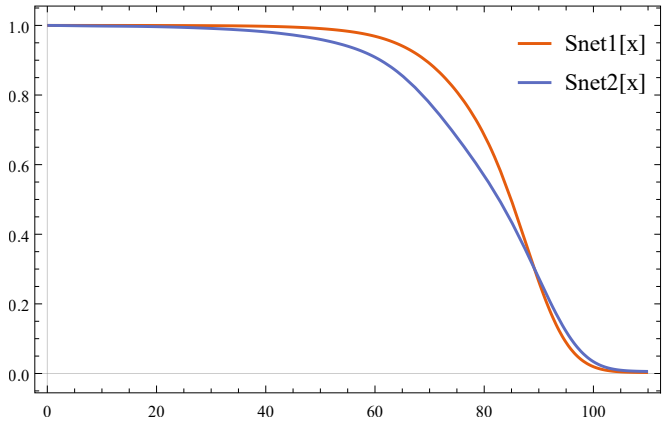
4.2 Outputs

In Chapter 1 we have already pointed out that potential lifetimes of an individual are unobservable in practise, however, it is reasonable to assume some degree of association between them in order to incorporate the dependence structure using copulas.

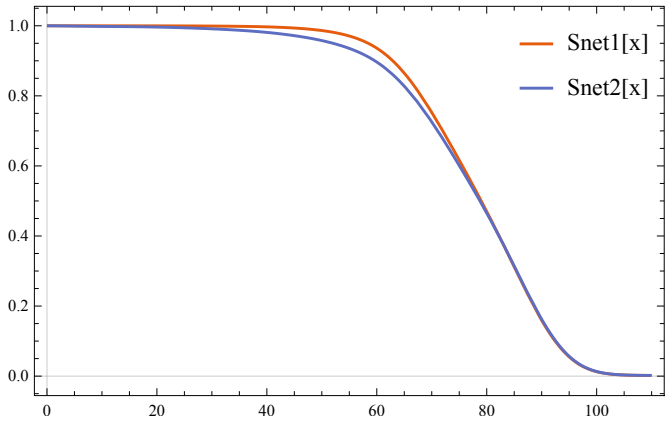
Let X_1 be a potential lifetime of an individual if he would die from circulatory system failure and X_2 be the corresponding lifetime if the death would occur due to other causes. We shall solve the system of differential equations 1.14 assuming $\tau(X_1, X_2) = 0.33$ (weak positive correlation) and $\tau(X_1, X_2) = 0.9$ (strong positive correlation) for the Clayton copula and $\tau(X_1, X_2) = -0.1817$ (weak negative correlation) for the AMH copula.

In Figure 4.2 we show net survival functions for the two mentioned values of Kendall's tau in the case of Clayton copula and in Figure 4.3 we provide the results in the case of AMH copula for which we consider negative dependence. While the results for Clayton copula seem to be generally in line with the assumption that almost no one is alive by the age of 110, the outputs for AMH copula suggest that there is still a considerable proportion of the population alive even at higher ages.

In fact, comparing all three plots, it seems to be the case that survival curves tend to be "higher" for lower correlations no matter which copula function was used. Due to technical issues, we were not able to apply more flexible Gaussian or t-copulas and thus we focused on Archimedean copulas.



(a) Net survival functions for $\tau(X_1, X_2) = 0.33$.



(b) Net survival functions for $\tau(X_1, X_2) = 0.9$.

Figure 4.2: Net survival functions in the case of Clayton copula.

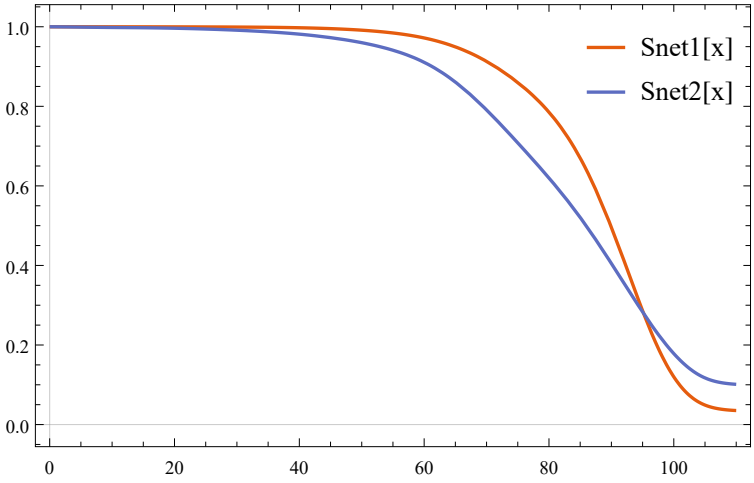
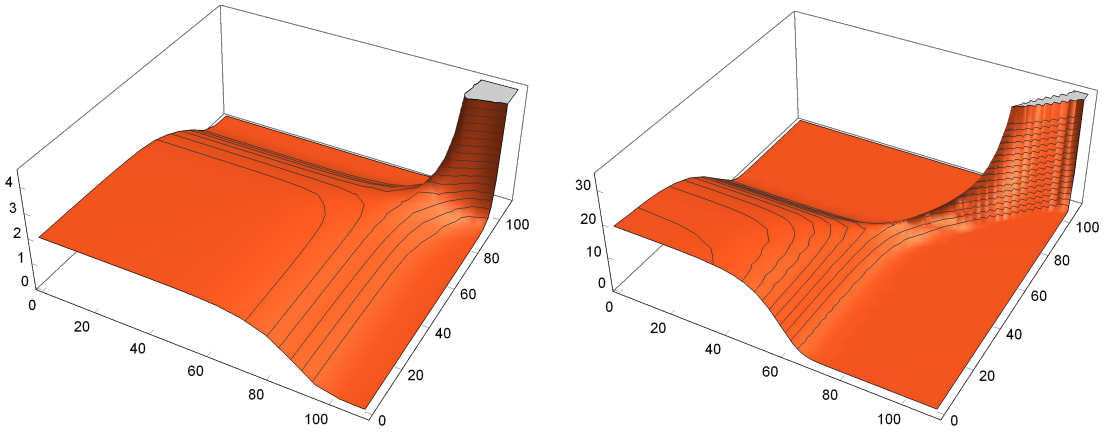


Figure 4.3: Net survival functions in the case of AMH copula.

In Figure 4.4 we also show the joint probability density functions of X_1 and X_2 in the case of Clayton copula and in Figure 4.5 we show the corresponding joint density in the case of AMH copula.



(a) Joint desity for $\tau(X_1, X_2) = 0.33$.

(b) Joint desity for $\tau(X_1, X_2) = 0.9$.

Figure 4.4: Joint densities in the case of Clayton copula.

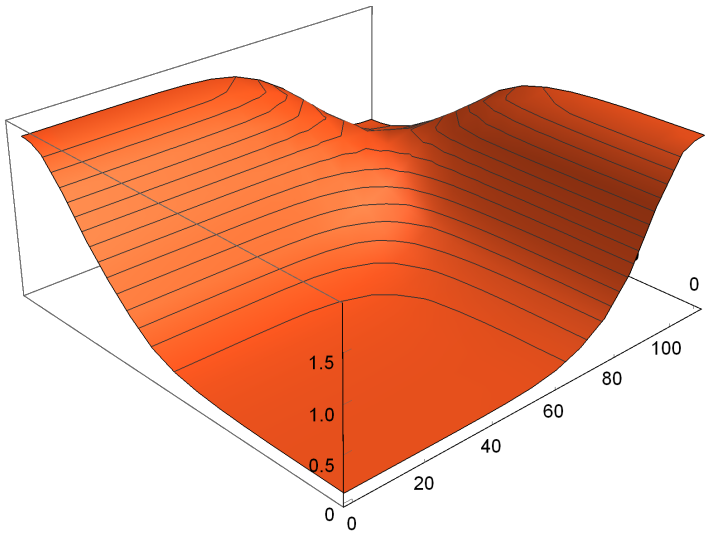


Figure 4.5: Joint density in the case of AMH copula.

Finally, we shall demonstrate the impact of eliminating circulatory system failures by means of the life expectancy at birth and at retirement age of 65. We note that this scenario is rather unrealistic, since it is highly improbable that in the near future such a frequent cause of death will be completely dealt with. The same scenario will be applied using the outputs from the regression model in Chapter 3. We recall that stress scenarios from Chapter 3 were calculated under the independence assumption of competing risks.

In Table 4.1 we compare life expectancies calculated for the central scenario (i.e. the observed one), MLR, Clayton and AMH copulas. Life expectancies for central scenario correspond to the ones presented earlier in Table 3.6. It appears that when assuming high correlation between causes of death, the life expectancy at birth as well as at retirement age even decreased, which probably does not

make much sense. We rather get more intuitive results for lower correlations and in the case of independence.

Table 4.1: Comparison of life expectancies.

	Central	MLR	Clayton copula		AMH copula
	-	-	$\tau = 0.9$	$\tau = 0.33$	$\tau = -0.1817$
at birth	78.81	82.14	77.89	80.92	84.87
at retirement	17.88	20.39	14.67	17.55	21.48

Conclusion

The aim of this work was to present different approaches to cause-of-death mortality analysis and to demonstrate the application of the selected method on real data.

In Chapter 1 we introduced the continuous model based on the force of mortality and presented the estimation method with respect to current population data. We also provided a brief overview of the method based on copula functions, which models the dependence between causes of death.

In Chapter 2 we presented the multinomial logistic regression formulated for cause-of-death mortality problem. We further discussed the construction of life tables given the central exposure to risk and age-cause-specific numbers of deaths.

In Chapter 3 we focused on the application of multinomial logistic regression on data from Czech Statistical Office and used the available 15 years history in our study. We first identified the appropriate regression model and discussed whether the assumptions of normal linear model were satisfied. Next we presented the outputs from the model including fitted life expectancies and predicted mortality rates. Later in this chapter, we considered several stress scenarios in order to demonstrate the impacts of shocked mortality rates on life expectancy. We first focused on the life underwriting shocks, namely mortality risk, longevity risk and Life CAT risk, assumed under Solvency II regulatory framework. Secondly, we considered two hypothetical stress scenarios, namely global climate change and drug resistance, which also simulate the adverse evolution of mortality rates. The latter scenarios might be useful for the purposes of so-called Own Risk and Solvency Assessment (ORSA) process within the second pillar of Solvency II when insurance companies are required to assess their own risk profile.

In Chapter 4 we focused on the application of copula functions in order to incorporate the dependence structure between the competing risks. We evaluated crude survival functions and solved the system of differential equations for unknown net survival functions. Lastly, we considered cause-elimination scenario for circulatory system failures and compared the outputs with calculations based on the multinomial logistic regression model from Chapter 3. We also confirmed that assuming high correlation between the two studied causes of death leads to slightly contradictory results.

Bibliography

V. Adámková. Antibiotická léčba. *Medicína pro praxi*, 2015.

Daniel H. Alai, Séverine Arnold (-Gaille), and Michael Sherris. Modelling cause-of-death mortality and the impact of cause-elimination. *Annals of Actuarial Science*, pages 167–186, 2015.

John Y. Campbell, Andrew W. Low, and Craig A. MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, 1997. ISBN 9780691043012.

J. Carriere. Dependent decrement theory. *Transactions of Society of Actuaries*, 46, 1994.

C.L. Chiang. *Introduction to Stochastic Processes in Biostatistics*. John Wiley and Sons, New York, 1968.

EIOPA. Technical specification for the preparatory phase. `https://eiopa.europa.eu/Publications/Standards/A_-_Technical_Specification_for_the_Preparatory_Phase__Part_I_.pdf`, 2014.

J. Fox. *Applied Regression Analysis and Generalized Linear Models*. 3rd edition. SAGE Publications, Inc, 2016. ISBN 978-1-4522-0566-3.

H.U. Gerber. *Life Insurance Mathematics*. Springer, Berlin, Germany, 1997. ISBN 978-3-662-03460-6.

V.K. Kaishev, D.S. Dimitrova, and S. Haberman. Modelling the joint distribution of competing risks survival times using copula functions. *Insurance: Mathematics and Economics*, pages 339–361, 2007.

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, 2005. ISBN 0-691-12255-5.

K. Pranesh. Probability distributions and estimation of ali-mikhail-haq copula. *Applied Mathematical Sciences, Vol. 4*, pages 657–666, 2010.

ČSÚ. Úmrtnostní tabulky za Čr. `https://www.czso.cz/csu/czso/umrtnostni-tabulky-za-cr-regiony-soudrznosti-a-kraje-2016-2017`, 2017.

List of Figures

List of Tables