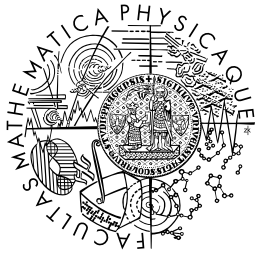

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

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 2019

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 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 review the approach using copula functions. In Chapter 2 we focus on the multinomial logit model formulated for cause-specific mortality data,discuss life tables construction and derive life expectancy. 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.

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. Next, we shall discuss the life tables construction and derive several necessary statistical quantities.

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.

1. Continuous model

1.1 Competing risks in mortality analysis

In this section we shall focus on the concept of competing risks in mortality analysis. We will present the methodology introduced in Chiang [1968] along with the estimation method based on the data about current population.

Let us assume a group of lives where every individual may die from one of n competing risks (causes of death). Let $X_1, \ldots, X_n$ 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 potential lifetime of an individual if death occurs from i-th cause corresponds to 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 1.3.

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

- The crude probability ... the probability of death from a specific cause in the presence of all other competing risks ($Q_{i\delta}$);
- The net probability ... the probability of death if a specific risk is the only one risk in effect in the population ($q_{i\delta}$) or in absence of all other competing risks ($q_{i,\delta}$);
- The partial crude probability ... the probability of death from a specific cause R_δ when another risk R_1 ($Q_{i\delta.1}$) or risks, e.g. R_1 and R_2 ($Q_{i\delta.12}$) are eliminated as a risk of death from the population.

Probabilities of death and survival in the interval (x_i, x_{i+1}) of an individual at age x_i are denoted as q_i and p_i, respectively, with $q_i + p_i = 1$.

In the presence of r causes of death, $R_1, \ldots, R_r$, for each risk R_δ there is a corresponding force of mortality $\mu(x;\delta)$ (cause-specific), which expresses the probability that an individual alive at age x will die from cause R_δ in the infinitesimal time interval $(x, x+dx)$ for $\delta = 1, \ldots, r$. The **total** force of mortality $\mu(x)$ then corresponds 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:

$$\mu(x) = \sum_{\delta=1}^{r} \mu(x;\delta).$$

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

$$q_i = 1 - \mathsf{exp}\left\{ -\int_{x_i}^{x_{i+1}} \mu(x)\, dx \right\},$$

$$p_i = \mathsf{exp}\left\{ -\int_{x_i}^{x_{i+1}} \mu(x)\, dx \right\}.$$

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

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

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

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

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

The crude probability of death discussed earlier is given by

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

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

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

And thus

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

1.2 Current population mortality analysis

As outlined in Chiang [1968], current population mortality analysis is a method used to estimate the probabilities described earlier in this chapter. Let in a given calendar year (x_i, x_{i+1}) be the age interval, $n_i = x_{i+1} - x_i$ the length of the interval, P_i the midyear population state (also referred to as central exposure), D_i the total number of deaths , a_i is the average fraction of the age interval that each of the individuals survive before dying and N_i the unobserved population state at x_i (initial exposure). Age specific mortality rate is

$$M_i = \frac{D_i}{P_i}. \tag{1.4}$$

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

$$\widehat{q_i} = \frac{D_i}{N_i}, \tag{1.5}$$

where the initial exposure N_i can be estimated from

$$\widehat{N_i} = (P_i + (1 - a_i)n_i D_i)/n_i.$$

Thus, (1.5) transforms into

$$\widehat{q_i} = \frac{n_i M_i}{1 + (1 - a_i)n_i M_i}. \tag{1.6}$$

The corresponding survival probability is therefore

$$\widehat{p_i} = \frac{1 - a_i n_i M_i}{1 + (1 - a_i)n_i M_i}. \tag{1.7}$$

Switching 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:

$$D_i = \sum_{\delta=1}^{r} D_{i\delta}.$$

The cause-specific mortality rate is then given by

$$M_{i\delta} = \frac{D_{i\delta}}{P_i}.$$

Similarly, the crude probability of death is estimated from

$$\widehat{Q}_{i\delta} = \frac{D_{i\delta}}{N_i},$$

which can be also expressed as

$$\widehat{Q}_{i\delta} = \frac{n_i M_{i\delta}}{1 + (1 - a_i)n_i M_i}. \tag{1.8}$$

Solving the equations (1.6) and (1.8) with respect to death rates M_i and $M_{i\delta}$, we get

$$M_i = \frac{\widehat{q}_i}{\widehat{q}_i a_i n_i + (1 - \widehat{q}_i) n_i},$$
$$M_{i\delta} = \frac{\widehat{Q}_{i\delta}}{\widehat{q}_i a_i n_i + (1 - \widehat{q}_i) n_i},$$

which implies that

$$\frac{M_{i\delta}}{M_i} = \frac{\widehat{Q}_{i\delta}}{\widehat{q}_i}.$$

Thus, the expression above is an analogy of the formula 1.3 which tells that the ratio of two mortality rates is equal to the ratio of the respective forces of mortality.

1.3 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. For the purposes of this work, 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 $\mathbf{X} = (X_1, \dots X_n)^\top$ assigned to an individual with respect to causes of death $i = 1, \dots, n$. In practise, however, we observe only actual lifetime, which is equal to min $(X_1, \dots X_n)$. The joint distribution function of $\mathbf{X}$ is given by

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

and the joint survival function is

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

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

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

and it obviously holds that

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

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

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

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

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

In analytic terms, 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*

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

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* $x_1, \ldots x_d \in [-\infty, +\infty]$,

$$F(x_1, \ldots x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \tag{1.9}$$

If the marginal distributions are continuous, then C *is unique. Otherwise,* C *is uniquely determined in* $\mathsf{Ran}(F_1), \ldots, \mathsf{Ran}(F_d)$, *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 (1.9) is a joint distribution function with marginals* $F_1, \ldots, F_d$.

Thus, 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 since marginal survival functions are of the form $S_i : \mathbb{R}_+ \to [0,1]$, the joint survival function of X_i is uniquely determined by

$$S(x_1, \ldots x_n) = \bar{C}(S'^{(1)}(x_1), \ldots, S'^{(n)}(x_n)), \tag{1.10}$$

where $\bar{C}$ is the survival copula with respected to copula C. 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 1.10 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*

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

where

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

Using the above lemma and applying the chain rule for 1.10, 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*

$$\begin{aligned}
\frac{d}{dx}S^{(1)}(x) &= \bar{C}_1(S'^{(1)}(x),\ldots,S'^{(n)}(x)) \times \frac{d}{dx}S'^{(1)}(x) \\
\frac{d}{dx}S^{(2)}(x) &= \bar{C}_2(S'^{(1)}(x),\ldots,S'^{(n)}(x)) \times \frac{d}{dx}S'^{(2)}(x) \\
&\vdots \\
\frac{d}{dx}S^{(n)}(x) &= \bar{C}_n(S'^{(1)}(x),\ldots,S'^{(n)}(x)) \times \frac{d}{dx}S'^{(n)}(x),
\end{aligned} \tag{1.11}$$

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 1.11 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 1.11 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 1.10.

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 $x \equiv k = 1,2,\ldots,120$. 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

2.1 Multinomial logistic regression

In this section we are going to focus on methodology by introducing the multinomial logistic model. 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, we shall stick with the notation introduced in Alai et al. [2015]:

- $D_i(x,t)$ ... 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

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

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

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

where $d_k(x,t)$ 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:

$$\ln \frac{q_k(x,t)}{p(x,t)} = \mathbf{X}_k \beta_k, \quad k = 1, \ldots, n \tag{2.2}$$

where $\mathbf{X}$ is model matrix and β_k 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. To obtain predicted probabilities we exponentiate and rewrite (2.2) in terms of the sequence of binary models:

$$\begin{aligned} q_1(x,t) &= p(x,t) e^{\mathbf{X}_1 \beta_1} \\ q_2(x,t) &= p(x,t) e^{\mathbf{X}_2 \beta_2} \\ &\vdots \\ q_n(x,t) &= p(x,t) e^{\mathbf{X}_n \beta_n}. \end{aligned}$$

Using 2.1 we derive:

$$p(x,t) = 1 - \sum_{k=1}^{n} q_k(x,t)$$
$$p(x,t) = 1 - p(x,t) \sum_{k=1}^{n} e^{\mathbf{X}_k \beta_k}$$
$$p(x,t) = \frac{1}{1 + \sum_{k=1}^{n} e^{\mathbf{X}_k \beta_k}}. \tag{2.3}$$

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

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

2.2 Life tables construction

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 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 multinomial logistic regression 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.

In order to construct the life tables for the cause-specific probabilities of death, we assume that the respective number of deaths $D_i(x,t)$ follows a binomial distribution $\mathrm{Bi}(E(x,t), q_i(x,t))$ with probability mass function

$$\mathbb{P}(D_i(x,t) = d_i) = \binom{E(x,t)}{d_i} q_i(x,t)^{d_i} (1 - q_i(x,t))^{E(x,t) - d_i},$$

where $E(x,t)$ is the measure of initial exposure. The likelihood function is given by

$$\mathcal{L}(\mathbf{q}(x,t) \mid \mathbf{d}) = \prod_{i=1}^{n} \binom{E(x,t)}{d_i} q_i(x,t)^{d_i} (1 - q_i(x,t))^{E(x,t) - d_i}.$$

The log-likelihood function is then of the form

$$\ell(\mathbf{q}(x,t) \mid \mathbf{d}) = \sum_{i=1}^{n} \left[\mathsf{ln} \binom{E(x,t)}{d_i} + d_i \, \mathsf{ln}(q_i(x,t)) + (E(x,t) - d_i) \mathsf{ln}(1 - q_i(x,t)) \right].$$

The maximum likelihood estimate of $\mathbf{q}(x,t)$ can be obtained from

$$\hat{q}_i(x,t) \equiv \arg\max_{q_i(x,t)} \ell(\mathbf{q}(x,t) \mid \mathbf{d}).$$

Taking the derivative of log-likelihood with respect to $q_i(x,t)$ and setting it to zero we get

$$\frac{\partial \ell(\mathbf{q}(x,t) \mid \mathbf{d})}{\partial q_i(x,t)} = 0 = \frac{d_i}{q_i(x,t)} - \frac{E(x,t) - d_i}{1 - q_i(x,t)},$$

which leads to

$$\hat{q}_i(x,t) = \frac{d_i}{E(x,t)}.$$

The above expression then corresponds to (1.5). It is essential to note that the likelihood function is concave and hence log-likelihood is indeed maximized at $\hat{q}_i(x,t)$. Therefore, it is reasonable to consider the estimator given by (1.5). It can be also shown that MLE estimate of $q_i(x,t)$ is unbiased:

$$\mathsf{E}\,\hat{q}_i(x,t) = \frac{\mathsf{E}D_i(x,t)}{E(x,t)} = \frac{E(x,t)\,q_i(x,t)}{E(x,t)} = q_i(x,t).$$

Consistency then follows from Chebyshev's inequality:

$$\begin{aligned}\mathbb{P}(|\hat{q}_i(x,t) - q_i(x,t)| \geq \varepsilon) &\leq \frac{\mathsf{var}\,\hat{q}_i(x,t)}{\varepsilon^2} \\ &= \frac{\mathsf{var}\,D_i(x,t)}{E^2(x,t)\,\varepsilon^2} \\ &= \frac{q_i(x,t)(1-q_i(x,t))}{E(x,t)\,\varepsilon^2} \xrightarrow[E(x,t)\to\infty]{} 0.\end{aligned}$$

For the purposes of practical part we introduce so-called cause-specific central mortality rate:

$$m_i(x,t) = \frac{D_i(x,t)}{E^c(x,t)}, \tag{2.5}$$

where $E_i^c(x,t)$ denotes the central exposure to risk, which describes a population in the middle of the year. The formula (2.5) can be viewed as an extension of (1.4). Assuming that deaths as well occur, on average, in the middle of the interval, the following relation can be obtained:

$$q_i(x,t) = \frac{m_i(x,t)}{1 + \frac{1}{2}m_i(x,t)}, \tag{2.6}$$

which is actually a special case of (1.6) for $a_i = 1/2$ and $n_i = 1$. The latter expression is also referred to as a relation between central and crude mortality rate; and, if in (2.6) we substitute $m_i(x,t)$ with (2.5), can be rewritten in a form

$$q_i(x,t) = \frac{D_i(x,t)}{E^c(x,t) + \frac{1}{2}D_i(x,t)}.$$

In the expression above, the denominator actually represents an estimator for $E(x,t)$ used in the binomial model earlier in this section.

Another important statistical measure that we will focus on in the practical part is life expectancy. 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 complete expectation of life is given by

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

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

$$\mathsf{E}K_x \leq \mathsf{E}T_x \leq \mathsf{E}(K_x + 1) \;\Leftrightarrow\; e_x \leq e_x^0 \leq e_x + 1.$$

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:

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

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

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

3. Practical part

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

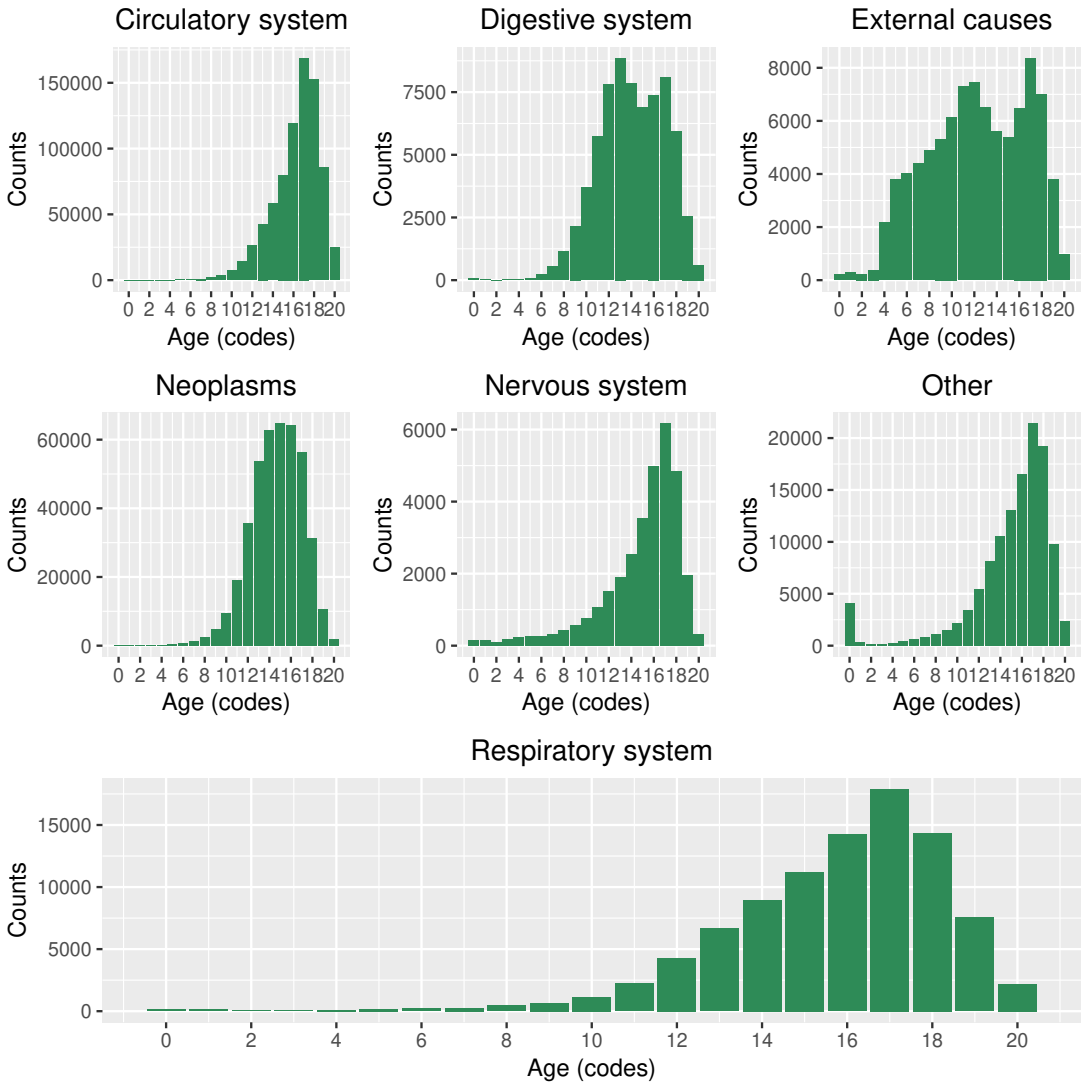

Figure 3.1: Histograms of the numbers of deaths

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 specified as follows:

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

which can be rewritten more formally as

$$\begin{aligned}\mathsf{logit}(mortality_i) =&\beta_0 + \sum_{i=1}^{6} \beta_i \cdot \mathbb{1}[cause = i] + \beta_7 \cdot t \\ &+ \sum_{i=8}^{27} \beta_i \cdot \mathbb{1}[age = i - 7] + \sum_{i=28}^{47} \beta_i \cdot t \cdot \mathbb{1}[age = i - 27].\end{aligned}$$

Proceeding with the fitting of 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]), i.e. we are going to test for whether the initial model is significantly better than the smaller one. We shall consider removing the interaction term from the initial model and test for

$$H_0: \quad \text{model } \mathsf{logit}(mortality) = cause + t + age \text{ is better}$$

against the alternative

$$H_1: \quad \text{model } \mathsf{logit}(mortality) = cause + t + age + t * age \text{ is better.}$$

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 (age-cause specific) formulas (1.6) and (1.7). In Table 3.4 we show the output from `R` software which includes the estimates of the regression coefficients, standard errors, values of

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**

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

Table 3.5: Coefficients of determination

Multiple R^2	Adjusted R^2
0.9246	0.9237

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 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 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). 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.

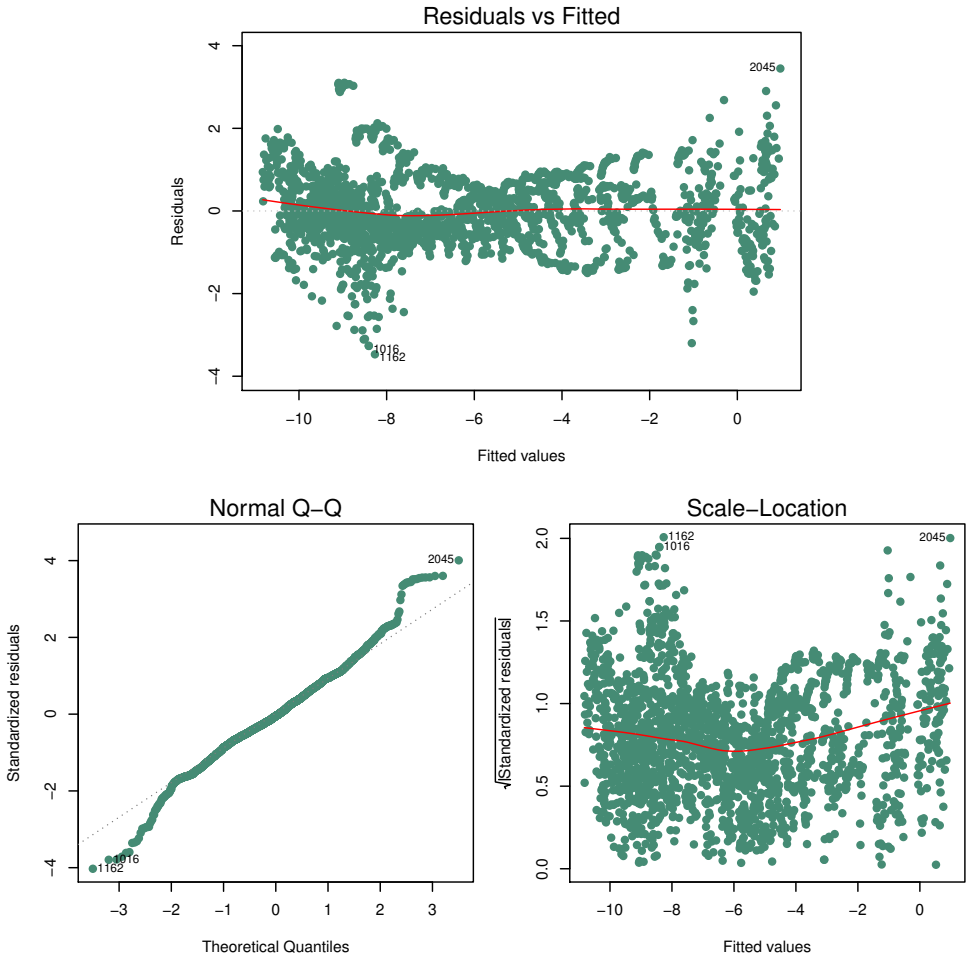

Figure 3.2: Diagnostic plots

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.

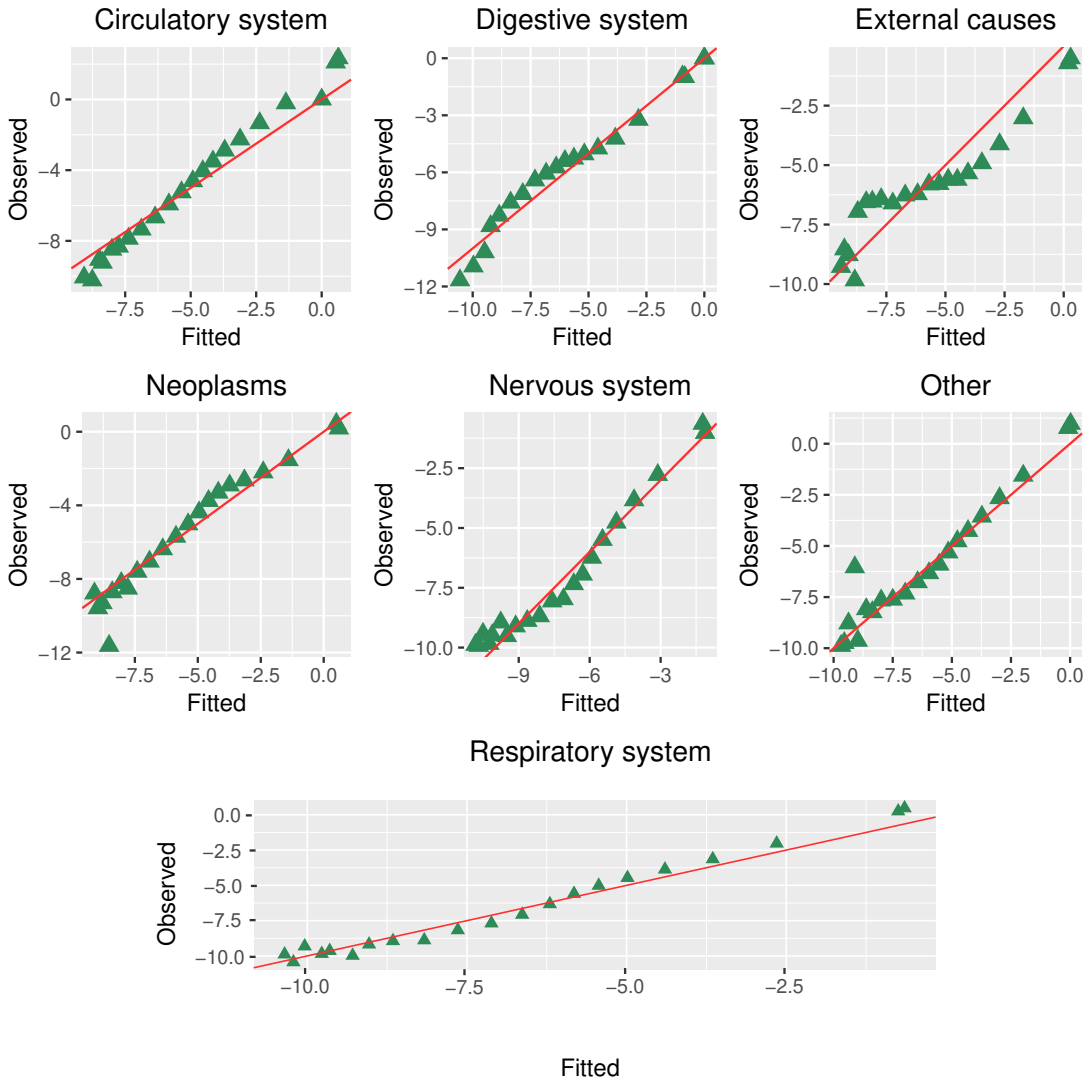

Figure 3.3: Observed vs fitted logit mortality rates in 2017

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.

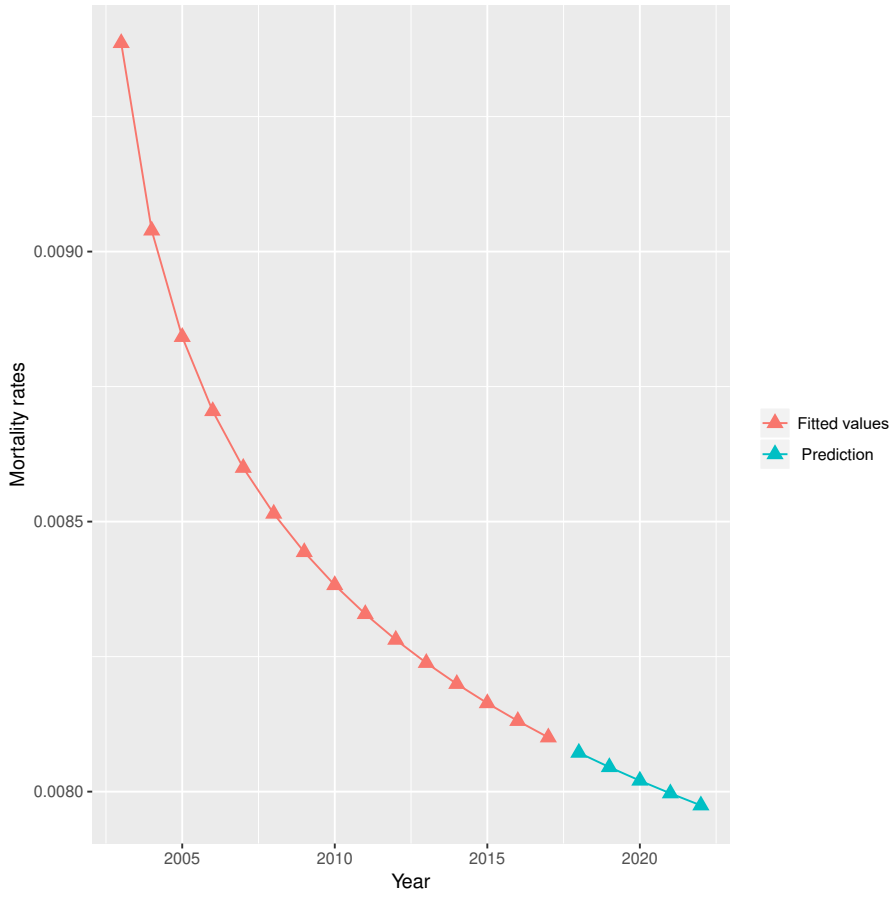

Figure 3.4: Fitted mortality rates with five-year outlook

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.

As outlined in Alai et al. [2015], it is 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. In the following

scenarios we shall assume the independence of competing risks.

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:

$$q_i(x,t) = \frac{\rho_{i,x} \cdot e^{\mathbf{X}_i\beta_i}}{1 + \sum_{k=1}^{n} \rho_{k,x} \cdot e^{\mathbf{X}_k\beta_k}} \tag{3.1}$$

$$p(x,t) = \frac{1}{1 + \sum_{k=1}^{n} \rho_{k,x} \cdot e^{\mathbf{X}_k\beta_k}}. \tag{3.2}$$

It can also be assumed that a shock factor is the same for all age intervals, therefore, we can rewrite (3.1) and (3.2) as

$$q_i(x,t) = \frac{\rho_i \cdot e^{\mathbf{X}_i\beta_i}}{1 + \sum_{k=1}^{n} \rho_k \cdot e^{\mathbf{X}_k\beta_k}} \tag{3.3}$$

$$p(x,t) = \frac{1}{1 + \sum_{k=1}^{n} \rho_k \cdot e^{\mathbf{X}_k\beta_k}}. \tag{3.4}$$

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:

$$q_i(x,t) = \frac{1.15 \cdot e^{\mathbf{X}_i\beta_i}}{1 + \sum_{k=1}^{n} 1.15 \cdot e^{\mathbf{X}_k\beta_k}}$$

$$p(x,t) = \frac{1}{1 + \sum_{k=1}^{n} 1.15 \cdot e^{\mathbf{X}_k\beta_k}}.$$

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

[1] Applied at the projection start date of insurer's liabilities

[2] Best estimate of liabilities

[3] Risk margin

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

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

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

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 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:

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

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

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

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 (projected) 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.

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

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.

Lastly, 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.

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.

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.

William H. Greene. *Econometric Analysis*. 7th edition. Boston: Pearson Education, 2012. ISBN 978-0-273-75356-8.

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*, 2007.

List of Figures

List of Tables