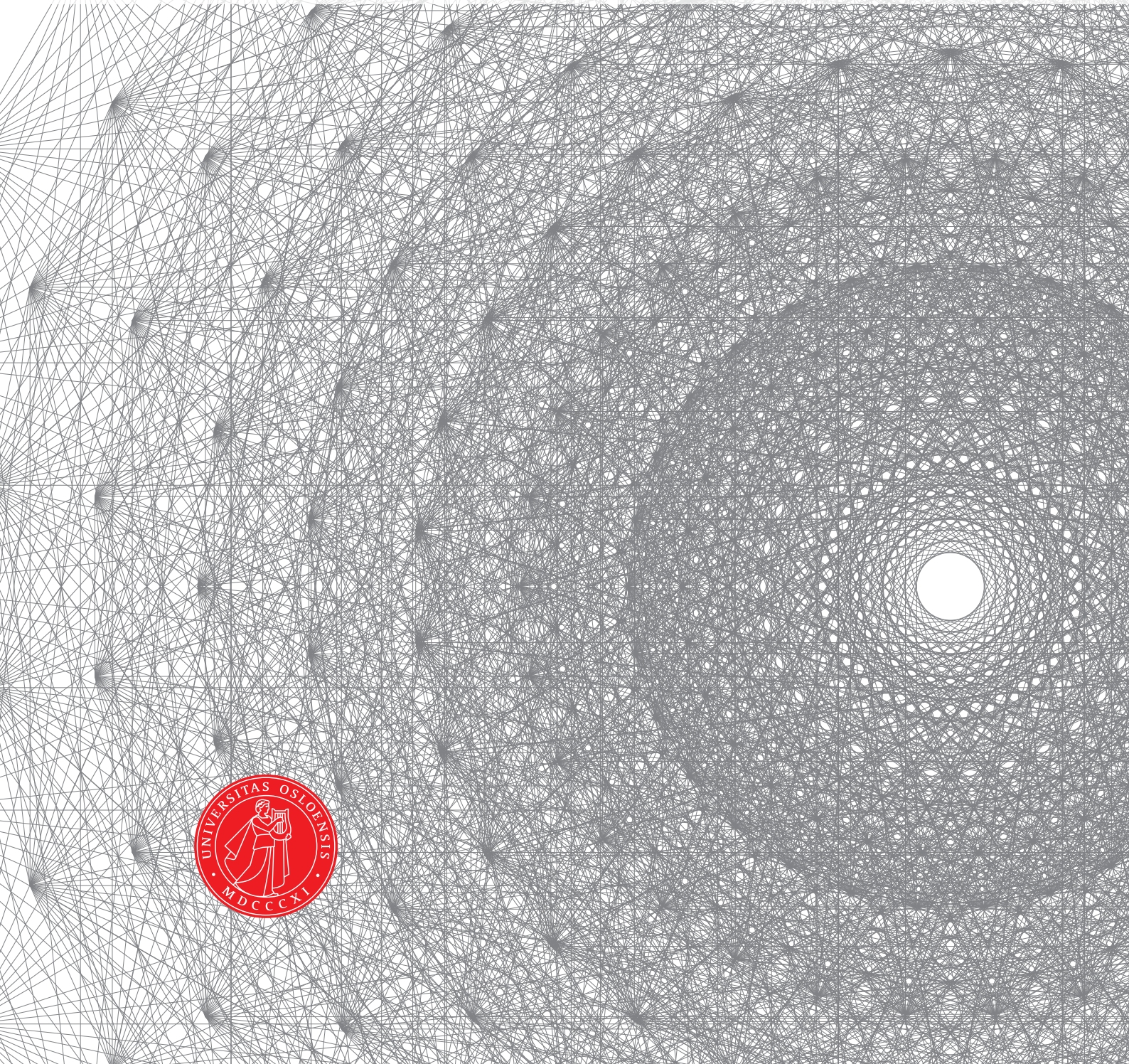
UiO : Department of Mathematics
University of Oslo

GLM and GAM modelling of life insurance data

Amanda Haugnes Rygg

Master's Thesis, Spring 2019



This master's thesis is submitted under the master's programme *Modelling and Data Analysis*, with programme option *Statistics and Data Analysis*, at the Department of Mathematics, University of Oslo. The scope of the thesis is 60 credits.

The front page depicts a section of the root system of the exceptional Lie group E_8, projected into the plane. Lie groups were invented by the Norwegian mathematician Sophus Lie (1842–1899) to express symmetries in differential equations and today they play a central role in various parts of mathematics.

Abstract

As an employee you can have a variety of insurances through your employer, one of them being life insurance covering death due to non-occupational illnesses. With such covers it is essential for the insurance company to know which factors impact risk and through this be able to predict the future risks for new policies. As companies enter and leave the portfolio from year to year, it induces shifts in the insured population and with that shifts in the observed death rates. This complicates the modelling of the death rates. In this thesis, we consider Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs) for prediction and smoothing of nonlinear death rate patterns. We will consider different costumer properties for modelling and discuss differences and similarities in smoothing and predictions done by GLMs and GAMs. We, of course, find that death rates due to non-occupational illnesses increase with age. We also find that the death rates decrease over time. We detect significant differences in death rates of people working in companies with different NACE-codes, also known as activity codes. This is a mandatory statistical classification of the economic activities of a company, put down and regulated by the European Union. Here one of the more surprising discoveries is a higher death rate for women engaging in financial and insurance activities, compared to women in other NACE-codes tested, which usually require less education.

Acknowledgements

I would first and foremost like to thank my supervisors, Anders Rygh Swensen and Ørnulf Borgan, for great assistance throughout the entire period, for patience and for always keeping the door open. I could not have asked for better supervisors. Anders, it is an honour to be your last student before retirement and I wish you all the best in the new time ahead.

I would like to thank IF for providing the data and especially Marianne Hartvig and Anders Klungre for support and help in making this thesis possible.

I would also like to thank my family for always being supportive of the things I do, as well as friends and fellow students for great company and positive distractions in my time as a student. Last, but not least, I would like to thank Gard for all encouragements, love and patience.

Amanda Haugnes Rygg
June 2019, Oslo

Contents

CHAPTER 1

Introduction

As an employee you can have a variety of insurances through your employer, one of them being life insurance. In Norway it is mandatory to have a workers compensation insurance for all employees, regardless of whether they are employed part-time or full-time. In a life insurance perspective, this insurance only apply to occupational death. Most death incidents are however due to non-occupational hazards or diseases. It is therefore common that businesses buy extensions to the mandatory death insurance, so that non-occupational deaths also are covered.

The sum payed out when a person dies varies greatly, depending on which covers the company of the deceased have for their employees. Regardless of which types of death a company choose to cover or how large death cover amounts a company wants for their employees, it is essential for the insurance company to have a good understanding of the risk involved for a potential insurance policy. It is also essential to know which factors impact risk and through this be able to predict the future risks for new policies. In this way the insurer is able to price a given policy according to the risk behaviour of a potential costumer, and the costumer is more likely to get a fair price, depending on the risks the costumer carries.

IF Skadeforsikring NUF (IF) sells an extension to the workers compensation insurance called death due to other illnesses (DOI), which covers death caused by a non-occupational disease. The company has a large DOI portfolio with small and large businesses leaving and entering the portfolio from year to year. This causes great shifts in the insured population and with that, shifts in the observed death rates. To truly know the risks of IF's (or any other similar insurers) future portfolio, it is therefore essential to know which properties of the insured population that drive the risks up or down. Finding these properties and using them in a wise way in modelling may however be challenging as the shifts in death rates may form nonlinear patterns.

Modelling over nonlinear patterns can be done in multiple ways, in this thesis we will look at two. We will look into IF's DOI portfolio to get a better understanding of which costumer properties impact the death rates in the portfolio, and use Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs) to smooth the nonlinear death rate patterns in the observations.

In Chapter 2 we look at the historical development of IF's portfolio. We also give an explanation of the available data on costumer properties and look at the differences in death rates given certain properties. In Chapter 3 we explain how Poisson generalized linear models can be used to study the death rates and

fit such models using costumer properties which we believe best explains the observed death rates.

In Chapter 4 we check if we have overdispersed data and see if we can find an alternative to the Poisson GLM which better explains the observed variation in death rates. In Chapter 5 we take a closer look on how smoothing of nonlinear relations can be done by using splines. In Chapter 6 we check how splines can be used on our data observations through GAMs to smooth the nonlinear death rates.

As different type of models may give a different answer to which costumer properties have the least and most impact, we discuss the similarities and differences between the GLMs and GAMs fitted along with a conclusion on what we have found in Chapter 7. In this last chapter we also give an overview of the analysis done, discuss possible extension and further work.

CHAPTER 2

Description of data

2.1 Data summary

All data used in this thesis has been provided to me by IF Skadeforsikring NUF (IF). The data file used in this thesis was received 29.10.2018 and is not open to the public. Historically IF has not always been structured the way it is today. In the calendar year period 1989 to 1998 IF did not exist, and the insurance was bought from what is now called Storebrand. Storebrand merged with UNI, increasing their portfolio in 1991. IF was established in 1999, when Storebrand merged with Swedish Skandia. IF taking hand of non-life insurance, as a subsidiary company of Storebrand. In the period leading toward 2004 stocks of IF was sold to Finnish Sampo, and 01.01.2004 IF was fully owned by Sampo. Storebrand kept most of the life insurance customers, and IF started to rebuild their life insurance portfolio. These merges and splits are mentioned because they may cause big shifts in the composition of the insured population and thereby big shifts in the observations made.

The data set is collected from companies which were life insurance clients of IF-insurance within the calendar year period of 1989 to 2018. Companies with a life insurance policy insure their employees with a lump sum compensation paid to the bereaved of the deceased employee if a death occurs before the age of 70 years old. In the case of no bereaved a compensation to pay for funeral costs is payed out. In this thesis we will focus on those companies which have insured their employees with a life insurance covering death due to other illnesses (DOI). With a DOI policy a compensation is paid if the employee dies as a result of a non-occupational disease before the age of 70.

The data observations of number of person years and deaths are organised as unique combinations of gender, age, insurance year, what region of the country the company of the insured was stationed in and the activity code of the company, also known as NACE-codes. Person years is here the sum of the duration, given as fractions of a year, in which people within that observation was insured. All variables are described in further detail later. There is a total of 783 749 observations in the data set, with number of person years ranging from 0.01 to 547. Table 2.1 gives a brief summary of the data file.

Adjustments and cleaning of the original data set have been made. Data which came from businesses outside Norway was removed. A decision to leave out the calendar years 1989 and 1990 was made due to lack of data for the periods. Calendar year 2018 is also left out due to the fact that data was

Table 2.1: Summary table of mean age, median age, total person years, deaths and death rates for men and women in the data file.

gender	age-range	mean age	median age	number of person years	number of deaths	death rates
male	20-69	42.28	42	1 567 417	2937	0.00187
female	20-69	42.11	42	672 408	708	0.00105

received within the calendar year, therefore the data may not be representative for the year as a whole. Hence we end up with data from the calendar year period of 1991 to 2017. Businesses which had no county location registered was given a county based on address when this was possible. When an address matched multiple counties or company had multiple addresses, county was reported as multiple. The original data set had an age range from 2 to 87 years old. This is a suspiciously wide range, especially in the lower end. In Norway you cannot, by law, work under the age of 15. This means businesses must have reported the wrong age of some employees or errors have occurred somewhere in the system. Considering this, and the fact that there is little data on insured population under the age of 20, this data is left out of the analysis. Data is also left out for the people of 70 years old or older as they are not covered by DOI. For 8 percent of the data observations the age of the insured was not available. For these observations age was estimated by those who provided the data, based on the age distribution in similar businesses and counties. These data are treated as all the other data in the analysis, but we will look into the possible effects they may have on our observations and look into their characteristics.

Data trouble

The first version of the data for this thesis was received 09.03.2018. The data did, however, have multiple errors causing three different versions to be delivered, where the final data set was received 29.10.18. The final data set used in this thesis has been checked multiple times and is thought to be of good quality. The first data set was at first thought to be good, but as time went by and I looked into the data I discovered otherwise. When analysing the data and fitting models to the observations, weird patterns appeared. I therefore contacted IF, who could confirm that the data I had been working with was wrong. There are big differences between the first data set and the final data set received, so already done tasks had to be redone and this caused quite a lot of extra work within the time available for this thesis.

2.2 Data variables

Gender:

It is natural to split the data according to male and female as we know, from many researches done before, that death rates vary between men and women. Aggregating over all other variables and keeping gender fixed, as done in Table 2.1, we also see a tendency of this in our data, males having 1.8 times higher total death rates than the females.

Age:

Age of a person is defined as the age of the insured the day the contract was signed. Likewise for death counts, age is the age of the person the day he or she died. As we get older the chance of dying changes, age is therefore a natural variable to include in this analysis. We use five-year or ten-year age groups in this part of the thesis to avoid too many observations with no deaths registered. Figure 2.1 shows an already known difference in death frequency between men and women, men dying more frequently through the whole age span. Log of death rates look approximately linear and death rates rise as age increase.

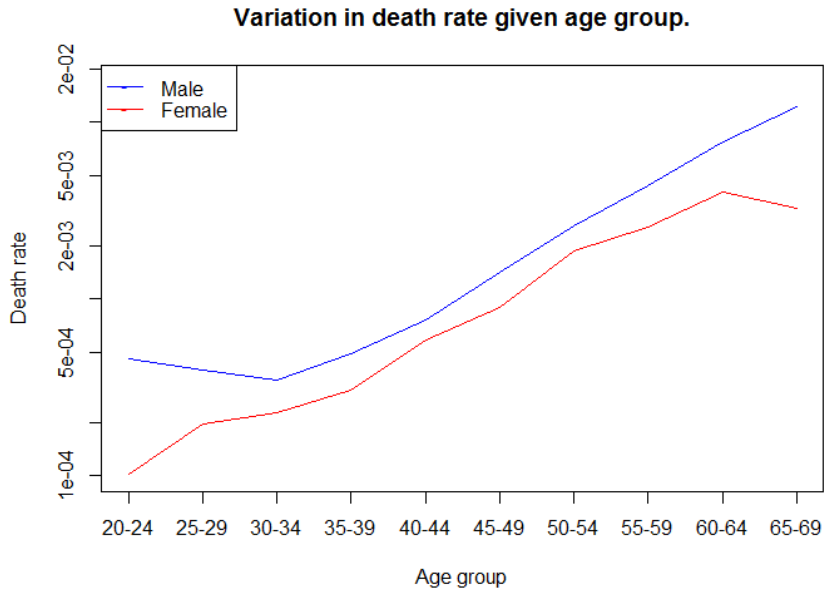


Figure 2.1: Observed variation in death rate given age group and gender. Each gender is represented by a line: Blue = male, red = female. The plot is on log-scale.

Year:

The calendar year or the period within a calendar year in which a person was insured is called the insurance year. Years are included as a variable as we know life expectancy has gone up over the years, and that there may be shifts in death rates for different age groups. Keeping insurance year fixed for different age groups and aggregating over all other variables available, as in Figure 2.2, there seems to be a tendency of decreasing death rates over the years. However, we must consider the historical changes of the portfolio, such as those mentioned in Section 2.1. The split between Storebrand and IF in 2004 make a clear appearance in Figure 2.2 with no deaths registered in any of the age groups. The number of people insured dropped massively in 2004, as can be seen in Figure 2.3, and so also the number of deaths observed. The uncertainty for this

year is therefore higher than for other years as this is the year with the least observations.

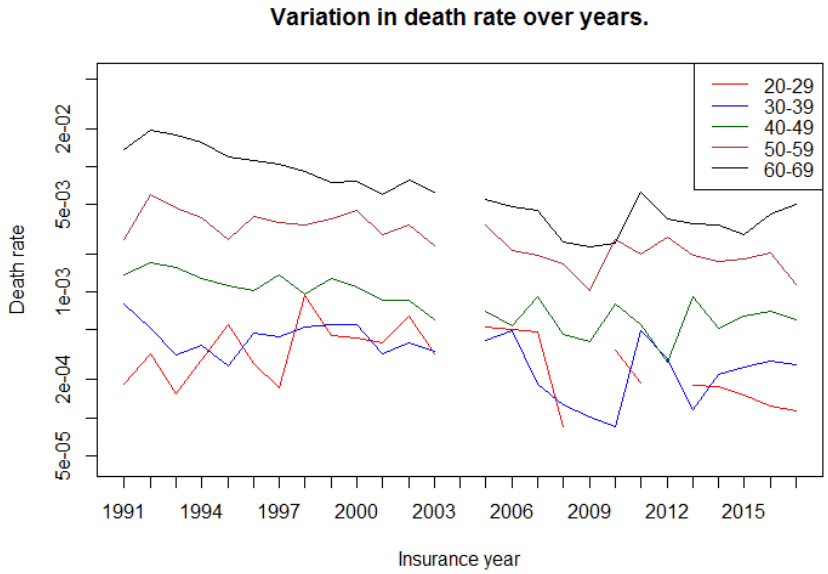


Figure 2.2: Observed variation in death rate given an age group over time. Each age group is represented by a line in the graph: red = 20 to 29 years old, blue = 30 to 39 years old, green = 40 to 49 years old, brown = 50 to 59 years old, black = 60 to 69 years old. When a line is cut there is no deaths observed whitin the age group for the given year. The plot is on log-scale.

From year to year it is normal to have changes in the insured population. Whether the observed tendency in death rate decrease is due to improved health over time or greater shifts in the insured population is therefore hard to tell. It may also look weird that the 20-29 year olds tend to die more frequently for some years than those of age 30-39. However the younger group, of age 20-29 year olds, has a bit higher percentage of males than the older group, of 30-39 years olds. It may also be caused by random variation, as the youngest group has few insured people for some years. If not caused by random variation, the difference may also be caused by other variables such as location or what type of work the younger groups engage in.

As the insured population changes over time, we also know that the number of insured people and the fraction of women insured most likely will vary over time. This becomes clear when we keep gender and year fixed and aggregate our data over all the other variables available, as in Figure 2.3. Also here the split between Storebrand and IF can clearly be seen as a jump in fraction of women and a drop in the number of people insured in the year 2004. The jump in 2013 is due to a big insurance agreement with multiple municipalities. This agreement lasted two years, causing the percentage of women and number of insured to drop back down in 2015 when it was lost.

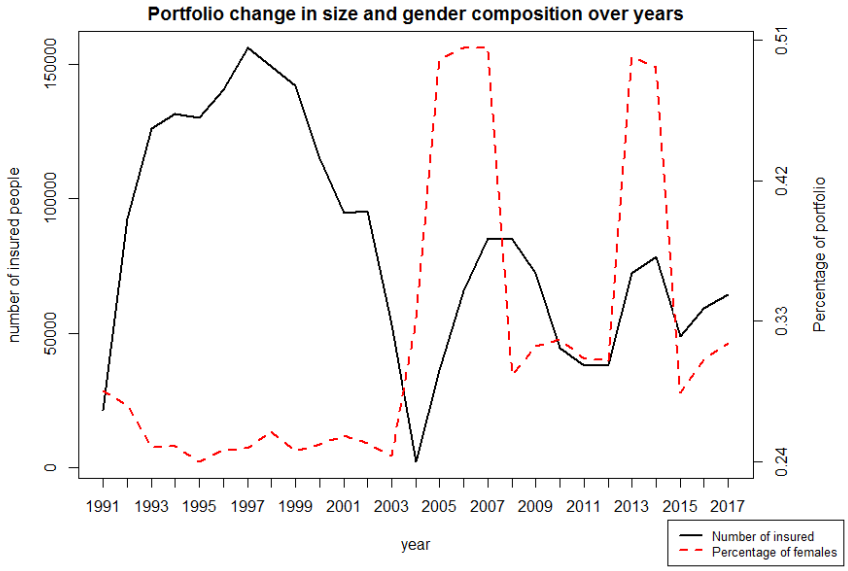


Figure 2.3: Variation in number of people insured and propotion of women in the portfolio over time. Solid black line: number of people insured, reference on left axis. Dotted red line: percentage of women in portfolio, reference on right axis.

Region:

What region of the country the business is located may affect the death rates of the people working there due to differences in lifestyle, focus on human resources and so on. When a company signs an insurance agreement, the company's address or county where it is located is registered. When a company is stationed in multiple counties this variable is reported as multiple. As counties may be too specific they were merged to country regions and the country regions again merged to three greater country parts, as explained by Figure 2.4. Businesses which originally had multiple counties was categorised in a fourth option, being multiple.

Keeping regions fixed and aggregating over all other variables available, as shown in Table 2.2, we see some variation in population structure and death frequency between regions. The two regions with highest mean age, 1: Østlandet and 3: Nord-Norge and Trøndelag, also have the highest death rates. A high fraction of women should lead to a lower death rate, as we know women die less frequent than the men. Looking at our regions however, this may not be the most influential factor. Region 1: Østlandet has the second highest fraction of women, after region 2: Vestlandet and Sørlandet, but has the highest death rates. Age distribution may be more important than fraction of women in the insured population. However this is not something we can tell for certain as it may also be caused by activities the businesses in each region operate in or other factors that the region variable may explain.

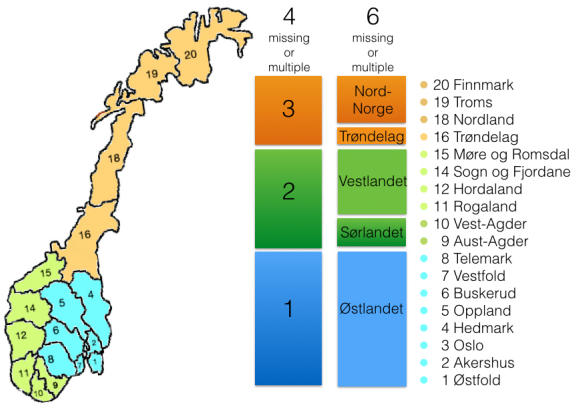


Figure 2.4: Map showing original classification of counties in which companies are registered and how they were merged to three larger regions with a forth category for multiple county registrations.

Table 2.2: Differences in mean age, proportion of women, person years, death counts and death frequency in different country regions.

Country Region	Mean age	Fraction of women	Person years	Deaths	Death rate
1: Østlandet	42.46	0.30	1438699.3	2604	0.00181
2: Vestlandet and Sørlandet	41.83	0.35	477162.4	556	0.00117
3: Nord-Norge and Trøndelag	42.23	0.19	49961.02	72	0.00144
4: Multiple	41.64	0.26	274001.99	413	0.00151

NACE-code:

When a company is established they register a NACE-code, also called an activity code. This is a mandatory statistical classification of economic activities put down and regulated by the European Union. To define the NACE-codes and how they are built up, we have used the definition from the European Union given in the manual from Eurostat 2008. Businesses are placed in the same category of NACE when they engage in the same kind of economic activity. Whether this activity is modern or traditional and where it is preformed, for example, factory or household, does not matter.

NACE-codes have an hierarchical structure. First level is an alphabetical code of letters A to U, specifying the business sections. All NACE-sections are in IF's DOI portfolio, except section T which is paid work in private households. The second, third and fourth level of the NACE-code follow as a two-digit, three-digit and four-digit numerical code specifying the division, group and class of a business. For example, K.65.11 is life insurance, structure and build up of this NACE-code is explained in Table 2.3.

Table 2.3: Explanation of NACE-code structure using life insurance as an example.

Section:	K	Financial and Insurance Activities
Division:	K.65	Insurance, reinsurance and pension funding, except compulsory social security
Group:	K.65.1	Insurance
Class:	K.65.11	Life Insuranse

When IF's system is not able to find the NACE-code of a company the company is given NACE-code "XX.99 Business not mentioned in activity register". The activity code is included as a variable because it may catch up on factors which may influence the death rates. These factors among other include differences in lifestyle and education level. Aggregating over NACE-letters, the business sections, as in Table 2.4, there are quite wide differences in death rates between NACE-sections. As an example we see in Table 2.4 that NACE-section S, has more than double the death rate of NACE-sections B and G. In Table 2.4 we have in chosen to only show NACE-sections with more than 170 deaths in total.

Table 2.4: Differences in person years, death counts and death frequency in chosen NACE-sections. NACE-sections shown are those with more than a totalt of 170 deaths in the data set.

	NACE-section	Person years	Deaths	Death rate
B:	Mining and quarrying	180380.19	258	0.0014
C:	Manufacturing	390604.96	699	0.0018
F:	Construction	114131.7	177	0.0016
G:	Wholesale and retail trade; repair of motor vehicles and motorcycles	295941.4	385	0.0013
H:	Transportation and storage	112972.9	180	0.0016
J:	Information and communication	182334.6	302	0.0017
K:	Financial and Insurance Activities	200851.6	431	0.0022
S:	Other service activities	169722.66	509	0.0030

The fraction of the portfolio consisting of people from a given NACE-section varies over time. The three largest NACE-sections in terms of total time people have been insured in IFs portfolio are:

C - Manufacturing (17.4% of total portfolio)
G - Wholesale and retail trade; repair of motor vehicles and motorcycles (13.2% of total portfolio)
K - Financial and Insurance Activities (9.0% of total portfolio)

The three NACE-sections constitutes 39.6% of the total portfolio. This percentage varies from year to year however, due to companies changing insurer from time to time. In 2005 the fraction in the portfolio that belonged to NACE-sections C, G and K was at its lowest, at 22.6%. It was at its highest in 2017 with 55.6% of the insured population working in the largest NACE-

sections. How these percentages varies over time for each of the three largest NACE-sections are given in Figure 2.5.

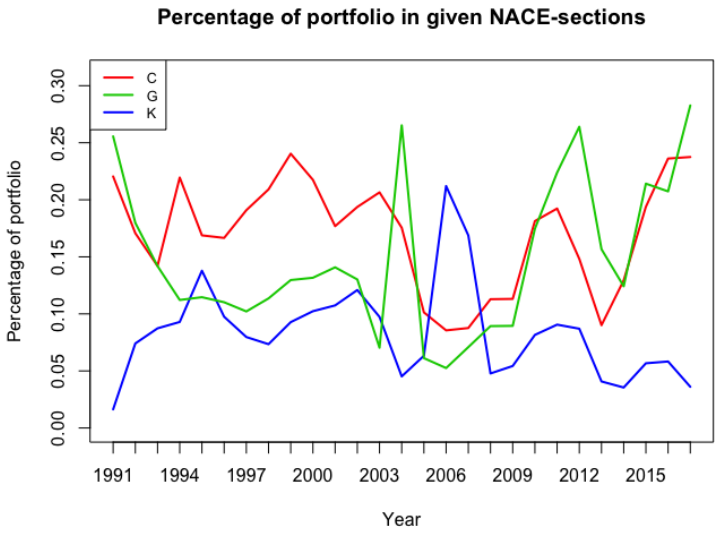


Figure 2.5: Percentage of portfolio belonging to each of the three largest NACE-sections over time. Each NACE-section is represented by a line. Red line: C - Manufacturing. Green line: G - Wholesale and retail trade; repair of motor vehicles and motorcycles. Blue line: K - Financial and Insurance Activities.

2.3 Characteristics and possible influence of data with estimated age

As mentioned in section 2.1, around 8 percent of all the observations in our data set have an estimated age variable, as the age of the insured was not available. To get a grip on what effects this may have on our analysis we now want to look into the characteristics of these observations.

Pulling out only those observations which have estimated age it becomes clear that the estimates make our total portfolio a little older. In Table 2.1, where we included all data available, the mean age of men was 42.28 and the median age was 42. The median age stays the same when only looking at the estimated ages, but the mean becomes slightly higher, with a mean age of 42.56. For the female population the difference is clearer, where mean age increases from 42.11 to 43.16 and the median age increase from 42 to 43 when looking at the data with estimated age vs all data as a whole.

The total death rates for men and women in the data set is pulled down by the observations with estimated age. The total death rate for men and women for the data with estimated age are the same and equal 0.0006. The difference in death rates between genders, observed in Figure 2.1 are almost gone for these data. As seen in Figure 2.6 a), males tend to die slightly more

than the females, but not as clear as before. In other words, the difference between genders becomes smaller when we include all data.

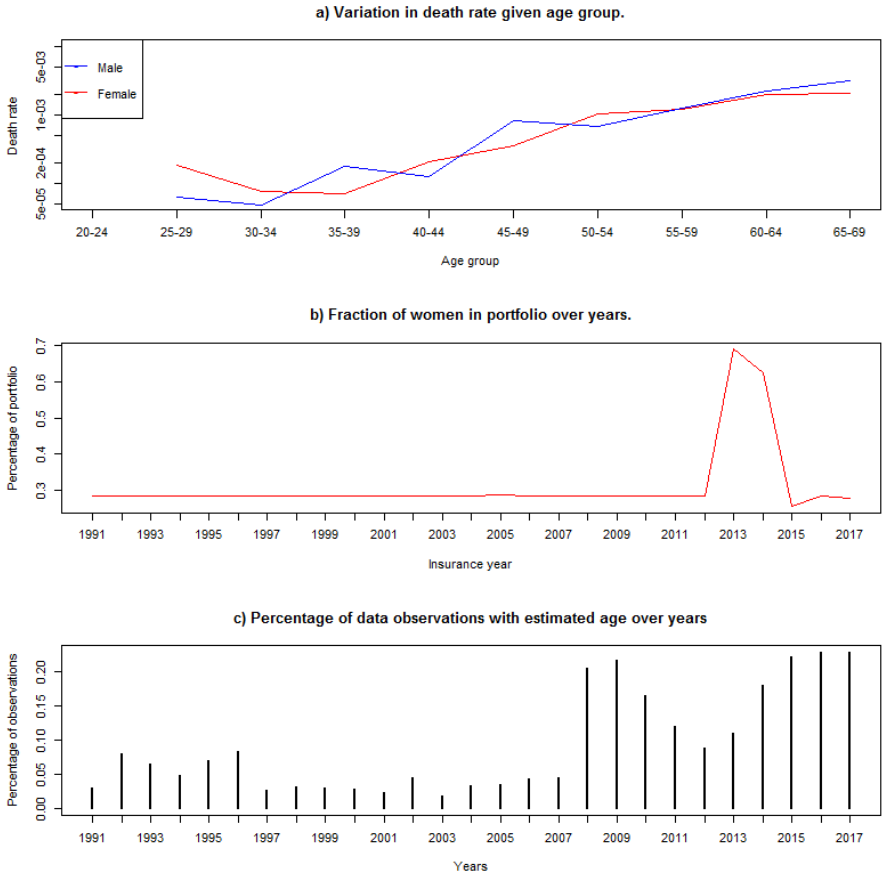


Figure 2.6: a) Observed variation in death rate given age group and gender for data observations with estimated age. Each gender is represented by a line: Blue = male, red = female. Death rates are plotted on a log-scale. b) Percentage of women in the portfolio over time for data observations with estimated age. c) Fraction of all observations which have estimated age.

The fraction of age estimated observations vary from year to year. For instance there is a major jump in percentage of women in the portfolio for data with estimated age in the years 2013 and 2014, as seen in Figure 2.6 b). This indicates that there is a higher percentage of the observations from women that have estimated age for these years. This is also the case if we look at the portfolio as a whole, where 11.5 percent of the women's observations and 6.5 percent of the mens observations have estimated age.

If we look at the total portfolio over years the percentage, and thereby the influence, of the observations with estimated age varies. As we can see in Figure 2.6 c), the percentage have been at its highest for the three last years at a

percentage close to 23 percent. If removing the data with estimated age we will therefore have the biggest cut in observations for these years.

CHAPTER 3

Generalized linear modelling

To model the mortality data described in Chapter 2 we will here use Generalized Linear Models (GLMs). We assume that each death count is independent and that the deaths counts have a Poisson distribution. Basic concepts of Poisson regression, parameter estimation, model fitting and model selection will be described.

3.1 Reasoning for choosing Poisson and basic consepts of Poisson Regression

Reasoning

Aggregating our data to only include ten-year age groups, gender and year as categorical variables we end up with 270 observations, being unique combinations of the three variables in use. We here consider the age groups, calendar year and gender as categorical variables.

The number of person years in each observation is large, with an average of 8296 person years. We think of the death counts as the sum of a large number of Bernoulli trials, where the success parameter, here the probability of dying, is small and the number of trials, person years, is large. The sum of successes from Bernoulli trials has a Binomial distribution (Casella and Berger 2002, p. 89). The probabilities in a Binomial distribution is again close to the probabilities in a Poisson distribution when there is a large number of trials with small success probabilities (Casella and Berger 2002, pp. 66–67). Hence it is reasonable to consider the death counts as Poisson distributed in our case.

Specification of model set-up

Our response variable X_{ijk} is the number of deaths for people of age group: i = 20-29, 30-39,..., 60-69, gender: j = 0,1, 0 indicating male and 1 indicating female, insured in year: k = 1991,..., 2017. The expected number of deaths in each group will depend on n_{ijk}, the sum of person years within the group. We would therefore like to model the rate per person year. Thus, $X_{ijk} \sim poisson(\lambda_{ijk} \cdot n_{ijk})$, λ_{ijk} being the chance of a death for a person that has been insured for the whole year and n_{ijk} being the total person years in the group. If not else is specified Agresti (2015, pp. 122,228-233) is used as reference in this section.

Using the canonical logarithmic link function $\eta_{ijk} = \log(\lambda_{ijk})$ or equivalent $\lambda_{ijk} = \exp(\eta_{ijk})$ we get:

$$\mu_{ijk} = E(X_{ijk}) = n_{ijk} \cdot \lambda_{ijk} = n_{ijk} \cdot \exp(\eta_{ijk}) = \exp\{\log(n_{ijk}) + \eta_{ijk}\} \quad (3.1)$$

$$\log(\mu_{ijk}) = \log(n_{ijk}) + \eta_{ijk} = \log(n_{ijk}) + (z_{ijk})^T\beta \quad (3.2)$$

Using the log link we see that $\log(n_{ijk})$ is an additive known constant term in the linear predictor, this will be used as an "offset" in our modelling. The vector z_{ijk} contains values of the explanatory variables for each age group i, gender j and year k combination. β is here a vector of our regression parameters, how these are found is explained in the next subsection.

The probability mass function (pmf) of the Poisson distribution is defined as (Casella and Berger 2002, p. 92):

$$P(X_{ijk} = x_{ijk}) = \frac{\mu_{ijk}^{x_{ijk}} \exp(-\mu_{ijk})}{x_{ijk}!} \quad (3.3)$$

Inserting the definition of μ_{ijk} from equation (3.1) above, we get the Poisson regression model:

$$P(X_{ijk} = x_{ijk}) = \frac{\exp\{\log(n_{ijk}) + \eta_{ijk}\}^{x_{ijk}} \exp(-\exp\{\log(n_{ijk}) + \eta_{ijk}\})}{x_{ijk}!} \quad (3.4)$$

Parameter estimation

Fitting the Poisson pmf to our data requires estimation of regression parameters β. This is done using the maximum likelihood (ML) method. The likelihood function of the Poisson pmf can be expressed as (Zuur et al. 2009, p. 214);

$$L(x;\mu) = \prod_{ijk} P(X_{ijk} = x_{ijk}) = \prod_{ijk} \frac{\mu_{ijk}^{x_{ijk}} \exp(-\mu_{ijk})}{x_{ijk}!} \quad (3.5)$$

and depends on β via equation (3.2). The maximum likelihood estimates of β are found by maximizing the likelihood function with respect to β. The estimated variance-covariance matrix of the β estimates is found by taking the inverse of the $-Hessian$ matrix. Here the $-Hessian$ matrix is found by taking the log of the likelihood function and differentiate it twice. The estimated variance-covariance matrix is used to obtain the standard errors of the β estimates (Agresti 2015, pp. 137–139).

3.2 Models selection

There are multiple models which may fit our data. We however, just want one model, the model that best describe the observed variation in the data set. We will now discuss three different ways we can compare possible models and how we choose our preferred model.

Selection criteria

We select a model by comparing deviances between our models using the likelihood ratio test (LRT). For Poisson distributed data the deviance equals (Zuur et al. 2009, p. 217);

$$D(x;\widehat{\mu}) = 2\{\log[L(x;x)] - \log[L(x;\widehat{\mu})]\} = 2\sum_{i=1}^{n}\{x_i \cdot \log(\frac{x_i}{\widehat{\mu}_i}) - x_i + \widehat{\mu}_i\} \quad (3.6)$$

where $L(x;x)$ is the maximum of the likelihood of the saturated model and $L(x;\widehat{\mu})$ is the maximum of the likelihood of a given fitted model. Here $n =$ the number of grouped observations and $p =$ the number of parameters.

Taking differences of deviances of nested models, where the parameter space of the smaller model, model 0, is contained in the bigger model, model 1, the terms involving the saturated model cancel out from the deviance equation. We get a likelihood-ratio statistic;

$$D(x;\widehat{\mu}_0) - D(x;\widehat{\mu}_1) = 2\{\log[L(x;\widehat{\mu}_1)] - \log[L(x;\widehat{\mu}_0)]\}$$

which has an approximate chi-squared distribution with df = p_1 - p_0, assuming that the smaller model holds (Agresti 2015, pp. 133–134). When testing the nested models against each other the null hypothesis is that the smaller model holds. By this we mean that the extra variables in the bigger model do not significantly improve the fit of the model. If we get a p-value less than 0.05, the null hypothesis is rejected and the bigger model is chosen as the adequate model.

Models which are not nested, as the models fitted in chapter 6, can not be compared using deviance differences and likelihood ratio testing. Two methods which then may be used to compare and select models are Akaike information criteria (AIC) and Bayesian information criteria (BIC). Fitting models using maximum likelihood (ML) the methods allow for comparison between models that are not nested, as long as they are fitted to the same data (Jong and Heller 2008, p. 63).

If p is the number of estimated parameters in the model and $L(x,\widehat{\mu})$ is the ML value of the likelihood function, AIC is in general defined as (Jong and Heller 2008, p. 62):

$$\text{AIC} = -2(\log(L(x,\widehat{\mu})) - p)$$

AIC judges which model is expected to have the sample fit close to the true model fit and gives a penalty based on the number of parameters in use. The lower the AIC the better the model. Comparing possible models for our data the preferred model is the one which has the lowest AIC (Jong and Heller 2008, p. 63).

AIC tend to prefer bigger models. We therefore also look at the Bayesian information criteria (BIC), which looks a lot like AIC, but gives a greater penalty for adding parameters. BIC is in general defined as (Jong and Heller 2008, p. 62):

$$\text{BIC} = -2\log(L(x,\widehat{\mu})) + \log(\text{number of observations}) \cdot p$$

Here, p is the number of estimated parameters in the model and $L(x,\widehat{\mu})$ is the ML value of the likelihood function. As in the AIC case, the model with the

lowest BIC-score is the preferred model. When the number of observations is large, as the case will be when we include more variables, the BIC tend to select models which may be too simple (Jong and Heller 2008, p. 63). When models are nested, model choice will mainly be based on deviance comparison and LRT. When the p-value of the LRT is close to the 0.05 threshold however, or we compare non-nested models, we will look at the AIC-score to choose a model. When the AIC-scores are similar, we will check significance of variable effects and use BIC to decide between models.

3.3 Fitting and comparing models

We first tried fitting a joint model for both genders. The joint model did however become really messy and difficult to interpret. To get a better overlook of the data and what happened when a model was fitted, the data set was split into two new sets, one for each gender with 135 observations each. A saturated model was then fitted for each data set separately in `R` using the `glm` function. The saturated models were specified in the following way:

```
> Saturated.model = glm(deaths ~ offset(log(personYears)) +
  ageGroup*year , family=poisson, data = subset for given gender)
```

Male - Model fitting

We first look at the data for males and fit a saturated model to the data, as explained above. Trying to remove the whole interaction effect between year and age group, as done in model $M2_m$, is rejected by the likelihood ratio test, see Table 3.1 for details. We therefore need to look at alternative ways to reduce the number of parameters.

When we look closer into the model summary of the saturated model, year stick out as a variable which has a lot of noise in the interaction estimates. It is also few observations within some of the age groups for certain years. We therefore create a new variable, Y_{group}, which groups together three and three years eg. Y_{1991}, Y_{1992}, $Y_{1993} \longrightarrow Y_{1991-1993}$, hence we get 9 year categories from the original 27.

Table 3.1: Deviance table showing model summaries and hypothesis testing of models fitted on male data with a total of 135 observations. Main components for variables are A_{group}: age group and Y: year. Y_{group} is a categorical year variable where three and three years are merged together. Δ = deviance and p = number of parameters.

model	variables	$2 \cdot$ log likelihood	p	Δ	Null hypothesis	p-value LRT
$M0_m$	A_{group}+Y + A_{group}:Y	-539,8	135	-	-	-
$M1_m$	A_{group}+Y + A_{group}:Y_{group}	-623.7	63	83.77	$Y_{1991} = Y_{1992} = Y_{1993}$; ... ;$Y_{2015} = Y_{2016} = Y_{2017}$:$-2 \log \frac{L(M1_m)}{L(M0_m)}$	0.1619
$M2_m$	A_{group}+Y_{group}	-703.0	31	163.11	A_{group}:Y=0 :$-2 \log \frac{L(M2_m)}{L(M0_m)}$	0.0002 ***

A new model, $M1_m$, is fitted to the male data set using Y_{group} in the interaction effects. When comparing the deviances of male models $M0_m$ and $M1_m$, as seen in Table 3.1, $M1_m$ is chosen as the adequate model.

Male - Deviance residual diagnostics

To check model fit we would like to look at the residuals of the model. There exists several types of residuals for GLMs, we will however just look at one type for now, deviance residuals, which is the default residuals used by R in GLM modelling.

The deviance residuals are elements of the unexplained variation by the fitted model. For Poisson distributed data the deviance residuals are given by (Zuur et al. 2009, pp. 229-230):

$$\text{Deviance residual} = \sqrt{d_{ijk}} \cdot sign(x_{ijk} - \widehat{\mu}_{ijk}) \tag{3.7}$$

where,

$$d_{ijk} = 2\{x_{ijk} \cdot \log(\frac{x_{ijk}}{\widehat{\mu}_{ijk}}) - x_{ijk} + \widehat{\mu}_{ijk}\}$$

The deviance of a model is the sum of d_{ijk}, which we recognize from equation (3.6). The residuals should be random and normally distributed, have a constant variance and zero mean (Agresti 2015, pp. 56–57).

We check for normality and randomness by plotting the fitted values against the residuals, as well as making a quantile-quantile (QQ) plot of the residuals from our chosen model. Looking at the plot of fitted values, Figure 3.1 a), the residuals of model $M1_m$ looks random, with an approximately equal proportion of positive and negative residuals spread around zero for all fitted values. Looking at our QQ-plot in Figure 3.1 b), the majority of the residuals lie close to the line, indicating that the residuals of our model is close to having a normal distribution (Agresti 2015, pp. 101–103).

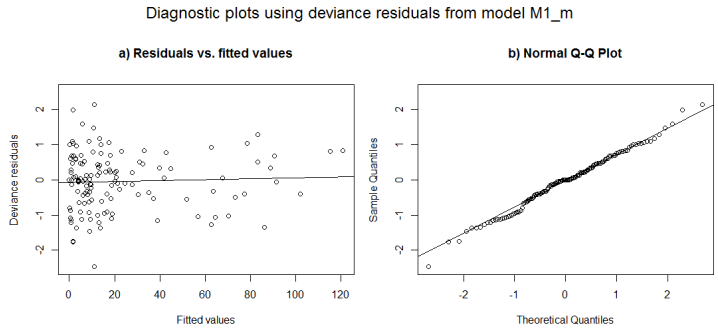


Figure 3.1: Residual diagnostic plots of model $M1_m$. a) Deviance residuals plotted over fitted values. Trend line of residuals is given as a black line in the plot. b) QQ-plot of deviance residuals.

Residuals may also form patterns over given variables. We therefore check if the residuals look random over the variables used in the model. Looking at the residuals over age groups, in the top panel of Figure 3.2, the median of the

residuals seems to be centred around zero in each group. This also seem to be the case when looking at the medians of residuals in given year groups, in the bottom panel of the same figure. Looking at the middle panel however, which is the residuals within each year, the medians of the residuals are less centred. The way the medians shift up and down do however not have a clear pattern and look random. All in all the residuals of the model seem to meet the model assumptions to a pleasent degree and $M1_m$ is chosen as an adequate model for the male data set observations.

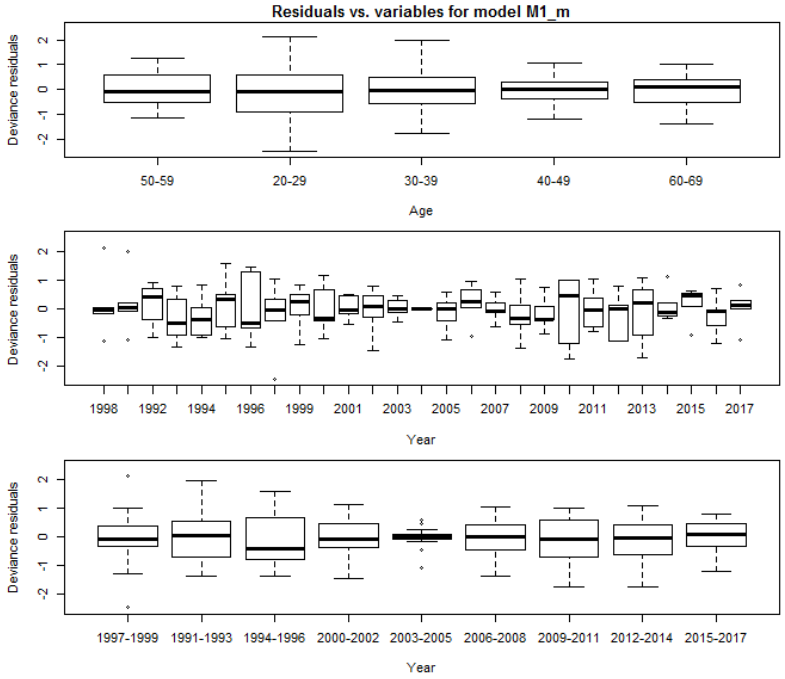


Figure 3.2: Residual diagnostic plots of model $M1_m$, deviance residuals are plotted against variables used in the model. Top panel: residuals over age groups. Middle panel: residuals over years. Bottom panel: residuals over year groups. Reference levels for each of the categorical variables are given to the far left in the corresponding variabel-panels.

Female - Model fitting

We now want to find an adequate model for the female population. Trying to remove the whole interaction effect between year and age group from the full model, as done in model $M2_f$, is as in the male case, rejected by the likelihood ratio test, see Table 3.2 for details. We therefore do the same here as we did for the males, where we try to use grouped years as variables for the interaction effects, redcuing the number of parameters.

Comparing the female models $M0_f$ and $M1_f$, as in Table 3.2, the simpler model is not preferred by the likelihood ratio test. The p-value is however not far from the 0.05 threshold. We therefore still choose $M1_f$ as a preferred model

Table 3.2: Deviance table showing model summaries and hypothesis testing of models fitted on female data with a total of 135 observations . Main components for variables are A_{group}: age group and Y: year. Y_{group} is a categorical year variable where three and three years are merged together. Δ = deviance and p = number of parameters.

model	variables	$2 \cdot$ log likelihood	p	Δ	Null hypothesis	p-value LRT
$M0_f$	A_{group}+Y + A_{group}:Y	-354.8	135	-	-	-
$M1_f$	A_{group}+Y+ A_{group}:Y_{group}	-451.0	63	96.25	$Y_{1991} = Y_{1992} = Y_{1993}$; ... ;$Y_{2015} = Y_{2016} = Y_{2017}$:$-2\log\frac{L(M1_f)}{L(M0_f)}$	0.0297 *
$M2_f$	A_{group}+Y	-510.6	31	155.77	A_{group}:Y=0: $-2\log\frac{L(M2_f)}{L(M0_f)}$	0.0008 ***

as we believe variables later added may explain the variation observed over the years for females within the different age groups.

Female - Deviance residual diagnostics

We use the same type of residuals for the female model, as we did for the male. To check for randomness in the deviance residuals of model $M1_f$, we plot the residuals over fitted values, as in Figure 3.3 a). The residuals form some bands in the lower left end of the plot. This is however due to a small number of deaths, and not to such degree that it is of concern. The trendline of the residuals also show that they are centred just below zero and have a close to flat slope over the fitted values.

Checking the distribution of the residuals through our QQ-plot, in Figure 3.3 b), the residuals look approximately normal. There is a slight bump around zero and an indication of a light tail in the higher end of our theoretical quantiles. The majority of the residuals do however lie close to the line, indicating that the residuals of the model are close to having a normal distribution.

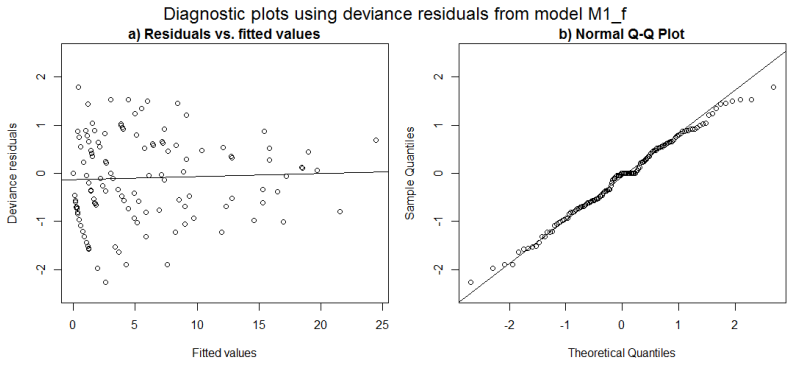


Figure 3.3: Residual diagnostic plots of model $M1_f$. a) Deviance residuals plotted over fitted values. Trend line of residuals is given as a black line in the plot. b) QQ-plot of deviance residuals.

We check if residuals look random over the variables used in the model in the same way as we did for the males. Looking at the residuals over both age groups and year groups, in the top and bottom panels of Figure 3.4, the medians of the residuals seem to be centred around zero in each group. Looking at residuals over years, in the middle panel of the same figure, we see the same tendency as we did for the male model residuals. The medians of the residuals fluctuate up and down from year to year, but there is no clear pattern in how they do so. The model seems to meet the model assumptions to a pleasant degree and is therefore considered an adequate model for the female data set observations.

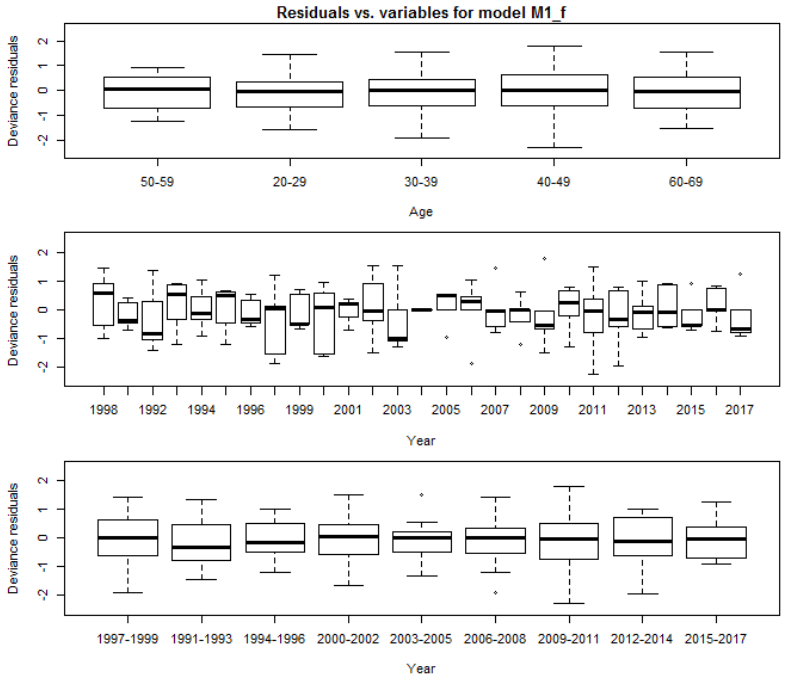


Figure 3.4: Residual diagnostic plots of model $M1_f$, deviance residuals are plotted against variables used in the model. Top panel: residuals over age groups. Middle panel: residuals over years. Bottom panel: residuals over year groups. Reference levels for each of the categorical variables are given to the far left in the corresponding variabel-panels.

3.4 Check and interpretation of chosen models

Prediction of death rates in each group is calculated by inserting the maximum likelihood estimates of models $M1_m$ and $M1_f$ into equation (3.1) (Agresti 2015, pp. 28–29). We now want to check if the predictions made are satisfactory, and that they match what we observed in Section 2.2.

Check of estimates

Looking at the predictions made for male and female over age, as in Figure 3.5, we see that the difference earlier observed between the genders in Figure 2.1 still is intact. The plot is made by aggregating over the predicted number of deaths in each group in our data instead of the observed number of deaths, which we did in Figure 2.1. Aggregating over all variables except gender in our predicted values the male death rate is 0.00187 and the female death rate is 0.00105. This is the same as we saw in our data summary, in Table 2.1, which means the overal difference between genders for predicted death rates is the same as for the observed death rates.

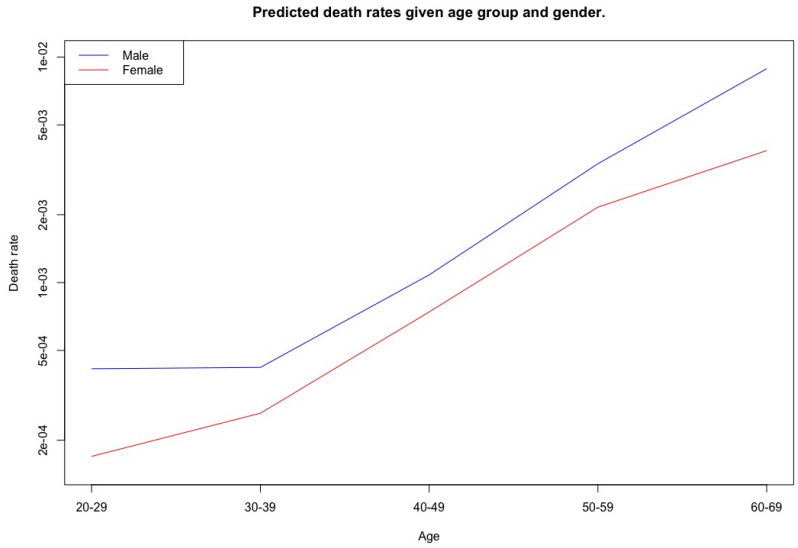


Figure 3.5: Predicted death rates by models $M1_m$ and $M1_f$ for males and females in given age groups. Each gender is represented by a line: Blue = male, red = female. Predictions are plotted on a log-scale.

Over the insurance years in our observations made in Section 2.2, we said that the group of 20-29 year olds may die more frequent than the group of 30-39 year olds due to the fact that the youngest group has a higher proportion of males. This hypothesis is partly true. For our female predictions the youngest group has the lowest death rate through most of the time span. For the male predictions however, this is not the case. Therefore when we aggregate our predictions keeping only age group and years fixed, as in Figure 3.6, the youngest group still tend to have a higher death rate up till the last eight years, meaning there must be another factor or random variation causing the higher death rates in the youngest age group.

The major drop in death rates in year 2004 which we observed in Figure 2.2, also appear in the prediction plot, Figure 3.6. The clear appearance of year 2004 may indicate that our models may not be smoothed enough over years. We do however know that the models now in use is true to the real data, in that there is no observed deaths in 2004. The models also seem to do predictions

matching what we earlier observed over years. Predictions have more defined differences between the age groups and smoother lines over years than what the observations have, which is something we want. We therefore keep models $M1_m$ and $M1_f$ as adequate models.

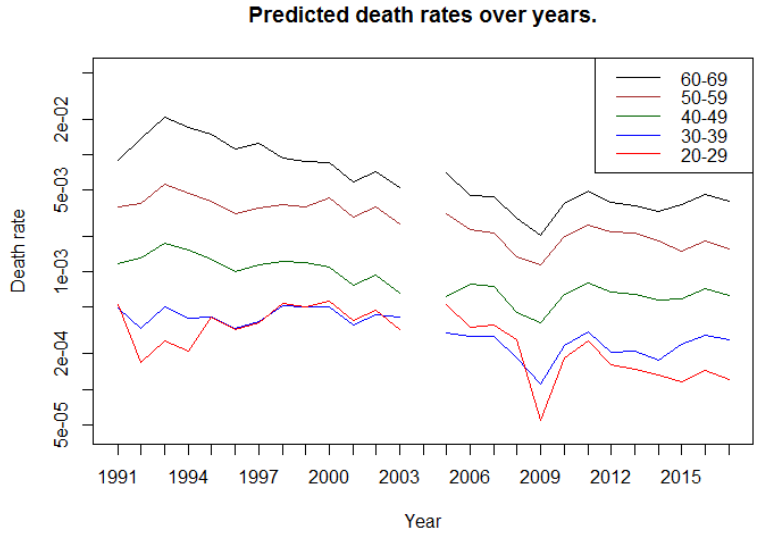


Figure 3.6: Predicted death rates over time in given age groups by model $M1_m$ and $M1_f$. Predictions were made using the covaraite values in the male and female data sets which $M1_m$ and $M1_f$ were fitted to. The predictions were then indluded in the data sets and the data from both models were aggregated together keeping year and age group fixed. Each age group is represented by a line in the graph: red = 20 to 29 years old, blue = 30 to 39 years old, green = 40 to 49 years old, brown = 50 to 59 years old, black = 60 to 69 years old. Predictions are plotted on a log-scale.

3.5 NACE-section effects

In the previous section we looked at models with grouped age, year and grouped years as variables. We now want to see what effects NACE-sections may have to the death rates and include this as a variable. We will not look at all the NACE-sections at once, but will for this thesis focus on the three largest NACE-sections in the data set. We will look at:

C - Manufacturing (699 deaths - Male: 613 and female: 86)
K - Financial and Insurance Activities (431 deaths - Male: 302 and female: 129)
G - Wholesale and retail trade; repair of motor vehicles and motorcycles.
(385 deaths - Male: 312 and female: 73)

Observations

Before we start with a more detailed modelling we will have a look at some summaries of the data observations. To look at the observed deaths rates we use the age groups of ten and ten years. We have done so to get a better overview of the trends between NACE-sections, as we know there are few observed deaths for each individual age when we include more variables. Over years we have used groups of three calendar years, as we will use this as a variable when later fitting models and it makes it easier to see the possible differences in trends between NACE-sections. Grouping the data observations in this way yields 135 observations (5 age groups x 9 year groups x 3 NACE-sections) for each gender.

Starting with the male observations, Figure 3.7, there are indications of differences between the NACE-sections, both over age and year. In the left panel of Figure 3.7 it looks like NACE-sections G and C have quite parallel trends over age, G having lower death rates than C over the whole age span. We see the same over the year span in the right panel of Figure 3.7. NACE-section G here has lower death rates than C for most year groups, except from 2003-2005 where we know there is little exposure. NACE-section K has a steeper curve over the age span than the two other sections, with one of the lowest death rates for the youngest group and the highest death rate for the oldest group. Over the year span NACE-section K has a curve shape differing from the curve shape of the other two sections.

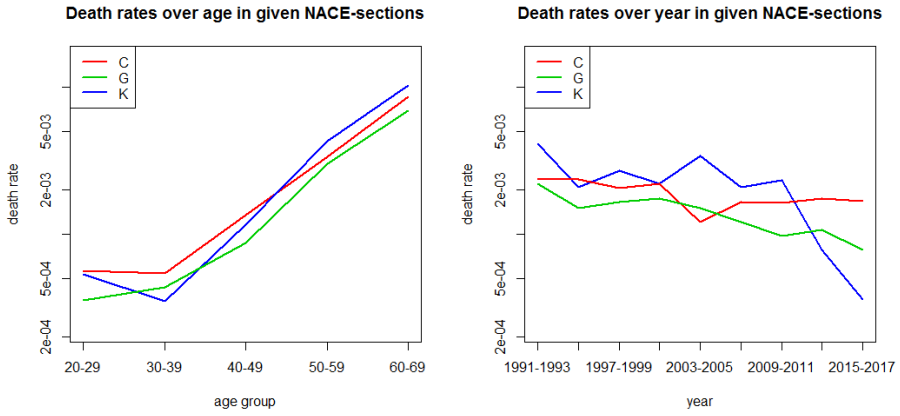


Figure 3.7: Observed male death rates over age (left panel) and year (right panel) in different NACE-sections, each section is represented by a line. Red line: NACE-section C - Manufacturing, green line: NACE-section G - Wholesale and retail trade; repair of motor vehicles and motorcycles, blue line: NACE-section K - Financial and Insurance Activities. Observations are plotted on a log-scale.

For the female observations we have different tendencies in trends between NACE-sections than what we had for the males. Over age, Figure 3.8 a), NACE-sections K and C look close to parallell. NACE-section G has low death rates for the two youngest age groups, but from age group 40-49 onwards the section has a death rate almost identical to NACE-section C. Over years, Figure

3.8 b), we see some of the same tendencies as we did over age. NACE-section K has a death rate higher than NACE-section C for most years, and NACE-section G has death rates close to the death rates of NACE-section C.

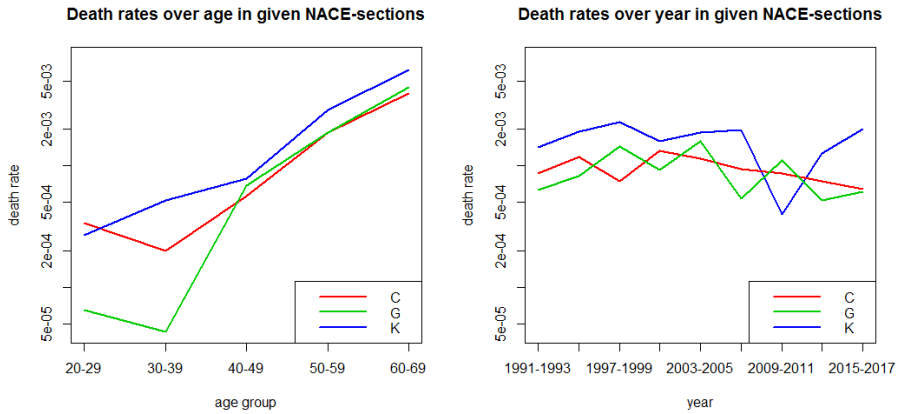


Figure 3.8: Observed female death rates over age (left panel) and year (right panel) in different NACE-sections, each section is represented by a line. Red line: NACE-section C - Manufacturing, green line: NACE-section G - Wholesale and retail trade; repair of motor vehicles and motorcycles, blue line: NACE-section K - Financial and Insurance Activities. Observations are plotted on a log-scale.

Before we start fitting models we change the categorical age variable to a numeric variable with age ranging from 20 to 69. Which means that the models will have a linear term in age. We do however keep the categorical three years span variable, Y_{group}, explained in the previous section and the categorical NACE-section variable. Aggregating our data with the three largest NACE-sections, grouped years and single age observations should leave us with 1350 observations for each gender (3 NACE-sections x 50 ages x 9 year groups). We do however not have exposures for each variable combination and end up with a male data set of 1349 observations and a female data set with 1347 observations[1].

Male - Model fitting and model selection

We fit models using forward selection, first specifying a model with the three main effects as:

```
> M1.nace.m = glm(deaths ~ offset(log(personYears)) +
  age + yearGroup + NaceMain , family=poisson, data = subset for males)
```

[1] Variable combinations with no exposure:
- Male of age 69 working in NACE-sections G insured at some point between 2003-2005.
- Female of age 69 working in NACE-section G insured at some point between 1997-1999.
- Female of age 69 working in NACE-section G insured at some point between 2003-2005.
- Female of age 20 working in NACE-section K insured at some point between 2006-2008.

We then choose which interaction to add to the model using the `add1()` command in `R`, specified as:

```
> add1(M1.nace.m, scope= ~age*yearGroup*NaceMain, test="LRT")
```

The function returns a list of all the first order interaction effects, the deviance and AIC of a model when an interaction effect is added, and the p-value for the LRT of the interaction effect. The interaction effect indicated as most significant, by having the lowest p-value for the LRT, is the interaction between year group and NACE-section. We therefore add this interaction to our model. We then run a new `add1()` command with our new model, and continue doing so till we have the full model;

```
> M5.nace.m = glm(deaths ~ offset(log(personYears)) +
  age*yearGroup*NaceMain, family=poisson, data = subset for males)
```

including all main effects, all first order interactions and the second order interaction between all included variables. Summaries of all the fitted male models are given in Table 3.3. Comparing deviances through LRT, we see in Table 3.3 that there are clear significant improvements in deviance up to model M3.nace.m. When we check wheter we should add the interaction between age and year group, in model M4.nace.m, however the choice of model is not as clear. The p-value of the LRT is 0.0457, close to the 0.5 threshold and the AIC-score of models M3.nace.m and M4.nace.m are almost the same [2]. The BIC-score do not prefer either of the two as the best model, comparing BIC-scores of the two models however M3.nace.m comes out as the preferred model.

Table 3.3: Deviance table showing model summaries and hypothesis testing of models fitted on male data with a totall of 1349 observations. Main components for variables are A: linear numeric age, N: NACE-section and Y_{group}: three and three years grouped together (as explained in section 3.3). Δ = deviance and p = number of parameters.

model	variables	$2 \cdot$ log likelihood	p	Δ	Null hypothesis	p-value LRT	AIC	BIC
M1.nace.m	A + Y_{group} + N	-2589.2	12	1191.74	-	-	2613	2676
M2.nace.m	A + Y_{group}+ N + Y_{group}:N	-2553.7	28	1156.22	Y_{group}:N=0: $-2\log\frac{L(\text{M1.nace.m})}{L(\text{M2.nace.m})}$	0.0034 **	2610	2756
M3.nace.m	A + Y_{group} + N + Y_{group}:N + A:N	-2543.2	30	1145.70	A:N=0: $-2\log\frac{L(\text{M2.nace.m})}{L(\text{M3.nace.m})}$	0.0052 **	2603	2759
M4.nace.m	A + Y_{group} + N + Y_{group}:N + A:N + A:Y_{group}	-2527.4	38	1129.92	A:Y_{group}=0: $-2\log\frac{L(\text{M3.nace.m})}{L(\text{M4.nace.m})}$	0.0457 *	2603	2801
M5.nace.m	A + Y_{group} + N + Y_{group}:N + A:N + A:Y_{group} + A:Y_{group}:N	-2516.5	54	1119.05	A:Y_{group}:N=0: $-2\log\frac{L(\text{M4.nace.m})}{L(\text{M5.nace.m})}$	0.8174	2625	2906

Predictions over age and year of model M3.nace.m are given in Figure 3.9. Predictions are made for each NACE-section using the `predict.glm()` function in `R`, here using NACE-section C as an example, through;

[2] Rounded of they are the same. Unrounded AICs; M3.nace.m: 2603.185 and M4.nace.m: 2603.409

```
> pred.age.c = predict.glm(model, newdata = data.frame(personYears=1,
  age=20:69, NaceMain="C", yearGroup="1997-1999"), type= "response")
> pred.year.c = predict.glm(model, newdata = data.frame(personYears=1,
  age=50, NaceMain="C", yearGroup=year.groups), type= "response")
```

We fix the year group to $1997 - 1999$ when making predictions over age and fix age at 50 when making predictions over year groups. This is done in order to see the effects of each variable in the models and avoid shifts in curves which may be caused by differences in distribution of age or year within NACE-sections. The default output from `predict.glm()` is predictions on the scale of the linear predictors, we therefore use `type="response"` to get predicted death rates (R Core Team 2017). We could also have gotten the death rates by using `type="link"`, which is the default and then take exp(resulting predictions), as we are working with the Poisson distribution and the default output therefore is log(death rates).

From Figure 3.9 it is clear that a model with none or just one interaction effect whould have been insufficient, to represent the trends observed in Figure 3.7. It looks like our chosen model, M3.nace.m, makes predictions representative for the observations we made earlier. We do not get the shifting patterns from one age group to the next, as we observed in Figure 3.7, because we use a linear numeric age variable. This is most clear comparing the observations and predictions for the younger population of the portfolio, under the age of 40. We will try to solve this in Section 6.6 by using a smoothed version of age.

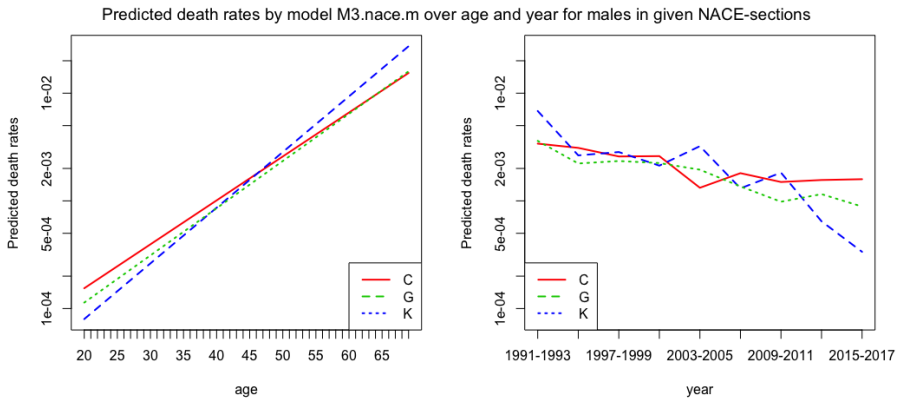


Figure 3.9: Predicted death rates over age (left panel) and year (right panel) of model M3.nace.m. Predictions over age are made with fixed year period: 1997-1999. Predicitons over year are made with fixed age: 50. Each NACE-section is represented by a line in each plot. Red line: NACE-section C - Manufacturing, green line: NACE-section G - Wholesale and retail trade; repair of motor vehicles and motorcycles, blue line: NACE-section K - Financial and Insurance Activities. Predictions are plotted on a log-scale.

Despite the already addressed differences in slope over given ages the prediction curves over age have a recognizable shape. NACE-section K has a steeper

slope than the two other sections and NACE-sections C and G have almost parallel lines over age. Prediction curves over year have a recognizable pattern for all NACE-sections. NACE-section K does however have death rates closer to the two other sections in the predictions, than what we saw in the observations. This may be due to an older polulation in NACE-section K, causing higher death rates in the observations than in the predictions, where we have fixed the age at 50 for all NACE-sections.

Male - Deviance residual diagnostics

As with the models in the previous section we want to check that our chosen model has deviance residuals that look random. Earlier we also checked if the residuals had an approximate normal distribution. We will also look at this now, but we cannot expect approximate normality as we have few deaths for many observations. This is due to the increase in number of groups in the data set and hence we have made each group smaller.

Looking at the deviance residuals of model M3.nace.m over the fitted values, Figure 3.10 a), we have bands of residuals in the plot. This is due to the small number of deaths in the observation. The lowest band of residuals belong to observations with zero deaths, the second band belong to the observations with one death and so on. It is therefore difficult to judge the plot by looking at the residuals alone. The trend line in the same plot however show that the residuals are centred just below zero and have a close to zero slope over the fitted value span, hence there is no clear trend in the residuals. The histogram of our residuals, in Figure 3.10 b), shows that the residuals have a higher peak and are a bit right skewed compared to what they would have been if they were normally distributed. Again, this may be due to a small number of observed deaths in our data groups.

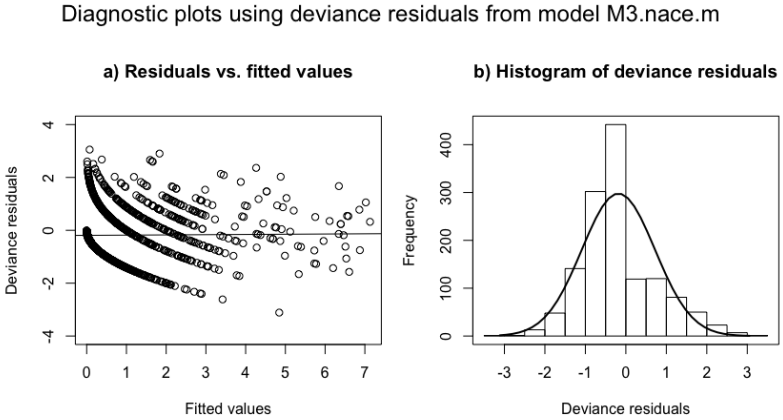


Figure 3.10: a) Deviance residuals of model M3.nace.m plotted over fitted values. Trend line of residuals is given as a black line in the plot. b) Histogram of the deviance residuals of model M3.nace.m with a normal distribution line drawn on top.

We want to check that our residuals do not make patterns over the variables included in the model. In the top panel of Figure 3.11 we check the deviance

residuals over age. We have many residuals and it can therefore be difficult to tell if there is a trend or not based on the residuals alone. The trend line of the residuals is however flat over the age span and centred just below zero, meaning there is no clear trends in the residuals over age. The residuals also look good over year groups and NACE-sections, middle and bottom panel of Figure 3.11. The residuals are centred around zero and have no clear patterns over either of the categorical variables.

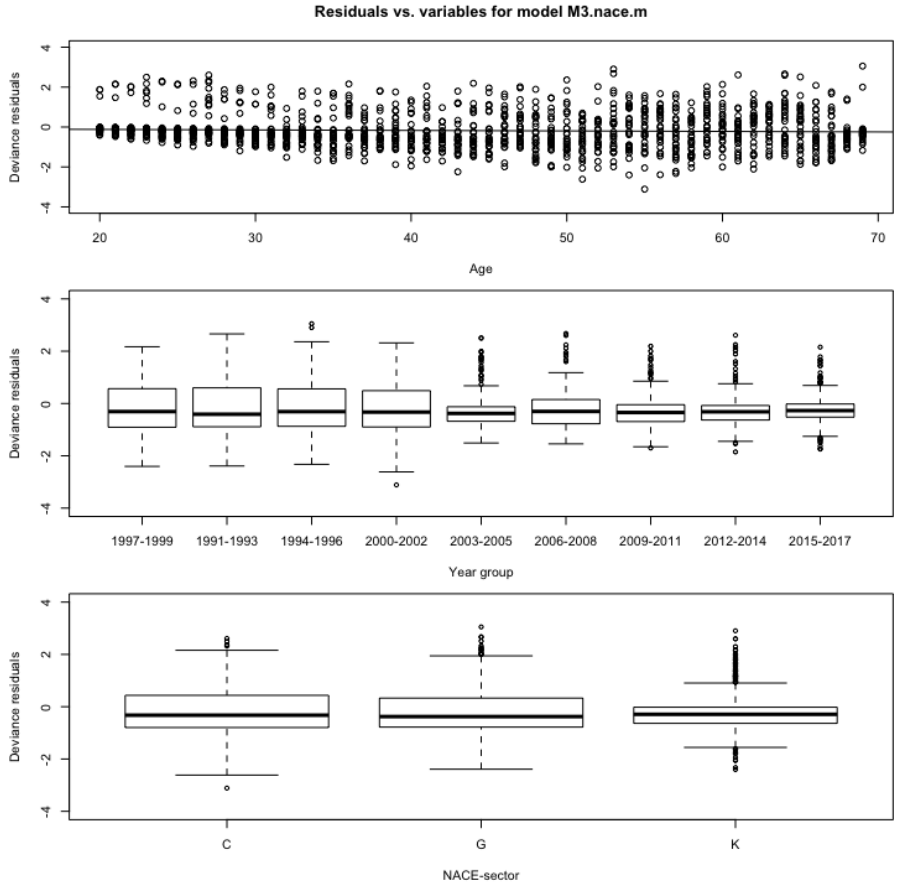


Figure 3.11: Residual diagnostic plots of model M3.nace.m, deviance residuals are plotted against variables used in the model. Top panel: residuals over age with a trend line drawn on top. Middle panel: residuals over year groups. Bottom panel: residuals over NACE-sections.

Female - Model fitting and model selection

We fit models in the same way as we did for the males, using `add1()`. Trying to add interactions however, none of the interaction variables have effects that are significantly different from zero. Starting with a full model and dropping

variables through `drop1()`, we end up with the same result, a model with main effects only. Table 3.4 show two of the models fitted to the female data set. Here model M2.nace.f is the model with the most significant interaction effect added. It is clear however from the LRT, AIC and BIC that the model with main effects only is the preferred model.

Table 3.4: Deviance table showing model summaries and hypothesis testing of models fitted on female data with a total of 1347 observations. Main components for variables are A: linear numeric age, N: NACE-section and Y_{group}: three and three years grouped together (as explained in section 3.3). Δ = deviance and p = number of parameters.

model	variables	2 · log likelihood	p	Δ	Null hypothesis	p-value LRT	AIC	BIC
M1.nace.f	A + Y_{group} + N	-1190.0	12	716.53	-	-	1214	1276
M2.nace.f	A + Y_{group}+ N + A:N	-1187.8	14	714.39	A:N=0: $-2\log\frac{L(\text{M1.nace.f})}{L(\text{M2.nace.f})}$	0.34	1216	1289

Death rate predictions over age and year by model M1.nace.f, from Table 3.4, are given in Figure 3.12. Prediction plots are made in the same way as the male predictions, by fixing year group to $1997-1999$ when making predictions over age and fixing age at 50 when making predictions over year groups. We here recognize some of the trends we saw earlier in our observations. NACE-sections C and G have almost identical death ratepredictions over both age and year groups.

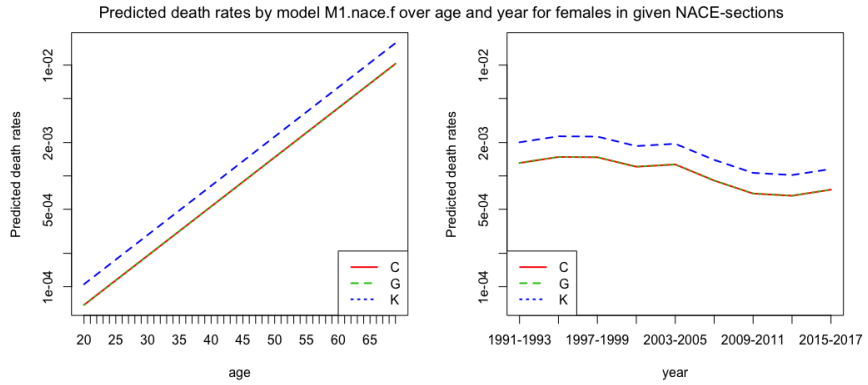


Figure 3.12: Predicted death rates over age (left panel) and year (right panel) of model M1.nace.f. Predictions over age are made with fixed year period: 1997-1999. Predicitons over year are made with fixed age: 50. Each NACE-section is represented by a line in each plot. Red line: NACE-section C - Manufacturing, green line: NACE-section G - Wholesale and retail trade; repair of motor vehicles and motorcycles, blue line: NACE-section K - Financial and Insurance Activities. Predictions are plotted on a log-scale.

The trend we saw in the observations, Figure 3.8, with lower death rates for the two youngest age groups in NACE-section G are gone in the predictions of the model. As in the male case, this is most likely due to the fact that we have the same slope for all ages, not using a categorical variable, but a linear

numeric one. It may also be due to differences in time insured, as we see a tendency of decreasing death rates over time. It may also be due to random variations, which is highly likely, as we have a small number of observations for the younger female population and the interaction effect bewteen NACE-section and age was insignificantly different from zero. NACE-section K have parallel death rate predictions to the two other NACE-sections over both age- and year, but with higher death rates than the other two. We get the parallel lines as we do not have interaction effects in our model. In all NACE-sections death rates increase with age, and decrease with years.

Female - Deviance residual diagnostics

In the female data set we have even fewer death observations in each group than what we had for the males. We will therefore most likely have residuals that are further away from having a normal distribution than the male deviance residuals. The histogram of the deviance residuals, Figure 3.14 a), show that this is the case. The residuals have a much higher peak than that of a normal distribution, are right skewed and have a heavier upper tail. In Figure 3.14 a) we check for randomness in the residuals. As in the male case, we get bands of residuals in the plot, due to a small numbers of deaths. The trend line is however flat and the residuals are centred just below zero, there is therefore no trends of concern.

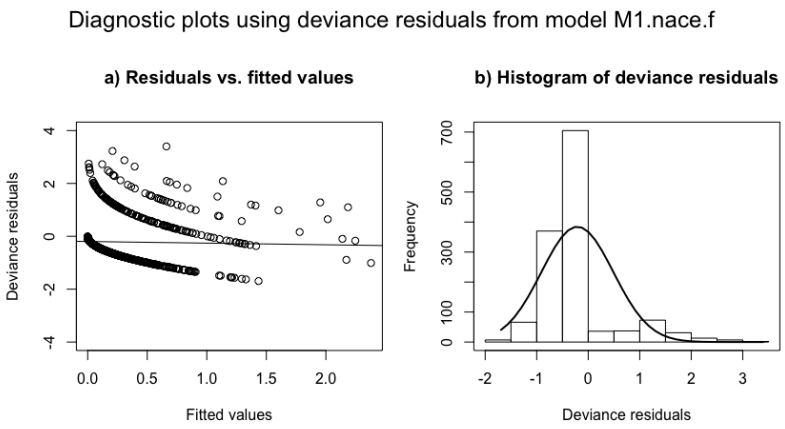


Figure 3.13: a) Deviance residuals of model M1.nace.f plotted over fitted values. Trendline of residuals is given as a black line in the plot. b) Histogram of the deviance residuals of model M1.nace.f with a normal distribution line drawn on top.

The deviance residuals plotted against variables used in model M1.nace.f, Figure 3.14, show no patterns of concern. The residuals are centred just below zero for all variables, and the trend line over age is flat.

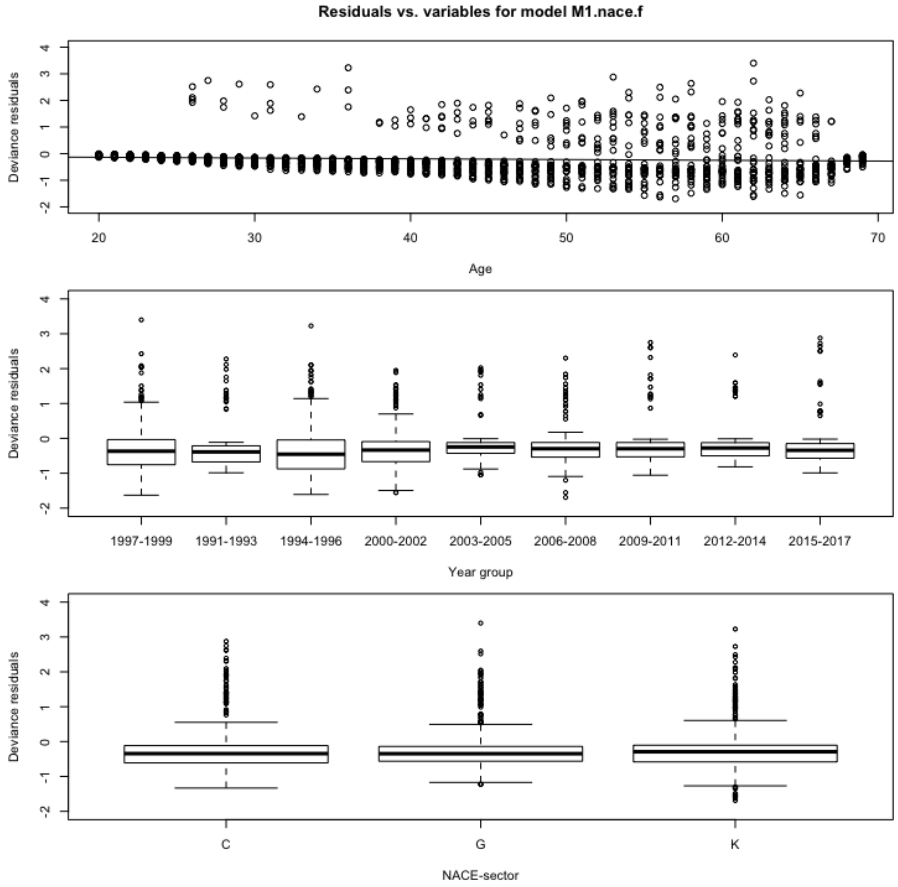


Figure 3.14: Residual diagnostic plots of model M1.nace.f, deviance residuals are plotted against variables used in the model. Top panel: residuals over age with a trend line drawn on top. Middle panel: residuals over year groups. Bottom panel: residuals over NACE-sections.

CHAPTER 4

Alternative model distributions

In the previous chapter, we assumed that death counts are Poisson distributed. Assuming this, we also assume that the variance in death counts is the same as the mean death count when the number of person years is the same for a given group of people. In practice however, this is not always the case. Death counts may have greater variation than that predicted by the Poisson distribution, i.e, we may have overdispersion. Our estimates made by the Poisson models are still consistent, but the standard errors of the estimates get too small (Agresti 2015, p. 248). We will now discuss how we can check if we have overdispersed data and look at two alternative models to the Poisson model which takes overdispersion into account.

4.1 Poisson dispersion test

Overdispersion commonly occur in two different ways (Zuur et al. 2009, p. 224). Firstly, it may occur due to the fact that we have left out some important explanatory variables or that the variables causing variation are unmeasurable. Secondly the observations may be correlated or clustered. We may also actually have real overdispersion. In this section we test for overdispersion by doing dispersion tests on the original data.

We let o_{ijk} be the number of non-aggregated data observations with age i = 20-29,...,60-69, with gender: j = 0,1, insured in year: k = 1991,...,2017. Then, given counts of death $x_1, ..., x_{o_{ijk}}$ we consider testing the null hypothesis that these counts are Poisson distributed with common μ_{ijk} versus the alternative hypothesis that they have different rates $\mu_{ijk_1}, ..., \mu_{ijk_{o_{ijk}}}$. The Poisson dispersion test is defined as (Rice 2007, p. 348);

$$D = \sum_{o=1}^{o_{ijk}} \frac{(x_{ijko} - \overline{x_{ijk}})^2}{\overline{x_{ijk}}} \qquad \text{for } i = \text{20-29,...,60-69, } j = \text{0,1 and } k = \text{1991,...2017.} \tag{4.1}$$

where $\overline{x_{ijk}}$ is the sample mean. D has an approximate chi-squared distribution with $df = o_{ijk} - 1$ degrees of freedom under the null hypothesis (Rice 2007, p. 349). We reject this hypothesis if $D > \chi^2_{o_{ijk}-1,0.95}$.

To test the hypothesis on our data, we have chosen to look at non-aggregated data within given age ranges in the year 1998. We have chosen 1998 as this year lack few exposures[1] and is one of the years with the highest number of

[1] In the non-aggregated data we do not have exposures for all possible variable combinations.

deaths. As the number of person years for each observation varies, we split the test in such a way that person years are approximately equal, as seen in Table 4.1. The results of the tests in each exposure range, for males of age 50-59 can be found in the samle table.

Table 4.1: Summary table of Poisson Dispersion Test done on males in the age group 50-59 in year 1998. D is the test statistic defined in equation (4.1). df is the degrees of freedom for each test statistic.

Exposure	mean	variance	D	df	p-value
Person years = 1	0.0012	0.0012	2528	2530	0.508
1< Person years <2	0.0187	0.0237	475	374	0.000 ***
Person years = 2	0.0029	0.0058	1370	685	0.000 ***
2< Person years $\leq$3	0.0046	0.0046	431	432	0.505
3< Person years $\leq$4	0.0087	0.0087	228	229	0.506
4< Person years $\leq$5	0.0157	0.0156	125	126	0.508
5< Person years $\leq$6	0.0345	0.0337	84	86	0.541

In most of the exposure groups in Table 4.1, there is no clear evidence of overdispersion. In two of the exposure groups however, the Poisson Dispersion Test (PDT) is marked as significant. We have performed the PDT for all age groups in both the male and female population, again in the year 1998. We have 5 age groups and 7 exposure groups, which leaves us with a total of 70 age and exposure groups, 35 for each gender. We do however end up with a total of 43 tests to use, as we have age-exposure combinations with no observed deaths[2]. Only considering PDT's with observed deaths, a total of 3 out of 43 test are marked as significant (in addition to the two significant PDT's shown in Table 4.1; males of age 20-29 in exposure group; Person years = 1), all tests are attached in Appendix B, page 103 and 104. To sum up the tests carried out, there is no clear evidence of overdispersion in the data for the majority of our observations, at least not in the year 1998. We could however have overdispersion for some of our data or for other years. We will therefore use two more methods to find out if the data is overdispersed or not.

4.2 The Quasi-Likelihood method

We group the data according to age groups: i = 20-29,...,60-69, gender: j = 0,1, year: k = 1991,...,2017 and NACE-sections: l = A,B,...,S,U,X. Why will be explained later in this section. Grouping data in this way should leave us with a total of 2835 observations for each gender (5 age groups x 27 years x 21 NACE-sections). We do however not have exposures for each variable combination, yielding 2729 observations in the male data set and 2691 observations in the female data set[3].

We do not give a full overview of which exposures are missing here. When exposures are missing and models are fitted to given data however, an overview of missing exposures have been attached. Throughout we treat missing exposures as missing completely at random.

[2] 7 group combinations with no deaths in the male data and 20 group combinations with no deaths in the female data. See Appendix B, page 103 and 104 to see which group combinations lack death observations.

[3] Overview of missing data can be found in Appendix B, page 101 - 102

With data grouped as explained above, the mean - variance relation of each exponential family response can be expressed as $var(X_{ijkl}) = \phi \cdot V(\mu_{ijkl})$(Jong and Heller 2008, p. 94). For a Poisson response we have $\phi = 1$ and $V(\mu_{ijkl}) = \mu_{ijkl}$. If $\phi > 1$ however, which is the case when we have overdispersion, the function for $var(X_{ijkl})$ above does not correspond to a Poisson distribution, nor an exponential family response (Jong and Heller 2008, p. 94). As a consequence the likelihood and therefore the maximum likelihood estimates cannot be found by using the methods of the previous chapter, as we no longer have an actual probability distribution to base the calculations on. As a solution we instead maximize the quasi-likelihood (QL).

Instead of assuming a particular distribution, the QL only assume a mean-variance relation for the distribution of X_{ijkl} (Agresti 2015, p. 268). In the next subsection, Jong and Heller (2008, Section 6.3) is used as reference if not else is specified.

Parameter estimation

For a model with link function $g(\mu_{ijkl}) = \eta_{ijkl}$, the quasi-likelihood estimate $\widehat{\boldsymbol{\beta}}$ is found in the same way as for an ordinary GLM, by solving the following equation with respect to $\boldsymbol{\beta}$:

$$u(\beta) = \sum_{ijkl} \left\{ \left(\frac{d\mu_{ijkl}}{d\boldsymbol{\beta}} \right)^T \frac{(x_{ijkl} - \mu_{ijkl})}{V^*(\mu_{ijkl})} \right\} = 0 \tag{4.2}$$

The prior definition of $V(\mu_{ijkl})$ is here replaced by $V^*(\mu_{ijkl}) = \phi \cdot V(\mu_{ijkl})$. We get the same estimates as in an ordinary GLM, as the dispersion parameter ϕ drops out of the equation. At the same time we take the empirical variablility into account as the the standard errors will be inflated by including ϕ.

For a Poisson model, we get the middle part of equation (4.2) by using the definition of μ_{ijkl} given in equation (3.1) and taking the derivative of $x_{ijkl} \log(\mu_{ijkl}) - \mu_{ijkl}$ with respect to β. The quasi-likelihood function for a Quasi-Poisson model is therefore defined as;

$$Q(\beta) = \sum_{ijkl} \{x_{ijkl} \log(\mu_{ijkl}) - \mu_{ijkl}\} \tag{4.3}$$

Maximized with respect to β, through equation (4.2), the QL gives identical regression parameter estimates to those from the usual Poisson regression. The standard errors will however be different as they are multiplied by a factor of $\sqrt{\phi}$.

We will be fitting Quasi-Poisson models in R, which estimates the size of the dispersion parameter by using the Pearson residuals. The Pearson residuals are the raw residuals, $x_o - \widehat{\mu}_o$, devided by the the estimated standard deviation of X, $\sqrt{V(\widehat{\mu}_{ijkl})}$ (McCullagh and Nelder 1989, p. 37). Remembering what we have discussed earlier in this section we have $V(\widehat{\mu}_{ijkl}) = \widehat{\mu}_{ijkl}$ for a Poisson response. Our Pearson residuals can therefore be defined as:

$$r_{per_o} = \frac{x_o - \widehat{\mu}_o}{\sqrt{\widehat{\mu}}}$$

With the Pearson residuals in place, the size of the dispersion parameter, ϕ, is estimated by (Zuur et al. 2009, p. 233);

$$\phi = \sum_{o=1}^{O} \frac{r_{per_o}^2}{O - p}$$

where O equal the total number of observations, p is the number of regression parameters and r_{per} is the Pearson residuals. The quasi-likelihood estimation is the same as maximum likelihood estimation when $\phi = 1$.

Comparing Poisson models with Quasi-Poisson models

The fitted estimates from a Quasi-Poisson model is the same as for a Poisson model. What varies is the standard errors of the estimates. We fit Poisson and Quasi-Poisson models to our male and female data, using the same variables as in Section 3.3. This time however, we fit the models on data sets aggregated as explained earlier in this section, so that we have NACE-section as an additional variable in the data sets, which is not included in the modelling.

We do this as it makes it easier to see variations and differences, both within and between groups in the data sets. It leads to more observations within each group of gender-age-year combination. The total number of observations is, as mentioned earlier, 2729 for males and 2691 for females. The combination classes of covariates used in the modelling, age and year is however much less (135: 5 age groups x 27 years). Hence, the μ_{ijkl}s will not depend on the NACE-sections (l)[4].

As we know from earlier in this section, the Quasi-likelihood equals the likelihood when the dispersion parameter $\phi = 1$. If the parameter is close to one there is no evidence of overdispersion. If the parameter is much greater than 1 however, we have evidence of overdispersion. If the parameter is much less than 1, we have clustering (Zuur et al. 2009, p. 225). For our fitted Quasi-Poisson models we get the following ϕ values:

qpM0$_m$: $\phi = 1.006$
qpM1$_m$: $\phi = 1.095$
qpM0$_f$: $\phi = 1.150$
qpM1$_f$: $\phi = 1.396$

The model names correspond to variables included in the Poisson models in Table 4.2, page 39. Models names with a m at the end are fitted on male data, model names with a f at the end are fitted on female data. We get somewhat different results for the two genders. The tests suggest that the male population is closer to Poisson distributed death counts than the females, neither of the genders is however far off. As a rule of thumb a ϕ larger than 1.5 indicates that something must be done in order to correct for the overdipersion (Zuur et al. 2009, p. 226). Here this is not the case and Poisson distributed models may still be considered adequate.

[4]For each of the 21 NACE-sections this means that $\mu_{ijk1} = \mu_{ijk2} = \ldots = \mu_{ijk21}$. The number of summands in the likelihoods and quasi-likelihoods, will depend on the NACE-sections as they increase the number of observations.

4.3 The Negative Binomial distribution

There are a few different ways of defining the Negative Binomial (NB) distribution. We will use one of the most common definitions, where we make a NB model as a Gamma mixture of Poisson.

Specification of model set-up

We use the same grouping of data as in Section 4.2, with two data sets, one for each gender (j), with age group (i), year (k) and NACE-section (l) as available variables. Given that mean $M_{ijkl} = m$, we assume that X_{ijkl} is Poisson distributed. In addition we regard M_{ijkl} as a gamma distributed countinous random variable with density function (Jong and Heller 2008, p. 28);

$$g(m) = \frac{m^{-1}}{\Gamma(v)}(\frac{m \cdot v}{\mu_{ijkl}})^v \exp(\frac{-m \cdot v}{\mu_{ijkl}}) \tag{4.4}$$

where $g(m) = 0$ for $m < 0$ and v is a shape parameter. The unconditional probability function of X_{ijkl} is then defined as (Jong and Heller 2008, p. 32);

$$P(X_{ijkl} = x_{ijkl}) = \int_0^\infty P(X_{ijkl} = x_{ijkl} \mid M_{ijkl} = m) \cdot g(m) dm$$

with $P(X_{ijkl} = x_{ijkl} \mid M_{ijkl} = m)$ equal to $P(X_{ijkl} = x_{ijkl})$ in equation (3.3) when grouping of data is the same as in Section 3.1. Solving this integral and substituting $\kappa = 1/v$ yields the Negative Binomial pdf (Jong and Heller 2008, pp. 32–33);

$$f(x_{ijkl}) = \frac{\Gamma(x_{ijkl} + 1/\kappa)}{x_{ijkl}!\Gamma(1/\kappa)} \left(\frac{1}{1 + \kappa\mu_{ijkl}}\right)^{1/\kappa} \left(\frac{\kappa\mu_{ijkl}}{1 + \kappa\mu_{ijkl}}\right)^{x_{ijkl}} \tag{4.5}$$

with $E(X_{ijkl}) = \mu_{ijkl}$ and $var(X_{ijkl}) = \mu_{ijkl}(1 + \kappa\mu_{ijkl})$ (Jong and Heller 2008, p. 25). We get a more flexible alternative to the Poisson distribution. κ is a dispersion parameter, controlling the deviation from the Poisson distribution by adjusting the variance independently from the mean. As $\kappa \to 0$, $f(x_{ijkl}) \to$ Poisson, in other words, the Poisson distribution is a special case of the NB distribution where $\kappa = 0$ (Agresti 2015, p. 248). A Poisson model is therefore a nested model of a NB model when fitted to the same data. Defining μ_{ijkl}, using the same link function as in equation (3.1);

$$\mu_{ijkl} = E(X_{ijkl}) = \exp\{\log(n_{ijkl}) + \eta_{ijkl}\}$$

and inserting the definition into equation (4.5) yields the Negative Binomial regression model.

Parameter estimation

Estimation of regression parameters β and κ is done in the same way as we did in Section 3.1, by using the maximum likelihood method. The likelihood function of the Negative Binomial pmf can be expressed as;

$$L(x, \mu) = \prod_{ijkl} f(x_{ijkl}) \tag{4.6}$$

with $f(x_{ijkl})$ equal (4.5). As in the Poisson case, the likelihood depends on β through (3.2)[5]. The values of β and κ which maximize the likelihood function are chosen as regression parameters in our models. Standard errors of the estimates are obtained in the same way as explained in Section 3.1.

Comparing Poisson models with Negative Binomial models

Despite dealing with two different distributions, we can compare our Poisson and NB models in the same way as we compared models in Section 3.2, by taking the difference in deviances, and carry out a LRT. We can do this because, as mentioned earlier in this section, Poisson is a special case of NB, and hence the models are nested. What is different for the LRT when comparing Negative Binomial and Poisson models is the distribution of the test statistic. The statistic has mass of 0.5 at zero, and a half-χ^2_1 distribution above zero (Agresti 2015, p. 250). This is because $\kappa \geq 0$, which means that $\kappa = 0$ is at the boundary of the parameter space (Jong and Heller 2008, p. 91). When carrying out the test, testing if $\kappa = 0$, we must therefore half the p-value we would get from the usual LRT. The null-hypothesis is that the Poisson model holds, if we get a p-value below 0.05 we reject the hypothesis and Poisson models are no longer considered having an adequate fit. This would indicate dispersion in our data.

We fit Negative Binomial models to our male and female data, again with NACE-section as a variable in the data sets, which is not used as a covariate in the models. This, as in the Quasi-Poisson models, means that the value of μ_{ijkl} will be independent of NACE-sections (l)[6]. Comparing Negative Binomial models with Poisson models for our male and female data sets respectively, we get quite different results for the two genders. This time however, we get the opposite suggestion regarding which gender is closest to having Poisson distributed death rates, than what we got in Section 4.2.

In Table 4.2 we see that the Negative Binomial model is preferred by the likelihood ratio test for the male models, with a significant p-value. For the bigger male NB model, nbM0_m, the p-value is not that far of the 0.05 threshold. For det smaller male model, nbM0_m, the p-value is small and clearly in favour of the NB model. For the models fitted on female data however, the Poisson model is preferred in both cases and we get κ close to zero in both NB models. We get a split in terms of which distribution is preferred for the two data sets.

4.4 Conclusion on distribution

Even though some of the results above may indicate mild overdispersion compared to Poisson, the main picture from the PDTs and the alternative models based on quasi-likelihood the Negative Binomial distribution, is that there is no clear sign of overdispersion. We therefore conclude that it is all right to proceed with the assumption that the number of deaths are Poisson distributed. One reason why there are no clear signs of overdispersion in our data may be that there is a change in the companies insured form year to year. The tests

[5] given that grouping of data is the same as in Chapter 3.

[6] The number of observations used to calculate the likelihood, (4.6), will depend on the NACE-sections as they increase the total number of observations.

Table 4.2: Deviance tables showing model summaries and hypothesis testing of Poisson- and Negative Binomial models. Main terms for variables are A_{group}: age group and Y: year. Y_{group} is a categorical year variable where three and three years are merged together. Δ = deviance and p = number of parameters. κ is the dispersion parameter described in Section 4.3. Male models are fitted on a data set with 2729 observarions. Female models are fitted on a data set with 2691 observarions.

Male models							
model	variables	$2 \cdot$ log likelihood	Δ	p	κ	Null hypothesis	p-value LRT
$pM0_m$	A_{group}+Y + A_{group}:Y	-4049.4	1727.94	135	-	-	-
$nbM0_m$	A_{group}+Y + A_{group}:Y	-4046.1	1665.67	136	0.0218	$\kappa = 0 : -2 \log \frac{L(M0_m)}{L(nbM0_m)}$	0.034 *
$pM1_m$	A_{group}+Y+ A_{group}:Y_{group}	-4133.2	1811.71	63	-	-	-
$nbM1_m$	A_{group}+Y+ A_{group}:Y_{group}	-4124.5	1707.83	64	0.0369	$\kappa = 0 : -2 \log \frac{L(M1_m)}{L(nbM1_m)}$	0.002 ***

Female models							
model	variables	$2 \cdot$ log likelihood	Δ	p	κ	Null hypothesis	p-value LRT
$pM0_f$	A_{group}+Y + A_{group}:Y	-2052.0	1093.81	135	-	-	-
$nbM0_f$	A_{group}+Y + A_{group}:Y	-2052.2	1092.73	136	0.0018	$\kappa = 0 : -2 \log \frac{L(M0_f)}{L(nbM0_f)}$	0.50
$pM1_f$	A_{group}+Y+ A_{group}:Y_{group}	-2148.3	1190.07	63	-	-	-
$nbM1_f$	A_{group}+Y+ A_{group}:Y_{group}	-2148.5	1187.16	64	0.0044	$\kappa = 0 : -2 \log \frac{L(M1_f)}{L(nbM1_f)}$	0.50

carried out in this chapter may have looked different if we had the same group of people being observed over longer time spans.

CHAPTER 5

Smoothing nonlinear relations

So far we have fitted models to our data using Generalized Linear Models (GLMs). But patterns may not always be linear, or one good linear fit for the whole data set may be hard to find. We now want to look at models that assume less about the effects the covariates have on the μ_i's than the parametric models do. One way of doing this is by fitting Generalized Additive Models (GAMs) of the form:

$$\mu(z_1, z_2) = \exp\{\alpha + f_1(z_1) + f_2(z_2)\} \tag{5.1}$$

when vi have two numeric covariates like age and year. These types of models are based on smoothing with splines f_j, α is as earlier an intercept. Before we look deeper into how we can use splines through GAMs on our data, in Chapter 6, we will use this chapter to explain what splines are and how smoothing can be done for numeric observations with one covariate.

5.1 Splines

A linear spline is a continuous function within a chosen interval, formed by linear segments which are connected (Hastie, Tibshirani, and Friedman 2009, p.141). Each point where the segments connect is called a knot of the spline. We have made an example of this in Figure 5.1 a). Here a linear spline is fitted to constructed data[1] within the interval $z \in [0, 6]$. This spline has two inner knots at $z = 3$, $z = 4$ and two boundary knots at $z = 0$ and $z = 6$. The segments between knots can also be polynomials (Hastie, Tibshirani, and Friedman 2009, p.141), as in Figure 5.1 b). Here the fitted splines have, as the first spline, knots at $z = 0, 3, 4, 6$. A spline made up of segments that are quadratic polynomials, as the green line in Figure 5.1 b), is said to have order 3. This means that it is made up of segments that are polynomials of degree 2, that it has 1 derivative, and it is continuous at the knots. A spline made up of segments that are cubic polynomials, as the red line in Figure 5.1 b) is said to have order 4. This means that it is made up of segments that are polynomials of degree 3, and that it has 2 derivatives, ensuring continuity. A spline of order 1 corresponds to using step functions.

In general we say that a spline has order K. This means that the spline is formed by connecting polynomial segments of degree $K - 1$, that it has $K - 2$

[1] Code used du construct data can be found in Appendix A, p.93

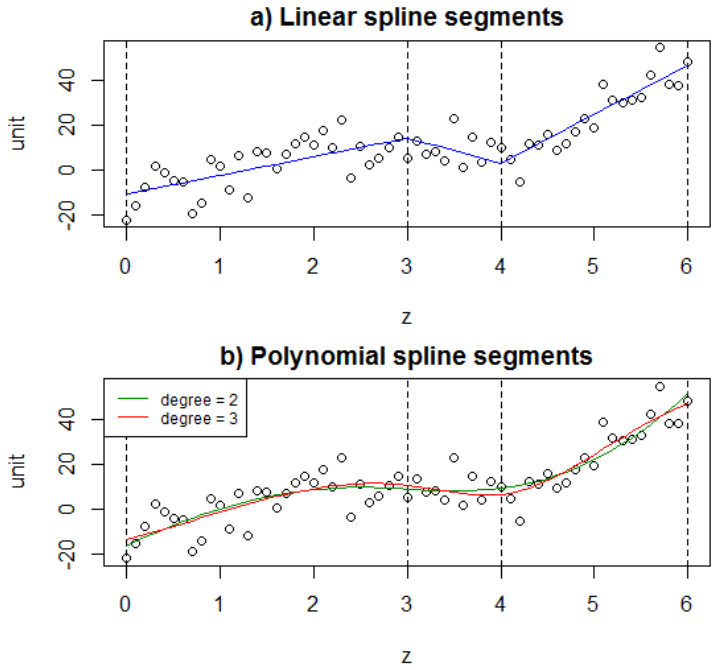


Figure 5.1: Example of splines fitted on constructed data plotted as points. a) Linear spline segments drawn as a blue line with two inner knots at z=3 and z=4. b) Two polynomial spline segments with inner knots at z=3 and z=4. Two polyniomials of different degrees are fitted to the data. Green line = quadratic polynomial, red line = cubic polynomial. Both figures have boundary knots at z=0 and z=6 marked by dotted vertical lines.

derivatives at the knots and it is continous (Hastie, Tibshirani, and Friedman 2009, p.144). By carefully choosing the number of knots, and their location we can control the shape of the spline. We can allow for flexibility when trends change quick and avoid overfitting when the trends have little change. The challenge is to choose the location and number of knots in such way that we manage to balance smoothing and overfitting.

We divide the domain of our covariate z into contiguous intervals, splitted by the knots, and represent $f(z)$ by a separate polynomial in each interval. This way we obtain a piecewise polynomial function (Hastie, Tibshirani, and Friedman 2009, p.139);

$$f(z) = \sum_{m=1}^{M} \beta_m B_m(z) \tag{5.2}$$

where M is the number of basis functions in a given type of spline. The number of knots limits the number of basis functions, M. We will now look into an example using B-splines.

5.2 B-splines example

To define B-splines we have, if not else is specified, used Hastie, Tibshirani, and Friedman (2009, pp. 186–187) as reference. B-splines are defined according to

given knots $\xi_1 \leq \xi_2 \leq \ldots \leq \xi_{N+1} \leq \xi_{N+2}$ with inner knots at $z = \xi_2, \ldots, z = \xi_{N+1}$ and boundary knots at $z = \xi_1$ and $z = \xi_{N+2}$. Out of these we make an augmented knot sequence $\tau_1 \leq \tau_2 \leq \ldots \leq \tau_{N+2K}$. Here N equals the number of inner knots and K equals the order which we want our final splines to have. The augmented knot sequence is made by repeating the boundary knots ξ_1 and ξ_{N+2} K times each, and then shift the starting point to the first value of ξ_1. This is done in order to have enough evaluation points for the B-splines, up to the desired final order K. The B-splines are defined recursively for every j, first defined by the j'th B-spline of order 1, in the following way;

$$B_{j,1}(z) = \begin{cases} 1 & \text{if } \tau_j \leq z < \tau_{j+1} \\ 0 & \text{else} \end{cases} \tag{5.3}$$

where z is a given input variable. If $\tau_j \leq z = \tau_{j+1}$ then $B_{j,1} = 0$ and if $\tau_j \leq z = \tau_{j+1} = \cdots = \tau_{j+m}$ then $B_{j,m} = 0$. For $1 < k \leq K$;

$$B_{j,k}(z) = W_{j,k}(z)B_{j,k-1}(z) + (1 - W_{j+1,k}(z))B_{j+1,k-1}(z) \tag{5.4}$$

where;

$$W_{j,k}(z) = \begin{cases} (z - \tau_j)/(\tau_{j+k-1} - \tau_j) & \text{if } \tau_j < \tau_{j+k-1} \\ 0 & \text{else} \end{cases} \tag{5.5}$$

We call $B_{j,k}(z)$ the j'th B-spline of order k and it is a polynomial of degree $k-1$. An example showing how B-splines are calculated can be found in Appendix A, p.93. We there calculate B-splines that are quadriatic polynomials with knots in $z = 0, 3, 4, 6$, which is the same as the basis of the fit for the green line in Figure 5.1 b). We will now look at how a function fit, using B-splines as basis functions, change when the number of knots used in the B-spline basis differs.

Using the definition of least squares and the piecewise polynomial definition (5.2), with B-splines as basis functions for a given knot sequence, we can fit a function over z for our constructed data[2]. We assume that all z_i values are unique due to the way the data observations in this thesis are organized and we know this is the case for our constructed data. We will from now on only consider B-splines of order 4, hence cubic polynomials, as we will build further on to these later in this thesis. The least squares are defined as (Hastie, Tibshirani, and Friedman 2009, p.30);

$$RSS(\beta) = \sum_{i=1}^{I} \{y_i - f(z_i)\}^2 = \sum_{i=1}^{I} \{y_i - \sum_{m=1}^{M} \beta_m B_m(z_i)\}^2 \tag{5.6}$$

where B_m is the m'th B-spline of order 4, I is the number of observations and M is the total number of B-spline basis functions. The number of basis functions for cubic B-splines is given by $M = K + N$, where K is the desired order of our splines and N is the number of inner knots. We can write (5.6) in matrix form as (Hastie, Tibshirani, and Friedman 2009, p. 45):

$$RSS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{B\beta})^T(\boldsymbol{y} - \boldsymbol{B\beta}) = \boldsymbol{y^T y} - 2\boldsymbol{\beta^T B^T y} + \boldsymbol{\beta^T B^T B\beta} \tag{5.7}$$

Here $\boldsymbol{B}$ is an $I \times M$ matrix of M B-spline basis functions evaluated at I values of z, for $M < I$. Hence is $\{\boldsymbol{B}\}_{im} = B_m(z_i)$. We then find the least square

[2]see Appendix A, p.93 for code used to construct data. This is the same data as mentioned earlier in this chapter. When mentioning constructed data, this is the data we refer to.

estimate $\widehat{\boldsymbol{\beta}}$ by differentiating (5.7) with respect to $\boldsymbol{\beta}$ and setting the resulting derivative equal to zero:

$$\frac{dRSS(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = -2\boldsymbol{B^T y} + 2\boldsymbol{B^T B\beta} \tag{5.8}$$

$$\frac{dRSS(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = 0 \Longrightarrow \widehat{\boldsymbol{\beta}} = \boldsymbol{(B^T B)}^{-1}\boldsymbol{B^T y} \tag{5.9}$$

We get the fitted vector of spline values $\widehat{\boldsymbol{f}}$ as a linear transformation of $\boldsymbol{y}$, using the result found for the least square minimizer $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{f}} = \boldsymbol{B}\widehat{\boldsymbol{\beta}} = \boldsymbol{B(B^T B)}^{-1}\boldsymbol{B^T y} = \boldsymbol{Hy} \tag{5.10}$$

The number of parameters in the fit, hence the number of β_ms, equals the number of basis functions M. This is the same as the rank of $\boldsymbol{H}$, also known as the hat matrix in statistics. As $\boldsymbol{H}$ is a projection matrix its rank is equal to its trace, the sum of the diagonal elements of the matrix (Agresti 2015, p. 35). In other words we have $trace(\boldsymbol{H}) = rank(\boldsymbol{H}) = M$ which equals the number of parameters used to fit a given function.

We fit four different functions to the constructed data with cubic B-spline basis functions using Equations (5.7) to (5.10). The only difference between the fits are the number of knots used in the basis functions. Function fits are given for 1,2,5 and 12 knots respectively. We see in Figure 5.2, that the more knots the basis function has, the more the fitted function fluctuates and follows the data. This is due to the fact that we get more parameters in the fitted function, the more knots we have. From what we have discussed above we know that $K + N = M = trace(\boldsymbol{H}) = rank(\boldsymbol{H})$. In our example it means that the function estimates, which are of order 4, with 1,2,5 and 12 inner knots get 5,6,9 and 16 parameters respectively. Choosing which fit is the preferred is not as clear here as for the fits we previously have looked at, as the functions no longer are nested. We want a function that fit the data well, at the same time as we do not want overfitting. We will discuss methods of dealing with this in Section 5.4.

5.3 Natural B-splines

A variant of polynomial B-splines are natural B-splines, we have used Hastie and Tibshirani (1990, p. 24) as reference to define these. Natural B-splines are, as the B-splines, defined as piecewise polynomials. They are however only defined as polynomials of odd degree. We will only look into natural cubic B-splines, as we will use this later in our thesis. The natural cubic B-splines are cubic splines which use B-splines as basis functions and have two additional constraints beyond the boundary knots: $f'''(z) = f''(z) = 0$. This makes the splines linear outside the boundary knots and less flexible at the boundaries. This is something we want, as polynomial splines have volatile fits near the boundaries of the data which makes predictions unreliable. The constraints in each boundary region reduce the dimension of the space, and hence the number of splines calculated from $M = K + N = 4 + N$, which we had for cubic B-splines, to $M = 2 + N$ for natural cubic B-splines (Hastie and Tibshirani 1990, pp. 24–25). Here N is the number of inner knots. We are not going to

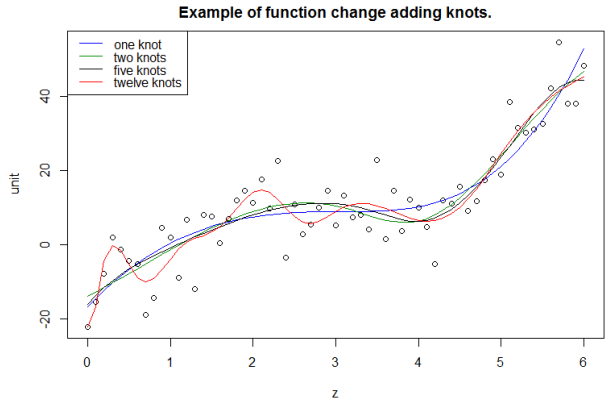


Figure 5.2: Plot of predictions made by function fitted to constructed data. All functions have B-splines of order 4 and boundary knots at $z = 0$ and $z = 6$. What differs between functions is the number of inner knots. Each fitted function is represented by a line. Blue: function with one inner knot in $z = 3$. Green: function with two inner knots in $z = 3, 4$. Black: function with five inner knots in $z = 1, 3, 3.5, 4, 5$. Red: function with twelve inner knots in $z = 0.1, 0.6, 1.1, \ldots, 5.1, 5.6$. Points in plot, the constructed data, is the same as in Figure 5.1.

look at the specific functions used to calculate natural cubic B-splines. We will however look at an example showing what the differences between cubic B-splines and natural cubic B-splines may look like.

We calculate cubic B-splines (BS) and natural cubic B-splines (NS), with 12 inner knots each, using the following functions in R:

```
> z = seq(-2,8,0.1)
> BS = bs(z,knots=seq(0.1,5.9,0.5),intercept=T, Boundary.knots=c(0,6))
> NS = ns(z,knots=seq(0.1,5.9,0.5),intercept=T, Boundary.knots=c(0,6))
```

This gives us a $[101 \times 16]$ BS matrix and a $[101 \times 14]$ NS matrix, with a column for each spline and a row for each value z. We then plot each column of the two matrices over $z \in [-2, 8]$ as in Figure 5.3 a) and b). For the cubic B-splines in Figure 5.3 a), we see that we get 4 lines in each boundary region which goes to extreme values, it is these tendencies we want to reduce by using natural cubic B-splines in Figure 5.3 b). We clearly see that the cubic B-splines in Figure 5.3 a) goes to extreme values outside the boundary knots quicker than the natural cubic B-splines in Figure 5.3 b). Hence the natural cubic B-splines are more stable than the cubic B-splines in these regions. Within the boundary, the natural cubic B-splines are the same as the cubic B-splines for all values, except from the first four and last four cubic B-splines near the boundary knots. Hence we get more stable splines outside the boundary knots, without much change within the boundaries. This is also reflected when the fitted functions using BS and NS as basis functions are extrapolated outside the region $z \in [0, 6]$, in Figure 5.4. The curve that use the BS basis drops rapidly in the upper boundary region, whereas the model that use the NS basis keeps a positive linear trend. Within the boundary knots the fitted curves look the

same.

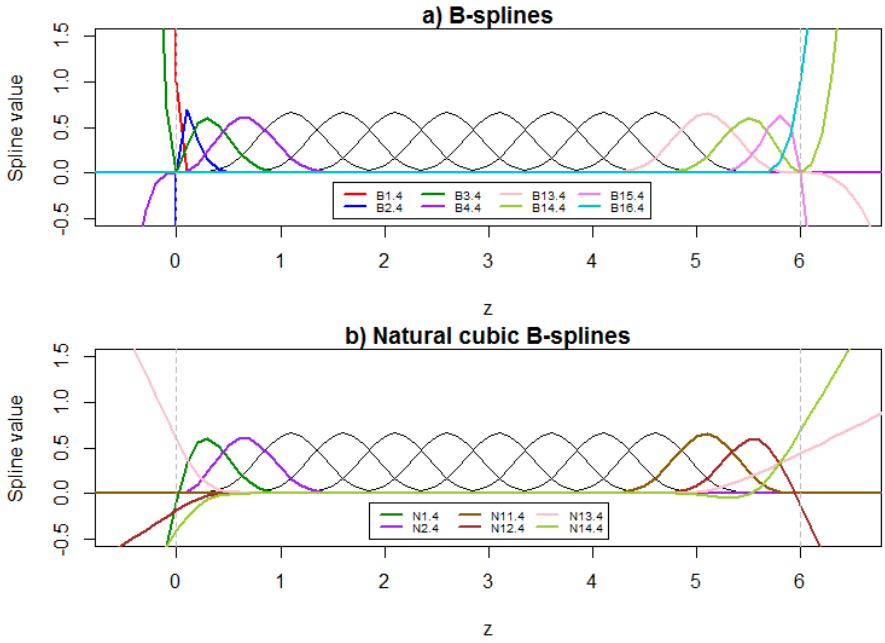


Figure 5.3: Spline exampes with boundary knots in $z = 0$ and $z = 6$ marked by grey vertical lines, and twelve inner knots at $z = 0.1, 0.6, \ldots, 5.1, 5.6$. a) Cubic B-splines plotted over $z \in [-2, 8]$, each spline is represented by a line in the plot. b) Natural cubic B-splines plotted over $z \in [-2, 8]$, each spline is represented by a line in the plot.

5.4 Natural cubic smoothing splines.

A natural cubic smoothing spline is a natural cubic B-spline with a knot in every uniqe value of z_i, which means we end up with $I - 2$ interior knots, where I equals the number of uniqe observations (Hastie and Tibshirani 1990, p. 27). It can be shown, that among all functions $f(z)$ with two continous derivatives, the natural cubic smoothing spline is the unique function which minimize (Hastie and Tibshirani 1990, p. 27):

$$\sum_{i=1}^{I} \{y_i - f(z_i)\}^2 + \lambda \int \{f''(z)\}^2 dz \tag{5.11}$$

Hence it arise as the solution to an optimization problem. We recognize the first part of (5.11) from the definition of least squares given in (5.6). It measures the closeness to the data. The second part of (5.11) penalize curvature in the function for a given constant λ. As with (5.6), we can write (5.11) in matrix form as (Hastie, Tibshirani, and Friedman 2009, p. 152):

$$(\boldsymbol{y} - \boldsymbol{N\beta})^T(\boldsymbol{y} - \boldsymbol{N\beta}) + \lambda \boldsymbol{\beta^T \Omega \beta} \tag{5.12}$$

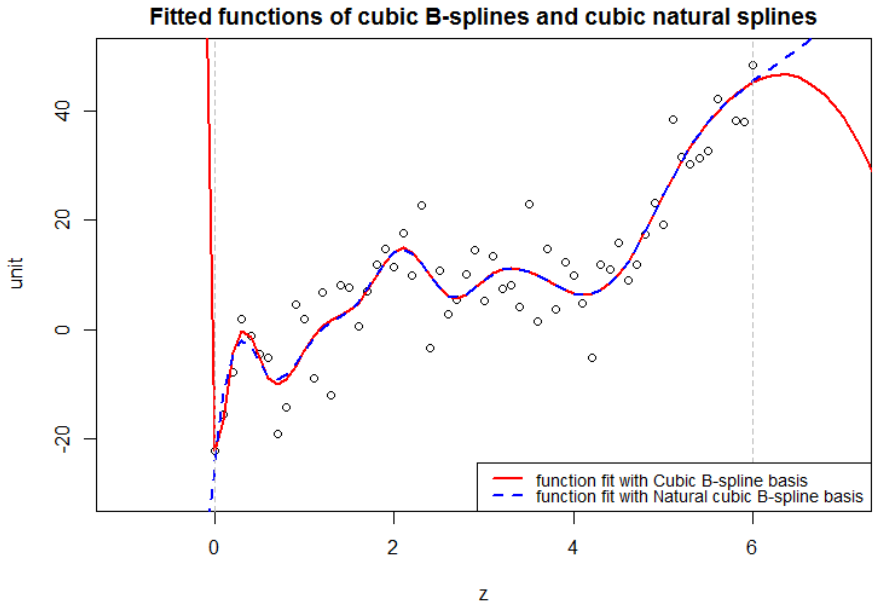


Figure 5.4: Fit of functions on constructed data, using splines from Figure 5.3 as basis functions. The constructed data is shown as points in the plot, over $z \in [-2, 8]$. Each type of spline basis function fit is represented by a line. Red: function with cubic B-spline from Figure 5.3 a) as basis function. Dotted blue: Function with natural cubic B-spline from Figure 5.3 b) as basis function.

The first part of this equation we recognize from (5.7). We have replaced the matrix of B-spline basis functions $\boldsymbol{B}_{[I\times(4+N)]}$ which we had for the cubic B-splines, Section 5.2, with $\boldsymbol{N}$. This is an $[I \times I]$ matrix with columns for each natural spline basis, as we get $M = 2 + N = 2 + (I - 2) = I$ basis functions using a knot in every observation. The second part of 5.12 we get from;

$$\begin{aligned}\int \{f''(z)\}^2 dz &= \int \{\sum_{m=1}^{I} \beta_m N_m''(z)\}^2 dz \\ &= \sum_{m=1}^{I}\sum_{t=1}^{I} \beta_m \beta_t \int N_m''(z) N_t''(z) dz = \boldsymbol{\beta^T \Omega_N \beta} \end{aligned} \tag{5.13}$$

where $\{\boldsymbol{\Omega_N}\}_{mt} = \int N_m''(z) N_t''(z) dz$ (Hastie, Tibshirani, and Friedman 2009, p. 152). The solution to the minimization is found in the same way as in (5.9). By differentiating (5.12) with respect to $\boldsymbol{\beta}$;

$$\frac{d}{d\boldsymbol{\beta}}\{(\boldsymbol{y} - \boldsymbol{N\beta})^T(\boldsymbol{y} - \boldsymbol{N\beta}) + \lambda \boldsymbol{\beta^T \Omega \beta}\} = -2\boldsymbol{N^T y} + 2\boldsymbol{N^T N\beta} + 2\lambda \boldsymbol{\Omega \beta} \tag{5.14}$$

setting the derivative equal to zero and solving for $\boldsymbol{\beta}$ gives:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{N^T N} + \lambda \boldsymbol{\Omega_N})^{-1} \boldsymbol{N^T y}$$

We then get the vector of fitted values through (Hastie, Tibshirani, and Friedman 2009, p. 153):

$$\widehat{\boldsymbol{f}} = \boldsymbol{N}\widehat{\boldsymbol{\beta}} = \boldsymbol{N}(\boldsymbol{N}^{\boldsymbol{T}}\boldsymbol{N} + \lambda\boldsymbol{\Omega}_{\boldsymbol{N}})^{-1}\boldsymbol{N}^{\boldsymbol{T}}\boldsymbol{y} = \boldsymbol{S}_{\boldsymbol{\lambda}}\boldsymbol{y} \tag{5.15}$$

We get the linear operatior $\boldsymbol{S}_{\boldsymbol{\lambda}}$, known as the smoother matrix. It is similar to the hat matrix $\boldsymbol{H}$ in that it is positive semidefinite and symmetric. It is however not indempotent and has rank $I = M = 2 + N$ instead of rank $M = 4 + N$. In Section 5.2 we found the number of parameters in a fit, and hence the degrees of freedom by taking the trace of $\boldsymbol{H}$. By analogy the effective degrees of freedom of a smooting spline is defined as $df_\lambda = trace(\boldsymbol{S}_{\boldsymbol{\lambda}})$ for a given λ (Hastie and Tibshirani 1990, p. 52).

Effects of changing lambda

Till now we have only mentioned λ as a given constant when our smoother matrix is calculated. The smoothing parameter λ establishes a tradeoff in (5.11) between fit to the data and smoothness of the function. We will now look at how different values of λ affects how curves are fitted to given observations through some examples. In R we fit three curves for $\lambda = 0.001, 0.01, 0.1$ respectively to our constructed data through;

```
> smooth.splines(y, lambda = given.value, all.knots = T)
```

and plot the results, as in Figure 5.5. The curves show that the smaller the lambda, the greater the curvature. The smoothing parameter can take any value $\lambda \in (0, \infty)$ (Hastie, Tibshirani, and Friedman 2009, p. 151). As $\lambda \to \infty$ the fit goes to a straight line, as any curvature in the function is endlessly penalized. Looking at the other end of the scale however, where $\lambda \to 0$, we no longer care about curvature and the fitted function can be any function interpolating the data.

As the smoother matrix is calculated for a given value of λ we know that λ will affect the trace of the matrix, and hence the effective number of parameters used to fit a curve. For our fitted curves with $\lambda = 0.001$, $\lambda = 0.01$ and $\lambda = 0.1$, we get $df_{\lambda=0.001} = 6.62$, $df_{\lambda=0.01} = 4.16$ and $df_{\lambda=0.1} = 2.78$. The greater the lambda the smaller the degrees of freedom. We get the same results taking the trace of the resulting $\boldsymbol{S}_{\boldsymbol{\lambda}}$ matrices[3]. As $df_\lambda = trace(\boldsymbol{S}_{\boldsymbol{\lambda}})$ is monotone in λ, the relationship can be inverted, fixing df in the model specifications to specify the amount of smoothing (Hastie, Tibshirani, and Friedman 2009, p. 158). This gives a more uniform way of dealing with comparison of different smoothing methods and is especially useful in generalized additive models, which we will look at in Chapter 6.

Choosing the optimal lambda

We have now seen the effect that different λ values may have on a fit, we do however not know which λ is the best and how we find the optimal λ. Hence, we still want to find the optimal amount of smoothing, as in Section 5.2, but instead of changing the number of knots we put a knot in each uniqe observation

[3] An example showing this can be found in Appendix A, p. 96

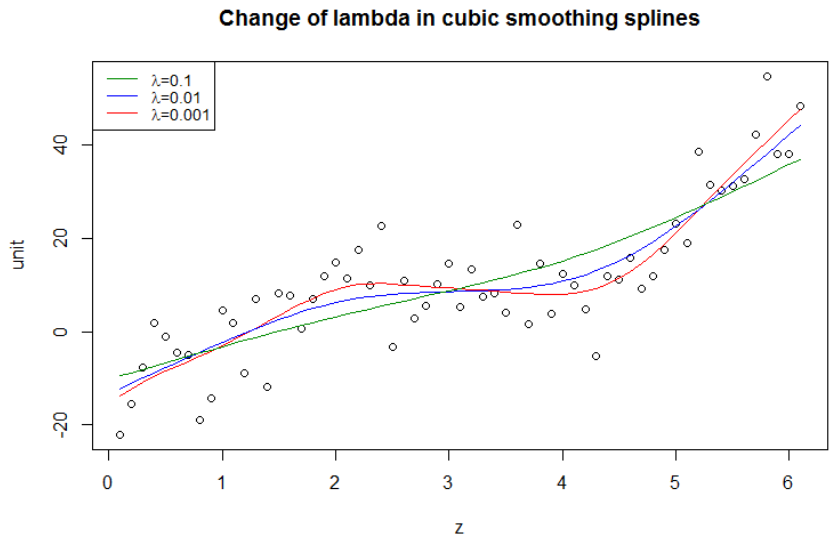


Figure 5.5: Curve fits on constructed data given different values of lambda. Each fit for a given lambda is represented by a line. Green: $\lambda = 0.1$, blue: $\lambda = 0.01$, red: $\lambda = 0.001$.

and try to find the optimal value of λ. To get the best fitted smoothers R uses generalised cross-validation (GCV), but before we go on to explain what this is we look at ordinary cross-validation (OCV)

The OCV-score, or the cross-validation sum of squares, is defined as (Hastie and Tibshirani 1990, p. 43):

$$\mathrm{CV}(\lambda) = \frac{1}{I}\sum_{i=1}^{I}\{Y_i - \hat{f}_\lambda^{[-i]}(z_i)\}^2 \tag{5.16}$$

Here all observations except from observation i are used to calculate the fit, $\hat{f}_\lambda^{[-i]}(z_i)$, at z_i, leaving $I-1$ observations to calculate each fit. The notation $[-i]$ indicates this and mimics the use of test- and training samples for predictions (Hastie and Tibshirani 1990, p. 43). It can be shown that (5.16) equals (Hastie and Tibshirani 1990, p. 48):

$$\mathrm{CV}(\lambda) = \frac{1}{I}\sum_{i=1}^{I}\left\{\frac{y_i - \hat{f}_\lambda(z_i)}{1 - S_{ii}(\lambda)}\right\}^2 \tag{5.17}$$

In this second definition the i'th observation is included when fitting $\hat{f}$. S_{ii} is the i'th diagonal element of the smoothing matrix $\boldsymbol{S_\lambda}$ and $\hat{f}_\lambda(z_i)$ is the optimized fitted value for observation z_i for a given value of λ. We will not go into explaining why (5.16) equals (5.17), but those curious to look deeper into it are referred to Hastie and Tibshirani (1990, pp. 46–48). To find the optimal λ the OCV-score (5.17) is computed for a range of suitable λ-values. The λ that minimizes the value of (5.17) is then chosen. OCV has a tendency of undersmoothing, this tendency is reduced in GCV (Hastie, Tibshirani, and

Friedman 2009, p. 245). GCV is therefore often preferred and is the default function used when fitting smoothing parameters in R.

The GCV approximation of OCV is defined as (Hastie and Tibshirani 1990, p. 49):

$$\text{GCV}(\lambda) = \frac{1}{I}\sum_{i}^{I}\left\{\frac{y_i - \hat{f}_\lambda(z_i)}{1 - \text{trace}(\boldsymbol{S}_{\boldsymbol{\lambda}})/I}\right\}^2 \tag{5.18}$$

The diagonal elements of the smoothing matrix, $S_{ii}(\lambda)$, in (5.17) have in (5.18) been replaced by the mean of the diagonal elements. GCV follows the same procedure as OCV to find the optimal λ value, by calculating a score for a range of different λ-values and then choose the value that minimize the GCV-score. We have shown this procedure over chosen values of λ and the corresponding effective degrees of freedom (edf) for our constructed data in Figure 5.6. Here, going from left to right in both plots, we se that the GCV-score decreases up to certain values of lambda and edf before the trends turn and the GCV-score increase. The minimum GCV-score for the fits we try here is GCV = 56.11. The preferred lambda and corresponding edf are those which are at the minimum value of the GCV-score. In this case this is $\lambda = 0.001517$, corresponding to edf = 6.05, marked as green points in the plots[4].

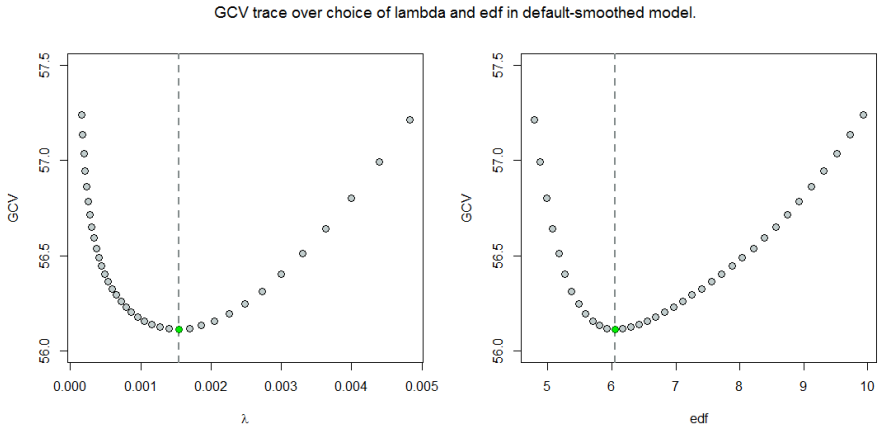


Figure 5.6: GCV-score plotted against λ (left panel) and edf, the effective degrees of freedom (right panel). Minimum value of the GCV trace with corresponding λ and edf values are marked by a vertical dashed line and a green point in both plots.

We let R go through the same procedure as we did above using `smooth.spline(z)`. This gives us edf = 6.06, with corresponding $\lambda = 0.001539$. Fitting a curve through the constructed data, smoothed over z with this default fit yields the fitted line in Figure 5.7.

We have now looked at splines for a numeric response, hence in the context of the classic linear regression domain. In our case however, we have a Poisson distributed response. In the next chapter we will look at how this can be handled with splines.

[4]Code for the plots can be found in Appendix A, p. 98

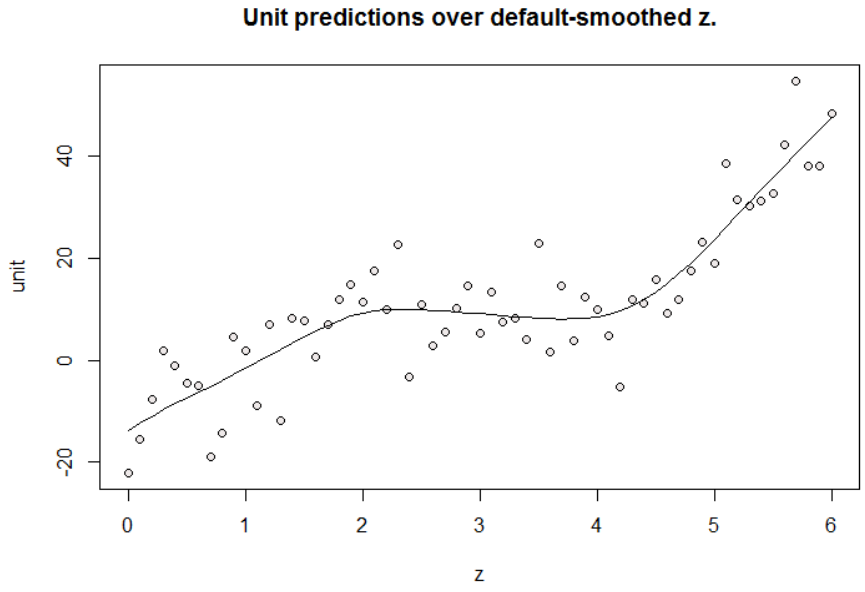


Figure 5.7: Default smoothed fitted function, smoothed over z for our constructed data. The data set has 61 observations, with observations in $z \in [0, 6]$ shown as points in the plot.

CHAPTER 6

Generalized additive modelling

Till now, we have discussed smoothing in a classical regression setting. In the GLM setting, Chapter 3, however we used Poisson regression. Considering our data, we would like to transfer what we have discussed about splines to a Poisson regression domain to see how splines can be used to smooth our numerical variables in Generalised Additive Models (GAMs). GAMs is an extension of GLMs using transformations of the input variables, creating a new space to fit linear models (Hastie, Tibshirani, and Friedman 2009, p. 139). We want to fit functions in the form (5.1), where $f_j(z_j)$ are smoothing splines.

6.1 Model set-up and model optimization

In Poisson regression we do not use the residual sum of squares (5.6) to measure goodness of fit. Instead the log-likelihood is used. A small residuals sum of squares corresponds to a large log-likelihood value. The greater the log-likelihood value the closer is the model fit to the data observations. Considering a case with only one input variable z, the Poisson log-likelihood can be defined as:

$$l(x,\mu) = \sum_{i=1}^{I}\{x_i \log(\mu_i) - \mu_i - \log(x_i!)\} = \sum_{i=1}^{I}\{x_i f(z_i) - \exp\{f(z_i)\} - \log(x_i!)\}$$

This is achieved by taking the log of the likelihood defined in (3.5) and defining $\log\{\mu(z)\} = f(z)$, which implies $\mu(z) = \exp\{f(z)\}$ (Hastie, Tibshirani, and Friedman 2009, p. 161). Hence, μ_i depends on z_i. The penalized log-likelihood corresponding to the penalized RRS, (5.11), is then constructed as (Hastie and Tibshirani 1990, p. 149):

$$l_{\text{pen}}(f,\lambda) = \sum_{i=1}^{I}\{x_i f(z_i) - \exp\{f(z_i)\} - \log(x_i!)\} - \frac{1}{2}\lambda \int \{f''(z)\}^2 dz$$

As with the penalized RSS, it can be shown for a given value of λ that the function f which maximize $l_{\text{pen}}(f,\lambda)$ is a natural cubic spline with knots at each value of z (Hastie, Tibshirani, and Friedman 2009, p. 162). This means that we can represent f, as we did in (5.13), by $f(z) = \sum_{m=1}^{I} \beta_m N_m(z)$.

To optimize λ we use `gam` from the `mgcv` package in `R`. For linear regression with numeric data the GCV-score, explained in Section 5.4, may be calculated in

order to find the optimal λ. For the Poisson distribution however the Unbiased Risk Estimator (UBRE) is calculated for a number of different λ values and the λ minimizing the UBRE-score is chosen. The UBRE-score is in general defined as (Wood 2018d);

$$\text{GCV}_{\text{UBRE}} = \frac{D_\lambda}{I} + \frac{2 \cdot s \cdot \text{edf}_\lambda}{I} - s \tag{6.1}$$

where the number of observations equals I, D is the deviance of the model, edf is the effective degrees of freedom and s is a scale parameter. In Poisson regression $s = 1$ and the deviance is $\sum\{y_i \log(y_i/\mu_i) - (y_i - \mu_i)\}$. The UBRE score is only used when the scale parameter is known (Wood 2018d) and is effectively just AIC rescaled. AIC can be defined as [1]:

$$\text{AIC} = -2 \cdot l(x, \widehat{\mu}_\lambda) + 2 \cdot \text{edf}_\lambda$$

The deviance of a model is, as discussed in Section 3.2, defined as:

$$D_\lambda = 2 \cdot l(x, x) - 2 \cdot l(x, \widehat{\mu}_\lambda)$$

We can rewrite the definition of GCV_{UBRE} (with $s = 1$) to show the connection between the UBRE-score and AIC-score as:

$$\begin{aligned} (\text{GCV}_{\text{UBRE}} + 1) \cdot I &= D_\lambda + 2 \cdot \text{edf}_\lambda \\ &= 2 \cdot l(x, x) - 2 \cdot l(x, \widehat{\mu}_\lambda) + 2 \cdot \text{edf}_\lambda \\ &= 2 \cdot l(x, x) + \text{AIC} \end{aligned} \tag{6.2}$$

The smaller the UBRE-score the smaller the AIC, as the rest of the components in (6.2) do not depend on the model. Hence, the λ that gives the smallest UBRE-score will also give the smallest AIC-score.

Assuming Poisson distributed death counts the GAM Poisson regression model looks like the GLM Poisson regression model (3.4) when grouping of data is the same. The definition of η_i is what makes the difference in the types of modelling. In the GLM case, (3.2), now with only one numeric covariate and offset $\log(n_i)$, we have:

$$\log(\mu_i) = \log(n_i) + \eta_i = \log(n_i) + z_i\beta$$

In the GAM case however, we have;

$$\log(\mu_i) = \log(n_i) + \eta_i = \log(n_i) + f(z_i) \tag{6.3}$$

where $f(z_i)$ is a smoothing spline fitted over the numeric variable. Hence, the shape of the model is defined through a smoothed version of our numeric variables instead of the numeric variable itself, or a categorical variable constructed from the numeric variable, as we did with age groups and year groups in Section 3.1.

In (6.3) it is common to include a constant term, α, so that the model becomes:

$$\log(E[deaths_i]) = \alpha + \log(n_i) + f(z_i)$$

[1] Explained in Section 3.2, p.15, here with number of parameters p replaced by the effective number of parameters edf_λ.

Working with functions rather than the variables themselves however introduce an identifiability problem: the function f is only estimable up to an additive constant (Wood 2017, p. 175). For example, we can subtract a constant from α and simultaneously add it to f without changing the predictions of our model. For the model to be well-defined we need to solve this identifiability problem. It can be solved by introducing the sum-to-zero constraint (Wood 2017, p. 175), for the function f so that:

$$\sum_{i=1}^{I} f(z_i) = 0$$

The only effect of this constraint is that it sets the mean value of the function to zero by shifting it vertically. Hence, the function still has the same shape and penalty value as before the constraint was implied.

6.2 Model diagnostics

To specify our GAMs we use the `gam` function from the `mgcv` package in `R`. As a reference Wood (2017, Chapter 4) is used throughout this section. We will illustrate how the `gam` function works, by an example using our own data. Aggregating all our observations, only keeping gender and age fixed, we get two data sets, one for each gender, with 50 observations each representing age 20 to 69. We fit one model for each of these data sets through:

```
> library(mgvc)
> default.male = gam(deaths ~ s(age), offset = log(personYears),
  family ="poisson", data = male.subset)
> default.female = gam(deaths ~ s(age), offset = log(personYears),
  family ="poisson", data = female.subset)
```

This fits the generalized additive model;

$$\log(E[deaths_i]) = \alpha + \log(n_i) + f(age_i)$$

where $deaths_i \sim Poisson$ and the smoothing function f is specified by `s( )`. By default the amount of smoothness on variable age is obtained by minimizing (6.1) with respect to λ. The models have $df_\lambda = 4.98$ for the male model and $df_\lambda = 3.82$ for the female, hence a total $edf_{\text{male model}} = 5.98$ and $edf_{\text{female model}} = 4.82$. We add one to get the total edf due to the sum to zero constraint and the constant intercept term α. Thinking back to what we saw in the previous chapter, a higher df_λ means less smoothing. In this case it means that the model fitted to the female data is smoothed more, hence has a greater λ giving the punishment for curvature in the fitted function (5.11) more weight, than in the male model case.

We have plotted the observations and the fitted values along with the smoothed functions fitted to both data sets in Figure 6.1. Left panel shows estimated deaths for the original data vs. what actually has been observed. The straighter the line, the closer the estimated values are to the observed values. The middle and right panels of this figure show the estimated effects of

the smoothing terms as solid curves, with 95% confidence limits [2] marked as shaded regions. The fitted smoothing line effects are, as explained in the last section, constrained to sum to zero (Wood 2017, p. 183). The partial residuals are plotted as points in the middle and right panels of Figure 6.1. For a smooth term the partial residuals are the residuals that would be obtained if we kept all other estimates fixed and dropped the term of concern (Wood 2018a). Hence it is the smooth term pluss the model residuals. If the model fits well, the partial residuals should be evenly scattered along the curve they belong to. In our case the residuals seem to do so to a pleasant degree.

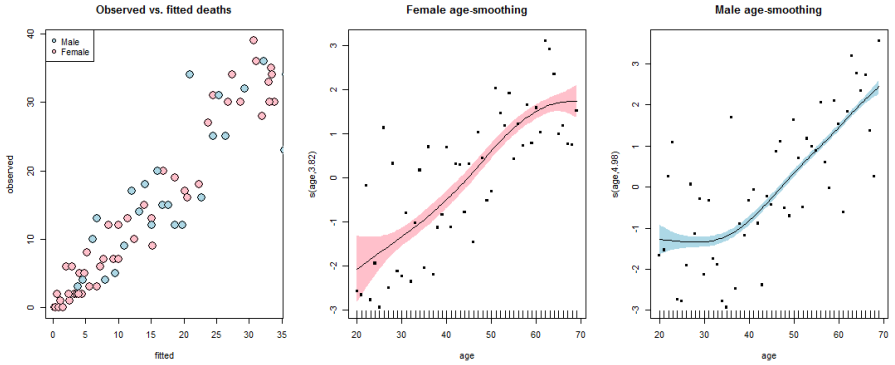


Figure 6.1: Left panel: Actual number of deaths observed plotted against predicted number of deaths. Predictions are made by default age-smoothed GAMs, fitted on male (blue) data and female (pink) data. Middle panel: Estimated smoothing curve for age in the female data set (solid line). The pink shade around the solid line is the 95% confidence interval. Right panel: Estimated smoothing curve for age in the male data set (solid line). The blue shade around the solid line is the 95% confidence interval. The black points in the middle and right panels of the figure are the partial residuals.

If we plot predicted death rates using these default smoothed models, along with the observed death rates for males and females, we get Figure 6.2. This figure has a recognisable pattern from what we observed in figures 2.1 and 3.5. The predictions and confidence intervals are made using `predict.gam()`, a function which work in the same way as `predict.glm()`, explained in Section 3.5, but for GAMs (Wood 2018e). We get the predictions and the upper and lower confidence intervals as follows, using the male population as an example:

```
> pred.age.m = predict.gam(model, newdata=data.frame(personYears = 1,
  age=20:69), se.fit=TRUE, type="link")
```

`se.fit=TRUE` allows us to get the standard error estimates of the predictions (Wood 2018e). Both predictions and standard errors are returned on the scale of the linear predictor. We create our confidence interval limits on the linear predictor scale through:

[2] The confidence intervals are strictly Bayesian credible intervals, we will not look further into these, but those curious are referred to Wood (2017, Section 6.10). The intervals are given by default running plot(GAMmodel) in R.

```
> upper = pred.age.m$fit + (2 * pred.age.m$se.fit)
> lower = pred.age.m$fit - (2 * pred.age.m$se.fit)
```

The resulting predictions, upper and lower confidence limits are then plotted on a log-scale by taking exp(pred), exp(upper) and exp(lower) and specifying `log = "y"` in the `plot()` function to get the plot on a log-scale.

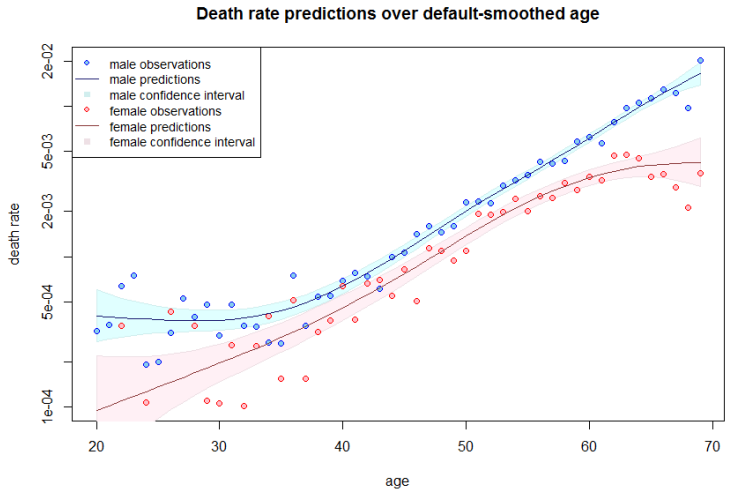


Figure 6.2: Predicted death rates (solid lines) by models default-smoothed for age, fitted on male (blue) and female (pink) data respectively. Data observations are drawn as points in the plot. The 95% confidence intervals are drawn for each model fit as shades around the model fit line, blue: male model confidence interval, pink: female model confidence interval.

We can carry out residual diagnostics for GAM in the same way as we did for GLM. The residuals should, as for GLM, be random, have a normal distribution[3], constant variance and zero mean. We check these assumption by plotting diagnostic plots of the deviance residuals, defined as (3.7)[4] in figures 6.3 and 6.4.

For the male model residuals, Figure 6.3, the residuals look random over both the fitted values, in panel b), and the age variable, in panel d). Looking at the QQ-plot and histogram of the residuals, panels a) and c) in the same figure, the residuals have a distribution which look close to normal in the QQ-plot, but not as close in the histogram. There are some peaks in the histogram, in the outer residuals regions (at residuals equal -2 and 1), that would not have been there with a perfectly normal distribution. The residuals do however meet the model assumptions to a pleasant degree.

In the female model case, Figure 6.4, the residuals look random over the age variable in panel d). There is however tendencies of a trend over the fitted values, panel b), when the fitted values are above 15. Looking at the QQ-plot and

[3]The residuals have a close to normal distribution only when $\widehat{\mu_i}$ is of reasonably large value (McCullagh and Nelder 1989, p. 38)

[4]Given that grouping of data is the same

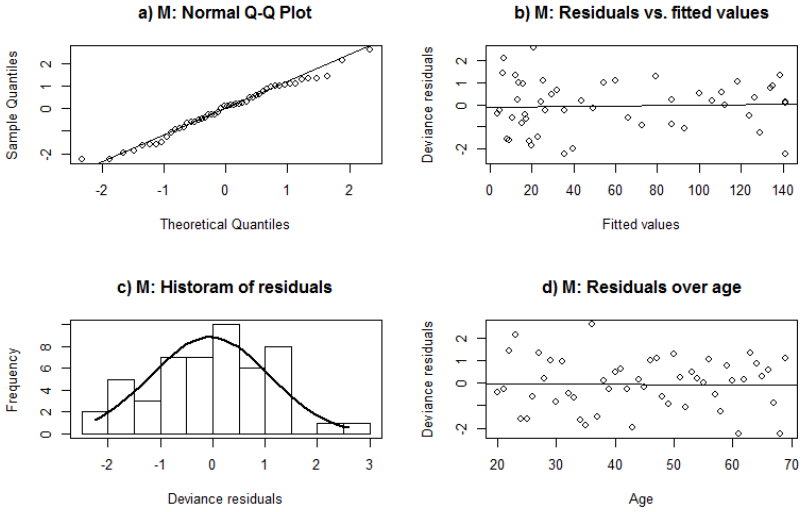


Figure 6.3: Residual diagnostic plots of male model default-smoothed for age. a) QQ-plot of deviance residuals. b) Deviance residuals plotted over fitted values. Trend line of residuals is given as a black line in the plot. c) Histogram of deviance residuals with a normal line plotted on top. d) Deviance residuals plotted over age with a trend line given as a black line in the plot.

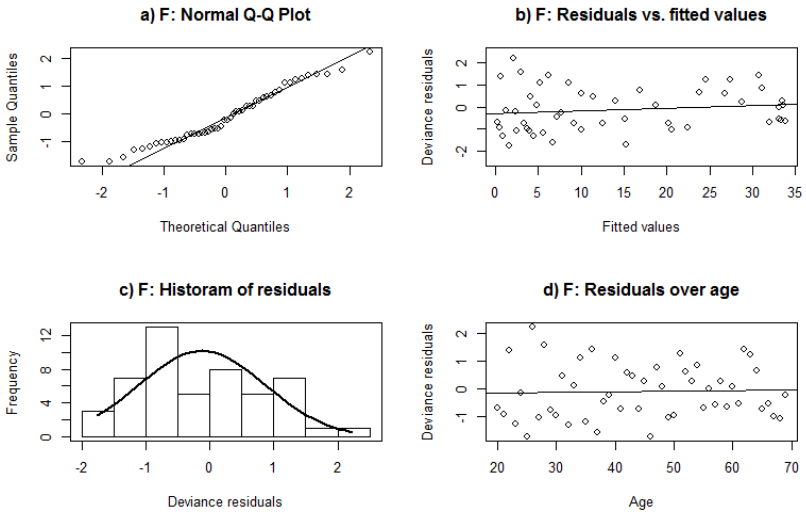


Figure 6.4: Residual diagnostic plots of female model default-smoothed for age. a) QQ-plot of deviance residuals. b) Deviance residuals plotted over fitted values. Trend line of residuals is given as a black line in the plot. c) Histogram of deviance residuals with a normal line plotted on top. d) Deviance residuals plotted over age with a trend line given as a black line in the plot.

histogram of the residuals, panels a) and c) in the same figure, the distribution of the residuals have a bit longer tail than what we would expect from a normal distribution. The residuals of the female model do not look as good as the residuals of the male model. This is however reasonable, as the $\widehat{\mu}_i$'s are greater across the age span for the male population. It is room for improvement in both models. We now go on to look at possible improvements of the models adding and smoothing over additional numerical variables.

6.3 Adding main effects

So far we have only looked at GAMs with one covariate. In this section however, we add year as a covariate, both in the data set and in the models. Aggregating our data keeping age, year and gender fixed we get two data sets, one for male and one for female, with 1350 observations each. Unless other is specified, Wood (2017, Section 4.3) is used as reference in this section.

We now want to fit models of the form

$$\log(\mathrm{E}[\mathrm{deaths}_i]) = \alpha + \log(n_i) + \mathrm{f}_1(\mathrm{age}_i) + \mathrm{f}_2(\mathrm{year}_i) \tag{6.4}$$

with $i = 1, ..., I$ and I equals the number of observations. Each function f_1 and f_2 is in the form we have discussed earlier, $f(z) = \sum_m^M \beta_m N_m(z)$, with parameters β_m and basis functions N_m for $m = 1, ..., M$, where M equal the number of basis functions. Each function is under the sum-to-zero constraint, as explained in Section 6.1, to avoid identifiability problems between the intercept and either of the functions or between the f_1 and f_2.

With this constraint in place we can find an optimal fit for each gender through a procedure called backfitting. To do so for a Poisson regression model, the backfitting procedure is used to maximize the penalized log-likelihood in combination with a likelihood maximizer. We will not do any further explanations regarding how this is done, but those interested are referred to Hastie, Tibshirani, and Friedman (2009, p.299–300).

6.4 Model selection criteria and model selection

Deciding on which model that fit our data best when we fit multiple models, is not as clear here as before. We can not use deviance comparison and LRT as the models are not nested. We may however, as mentioned in Section 3.2 use AIC and BIC to compare our models. Dealing with models containing splines we have to make small adjustments to the traditional definition of AIC and BIC given in Section 3.2 (Wood 2017, pp. 301–304);

$$\begin{aligned}\mathrm{AIC} &= -2(\log(L(x, \widehat{\mu})) - \mathrm{edf}_\lambda) \\ \mathrm{BIC} &= -2\log(L(x, \widehat{\mu})) + \log(\text{number of observations}) \cdot \mathrm{edf}_\lambda\end{aligned}$$

where the number of parameters p is replaced by edf, the effective degrees of freedom, or the effective number of parameters. This is done as the coefficients, as explained in Section 6.1, are fitted using penalized likelihood. In order to account for the penalized estimates we therefore use edf in the penalty terms of the definitions (Wood 2017, p. 301). The way we judge the AIC- and BIC-score is however the same as before. Comparing models fitted to the same data, the lower the score the better the model fit.

Male model selection

The models fitted to the male data set in this section is given in Table 6.1. Comparing AIC and BIC values it is clear that a model with age or year as only covariate is insuffisient, as the model with both smoothed age and smoothed year has the lowest UBRE-, AIC- and BIC-score. The edf of a model fitted with one covariate is, as mentioned earlier, the trace of the smoothing matrix $\boldsymbol{S}_\lambda$. The total edf of a GAM with more than one smoothed term is found by summing the degrees of freedom in each smoothing term and add one for the intercept (Wood 2017, p. 252).

Table 6.1: Table showing model summaries of GAMs fitted on male data with age and year as available variables. Variables are numerical and called A: age and Y: year. s(variable) means that smoothing splines are used on the variable. edf is the effective degrees of freedom in the model. UBRE, AIC and BIC colums give the corresponding test-scores for each model. Models are fitted on a data set with a total of 1350 observations.

model	variables	edf	UBRE	AIC	BIC
m.a.mod	s(A)	5.98	0.40	4139	4170
m.y.mod	s(Y)	9.39	2.82	7413	7462
m.ay.mod	s(A)+ s(Y)	14.14	0.06	3682	3756

The left panel of Figure 6.5 shows the observed number of deaths versus the fitted number of deaths for the male data set. The fit looks quite good up till around 12 deaths. There seem to be some factors or variations which lead to the model not predicting the highest death numbers. We will not look at the residual diagnostic plots of the model, as this is not the focus in this section, those curious however can find them in Appendix B, p.59, Figure B.1, panels a) to e).

Results of our smoothing procedure on the male data set, middle panel Figure 6.5, show less smoothing over age when year is included than with just age alone, with $df_{\lambda_{age}} = 5.13$ now, compared to $df_{\lambda_{age}} = 4.98$ which we have for the simpler model in both the previous section and this section. It is also clear from the right panel of Figure 6.5 that different degree of smoothing is required over age and year, as year is less smoothed with $df_{\lambda_{year}} = 8.01$. Hence, we have a more wiggly tendency across the year span, than we have for the age span. Except from the younger population, under 35, the death trend looks linear and the higher the age the higher the death rate. Over the year span we see a decreasing death rate going from the first years in the 1990's to 2017.

Female model selection

The models fitted to the female data set in this section are given in Table 6.2. As in the male case, it is clear that a model with only smoothed age or smoothed year is insuffisient. The difference is however not as clear as in the male model case, as the UBRE-, AIC- and BIC-score of the models are closer for the female models than they were for the male models.

The left panel of Figure 6.6 show observed number of deaths versus predicted number of deaths in the female data set. The fit here does not look as good as it did for the male model, as we have more points falling out from the trend

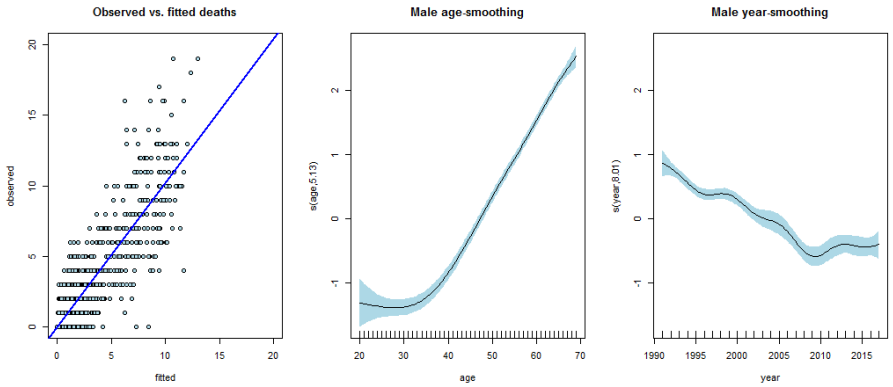


Figure 6.5: Left panel: Actual number of deaths observed plotted against predicted number of deaths. Predictions are made by a default age- and year-smoothed GAM, fitted on male data. Middle panel: Estimated smoothing curve for age in the male data set (solid line). Right panel: Estimated smoothing curve for year in the male data set (solid line). The blue shades around the solid lines in the middle and right panel of the figure is the 95% confidence intervals of the smoothed curves.

Table 6.2: Table showing model summaries of GAMs fitted on female data with age and year as available variables. Variables are numerical and called A: age and Y: year. s(variable) means that smoothing splines are used on the variable. edf is the effective degrees of freedom in the model. UBRE, AIC and BIC colums give the corresponding test-scores for each model. Models are fitted on a data set with a total of 1350 observations.

model	variables	edf	UBRE	AIC	BIC
f.a.mod	s(A)	4.84	-0.15	2149	2174
f.y.mod	s(Y)	3.21	0.34	2815	2832
f.ay.mod	s(A)+ s(Y)	13.25	-0.20	2083	2152

line and they fall out at an earlier stage of the death count span. We will not look at the residual diagnostic plots of this model either but those curious can find them in Appendix B, p.59, Figure B.1, panels f) to j).

Results of our smoothing procedures in the female data, middle panel Figure 6.6, show a bit more smoothing over age when year is included than with just age alone. We now have $df_{\lambda_{age}} = 3.57$, when both age and year is included, compared to $df_{\lambda_{age}} = 3.82$ which we had for the simpler model with age only in the previous section. As in the male model case, it is also clear from the right panel of Figure 6.6 that different degree of smoothing is required over age and year, as year is less smoothed with $df_{\lambda_{year}} = 8.68$. We have a more wiggly tendency across the year span in the female case, than what we had for males. We do however still see a decrease in death rates over years. Over age we see the same tendency as we did for the males, the higher the age the higher the death rate. The smoothed curve does however look non-linear for the older population, 60+, rather than for the youngest population, which was the male case.

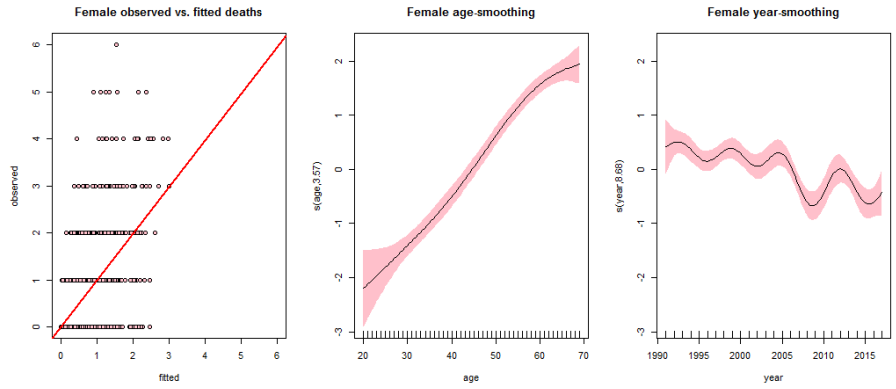


Figure 6.6: Left panel: Actual number of deaths observed plotted against predicted number of deaths. Predictions are made by a default age- and year-smoothed GAM, fitted on female data. Middle panel: Estimated smoothing curve for age in the female data set (solid line). Right panel: Estimated smoothing curve for year in the female data set (solid line). The pink shades around the solid lines in the middle and right panel of the figure is the 95% confidence intervals of the smoothed curves.

6.5 Adding interaction effects

We now have two main effects in our models, one for age and one for year. What is not included however is the possible interaction between age and year. We therefore also want to check the relevance of the interaction term, and fit a model of the form:

$$\log(\mathrm{E}[\mathrm{deaths}_i]) = \alpha + \log(n_i) + \mathrm{f}_1(\mathrm{age}_i) + \mathrm{f}_2(\mathrm{year}_i) + \mathrm{f}_3(\mathrm{age}_i, \mathrm{year}_i) \quad (6.5)$$

The interaction function, f_3, is under the sum-to-zero constraint which means that $\sum f_3(\mathrm{age}_i, \mathrm{year}_i) = 0$ (Hastie and Tibshirani 1990, p. 266).

The interaction term function $\mathrm{f}_3(\mathrm{age}_i,\mathrm{year}_i)$ can be fitted using isotropic smooths or tensor product smooths. Isotropic product smooths are defined through `s(age,year)` in R. Isotropic smooths is a good choice when smoothing two variables where we expect the same degree of smoothness over the covariate axes and when the covariates naturally are on the same scale (Wood 2017, p. 334). However, we have age and year as covariates and saw different degree of smoothing across them in the previous section. We also saw in Section 2.2 that the portfolio development and death rate variation look different across the two covariate spans. Isotropic smooth is therefore not preferred. Instead we use tensor product smooths interactions which allows different degrees of smoothing in different directions (Wood 2017, p. 227).

Tensor product smooths of multiple variables are built up from the smooths of fewer variables. We will explain this by showing an example using our own variables. The example is a recreation of an example in Wood (2017, Section 5.6.1) on the same topic. Assume we can represent smooth functions

for age and year by low rank basis functions through:

$$\mathrm{f}_1(age) = \sum_{i=1}^{I} \beta_i N_i(age) \quad \text{and} \quad \mathrm{f}_2(year) = \sum_{j=1}^{J} \beta_j N_j(year)$$

We now want a function f_3 which depends on both age and year. To achieve this we want to combine f_1 and f_2 in such way that we get a function which varies smoothly over both variables. We first allow the parameters of f_1 to vary smoothly with year:

$$\beta_i(year) = \sum_{j=1}^{J} \beta_{ij} N_j(year)$$

Inserting this new definition of the parameters of f_1 back to the definition of f_1 gives us f_3:

$$\mathrm{f}_3(age, year) = \sum_{i=1}^{I} \sum_{j=1}^{J} \beta_{ij} N_j(year) N_i(age)$$

This method can also be used to smooth over more than two variables following the same method as above.

In `R` we will define the tensor product smooths through functions `ti()` and `te()`. Function `ti()` produce an interaction effect with any lower interactions and main effects excluded (Wood 2018b). Hence, assuming that the lower terms are included in the model specifications. Function `te()` on the other hand produces a full tensor product smooth (Wood 2018b). Fitting models as (6.5) the models using `ti()` and `te()` both fit full models with interaction terms, the fit will however not be the same. This is due to the fact that they do not have the same penalty structure, where models using `ti()` have extra separate penalties of the basis representing the main effects (Wood 2017, p. 335), f_1(age) + f_2(year). We will use `ti()` for ANOVA decomposition of our models to check significance of interaction terms, then if the interaction term is significant we may fit a new model using `te()`.

We build on to the male and female models in the previous section, one model for each gender specified as:

```
> model = gam(deaths ~ s(age) + s(year) + ti(age,year),
  offset=log(personYears),family=poisson,data=subset for given gender)
```

in `R`. Summaries for the fitted male and female models are given in row 1 and 4 of Table 6.3 respectively. We want to check if the interaction effect is significant for each gender. We can do this using `anova(model)` in `R` (Wood 2017, pp. 335–336). For the male model we get a p-value equal $5.96 \cdot 10^{-09}$ for the interaction term, indicating that the interaction term is needed. That being said, the p-values for regression splines with estimated edf, as we do in this thesis through UBRE, can be misleading due to neglected uncertainty associated with estimation of λ (Zuur et al. 2009, p. 67). It is mentioned by Zuur et al. (2009, p. 67) that smoothers with p-values under 0.001 and above 0.2 can be trusted, but that extra considerations should be taken if the p-value is close to, or within the uncertain interval of half the 0.05 threshold. The p-value of $5.96 \cdot 10^{-09}$ is in this case well outside this interval. Comparing UBRE-,

AIC- and BIC-values of model m.ay.ten.mod, Table 6.3, with model m.ay.mod, Table 6.1, it is also clear that the model with the interaction effect included is preferred for the male data.

For the female data the choice of model is not as clear. The `anova` function returns a p-value equal 0.0711 for the interaction effect, which is within the uncertainty interval of the p-values. The AIC-score of model f.ay.ten.mod, Table 6.3 suggest that this model should be chosen over f.ay.mod, Table 6.2. The BIC-score, in the same tables for the same models, suggest the opposite however, preferring the model with no interaction effect. The residual plots of the female models look very similar[5], so it is difficult to choose a model based on residual looks.

Table 6.3: Table showing model summaries of GAMs fitted on male and female data with age and year as available variables. Variables are numerical and called A: age and Y: year. s(variable) means that smoothing splines are used on the variable. ti(variable) and te(variable) means the ti() or te() function explained in the text is used on the variales. edf is the effective degrees of freedom in the model. UBRE, AIC and BIC colums give the corresponding test-scores for each model. Models are fitted on a data sets with a total of 1350 observations each. Model name starting with m means model is fitted on male data. Model name starting with f means model is fitted on female data.

model	variables	edf	UBRE	AIC	BIC
m.ay.ten.mod	s(A)+ s(Y)+ti(A,Y)	23.42	0.02	3630	3752
m.ten.mod	te(A,Y)	17.41	0.03	3641	3732
f.ay.ten.mod	s(A)+ s(Y)+ti(A,Y)	21.19	-0.21	2076	2186
f.ten.mod	te(A,Y)	13.86	-0.20	2087	2160

Plotting the observed number of deaths over the fitted number of deaths by model f.ay.ten.mod for females, left panel Figure 6.7, there are no major changes to what we saw for the simpler model f.ay.mod, left panel Figure 6.6. This may suggest that the interaction term not contributes much to the fit of the model. What is different between the models, is the degree of smoothing over age and year. The model with interaction effect, f.ay.ten.mod, smooth less over age and more over year, than the model with no interaction term. The degree of smoothing for the main effects are read of the y-axis labels of middle and left panels in figures 6.6 and 6.7.

We also fit single tensor product smooths to the female and male data sets to see how smoothing over age and year at the same time affects the degree of smoothing and fit of the models. We specify the tensor product models for each gender through:

```
> model = gam(deaths ~ te(age,year), offset=log(personYears),
  family=poisson, data=subset for given gender)
```

in R. Model summaries are given in Table 6.3, rows 2 and 4 for male and female respectively. What is clear is that there is much more smoothing in these models. The edf of the male model has gone down from 23.42 to 17.41 and the

[5]See Appendix B, Figure B.1 panels f) to j) p.105 and Figure B.3 p.106 for residual diagnostic plots.

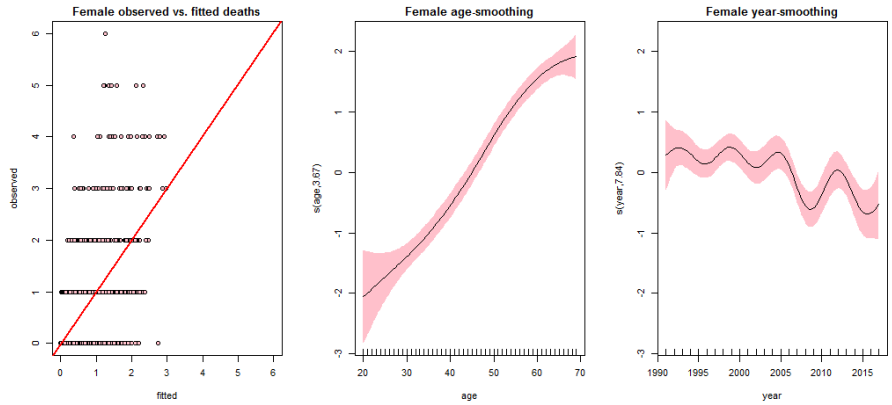


Figure 6.7: Left panel: Actual number of deaths observed plotted against predicted number of deaths. Predictions are made by GAM, f.ay.ten.mod, default-smoothed over age, year and the interaction of age and year. The model was fitted on female data. Middle panel: Estimated smoothing curve for age in the female data set (solid line). Right panel: Estimated smoothing curve for year in the female data set (solid line). The pink shades around the solid lines in the middle and right panel of the figure is the 95% confidence intervals of the smoothed curves.

female model has gone down from edf = 21.19 to edf=13.86. The penalty of "wigglyness" is in other words greater for the models specified through `te()` than for `ti()`, even though they both model:

$$\log(\mathrm{E}[\mathrm{deaths}_i]) = \alpha + \log(n_i) + \mathrm{f}_1(\mathrm{age}_i) + \mathrm{f}_2(\mathrm{year}_i) + \mathrm{f}_3(\mathrm{age}_i, \mathrm{year}_i)$$

The greater penalty on curvature is also clear when looking at the fits of the interaction models fitted through `ti()` versus the models fitted through `te()`.

Starting with the male model fits, Figure 6.8, it is hard to tell which fit is better. We can say on one hand that the left panel of Figure 6.8 look more realistic and that the model in the right panel of the same figure is too smoothed. On the other hand we can say that the left panel of Figure 6.8 is to sensitive to changes, specially over the year span, and that we believe the model in the right panel is a more stable model. The residual diagnostic plots of the two models look almost identical, basing the choice of model on these plots is therefore difficult [6].

For the female model fits, Figure 6.9, it is easier to prefer the more smoothed model, hence the single tensor product model f.ten.mod. Comparing left and right panel of Figure 6.9, the wiggly pattern over years is smoothed away all together in f.ten.mod, making a more stable predictions over years than f.ay.ten.mod. As in the male case, the residual diagnostic plots look the same for the two models, choosing a model based on residual looks is therefore difficult [7].

[6] See Appendix B, Figure B.2 p.106 and Figure B.4 p.107 for residual diagnostic plots.
[7] See Appendix B, Figure B.3 p.106 and Figure B.5 p.107 for residual diagnostic plots.

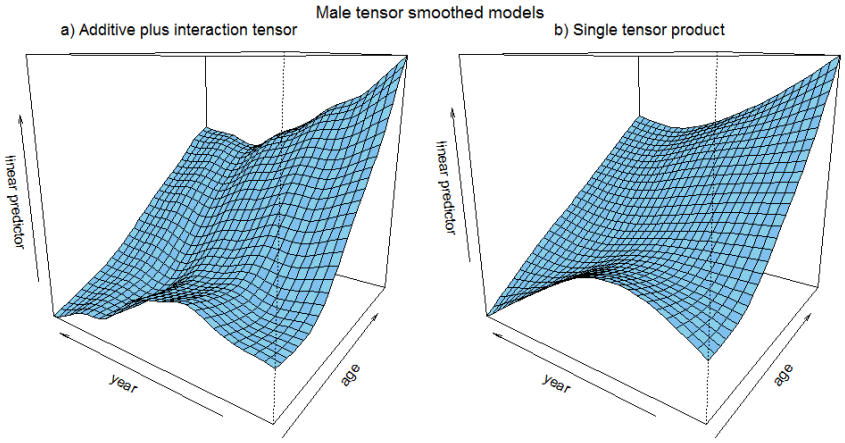


Figure 6.8: Death rate predictions of male models with smoothed interaction terms over age and year, plotted on the scale of the linear predictor (log-scale). a) Predictions of model m.ay.ten.mod, model with smoothed main terms pluss smoothed interaction. b) Predictions of model m.ten.mod, model with a single tensor product smoother.

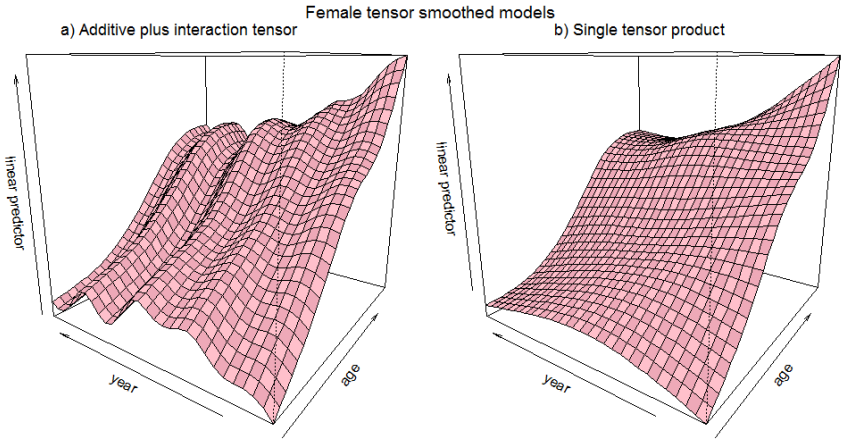


Figure 6.9: Death rate predictions of female models with smoothed interaction terms over age and year, plotted on the scale of the linear predictor (log-scale). a) Predictions of model f.ay.ten.mod, model with smoothed main terms pluss smoothed interaction. b) Predictions of model f.ten.mod, model with a single tensor product smoother.

6.6 NACE-section effects

In Section 3.5 we looked at the effect of adding NACE-section to our GLMs with age as a numeric variable and year as categorical year groups. In Section 3.5 the NACE-section effects turned out to be significant for the death rates of both males and females. In the male model case we got a more complex model, than in the female case, with interaction between NACE-section and year groups, as well as between NACE-section and age. We did however see that we lost some trends over the age span using a numeric age and that there were room for improvements in terms of smoothing over years. Using smoothed age and year, as we do in this section, we will most likely capture more of the variations that exist over the age and year span, as we allow for change in slope across the variable spans. The question is then if NACE-sections still have an effect.

We here, as in the previous section, treat age and year as numeric variables. In addition we add a categorical variable which is NACE-section. We will look at the same NACE-sections here as we did in Section 3.5. As a reminder, these sections are:

C - Manufacturing (699 deaths - Male: 613 and Female: 86)
K - Financial and Insurance Activities (431 deaths - Male: 302 and Female: 129)
G - Wholesale and retail trade; repair of motor vehicles and motorcycles.
(385 deaths - Male: 312 and Female: 73)

Grouping the data with single years, single ages and these three NACE-sections should give us two data sets with 4050 observations for each gender (3 NACE-sections x 50 ages x 27 years). As in Section 3.5 however, we do not have exposures in each observation group. We have 4017 observations for the male population and 3995 observations for the female population[8].

Smoothing function comparisons

To look at the differences and similarities between NACE-sections we first split each data set made for male and female in three. This gives us six data sets, one per gender within each NACE-section. We then fit six models in the following way:

```
> model = gam(deaths ~ s(age) + s(year),offset=log(personYears),
  family=poisson,data=subset for given gender-NACE-combination)
```

This corresponds to a model with interaction between NACE-section and smoothed age, and an interaction between NACE-section and smoothed year. Predictions and confidence intervals are made in the same way as in the last section, through `predict.glm()`. Using males in NACE-section C as an example:

```
> pred.year.c = predict.gam(model, newdata=data.frame(personYears=1,
  age=50, year=1991:2017), se.fit=TRUE, type="link")
> pred.age.c = predict.gam(model, newdata=data.frame(personYears=1,
```

[8]Total number of missing exposures in each NACE-section; C: 15, K:53, G:20. See Appendix B page 108 and 109 for missing male and female variable combinations respectively.

```
age=20:60, year=1997), se.fit=TRUE, type="link")
```

We get the confidence interval limits for each of the two predictions in the same way as we did in the previous section, by adding and subtracting 2-times the standard errors from the predictions and taking exp of the results. Doing so for each NACE-section in the male data sets and plotting predictions on a log-scale with shaded confidence intervals, we get Figure 6.10.

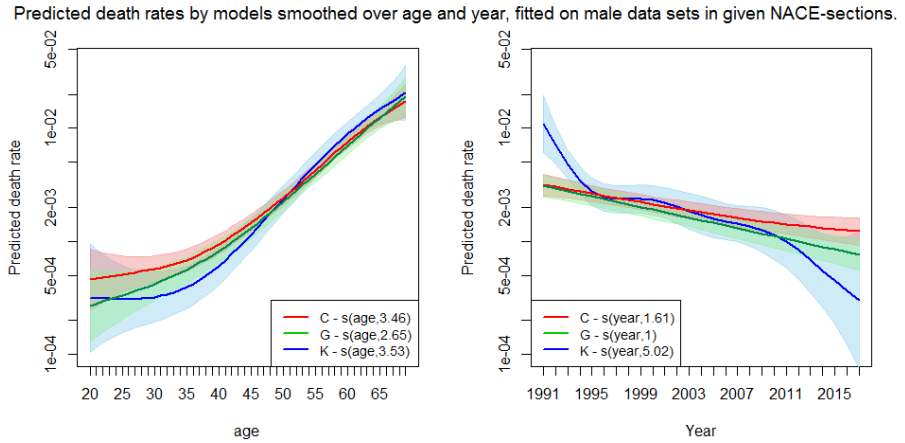


Figure 6.10: Death rate predictions and confidence intervals of models default-smoothed over age (left panel) and year (right panel). Models are fitted on three sets of male data, one set for each NACE-section. Each NACE-section is represented by a solid line with shaded 95% confidence intervals in each plot. Red: NACE-section C - manufacturing, green: NACE-section G - Wholesale and retail trade; repair of motor vehicles and motorcycles, blue: K - Financial and Insurance Activities. Predictions over age are made by fixing year at 1997 and predictions over years are bade by fixing age at 50. Both plots are on a log-scale.

Comparing the predictions of Figure 6.10 with what we observed in Figure 3.7 we have a recognizable pattern over both age and year. If we also compare the predictions of Figure 6.10 with the predictions made by the preferred male GLM in Section 3.5, M3.nace.m, from Figure 3.9 we get a better representation of the observations using a smoothed versions of age and year, rather than linear numeric age and categorical year groups.

In the left panel of Figure 6.10 we see that the degree of smoothing vary between the NACE-sections. NACE-section G is smoothed the most with $df_{\lambda_{age}} = 2.65$, followed by NACE-sections C and K with $df_{\lambda_{age}} = 3.46$ and $df_{\lambda_{age}} = 3.53$ respectively. If we look at the right panel of Figure 6.10 we get the same order of NACE-sections considering which has the highest and lowest degree of smoothing. The difference between the NACE-section which is smoothed the most, G, and smoothed the least, K, is however greater over year, than what it was for age. NACE-section G has a smoothed function equal a straight line with $df_{\lambda_{year}} = 1$ over years, where as NACE-section K has a much more fluctuating fitted curve with $df_{\lambda_{year}} = 5.02$. NACE-section C is also smoothed to a great extent, with a smoothed function close to a straight line,

with $df_{\lambda_{year}} = 1.61$.

The confidence intervals of the NACE-sections over both age and year are to a great extent overlapping. This does however not mean that the differences between them are insignificant. What we can say is that NACE-section K for the male population has a confidence interval for its predictions which is the furthest away from the two other sections.

We make the same prediction plots for the female data set, with predictions plotted on a log-scale and shaded confidence intervals in Figure 6.11. If we compare these predictions with the observations we made in Figure 3.8 we have recognizable patterns in our predictions. NACE-section K has higher death rates than the two other sections. The smoothed functions of NACE-sections C and G follow close trends over years, but have different shapes over age. The little dip in death rates for from the youngest population to those of age 30-39 in NACE-section C, seen in the observations, are still visible in the prediction plot. For NACE-section G however the corresponding dip is gone. This however builds up the hypothesis made in Section 3.5, that the dip in the observed death rates was due to random variation.

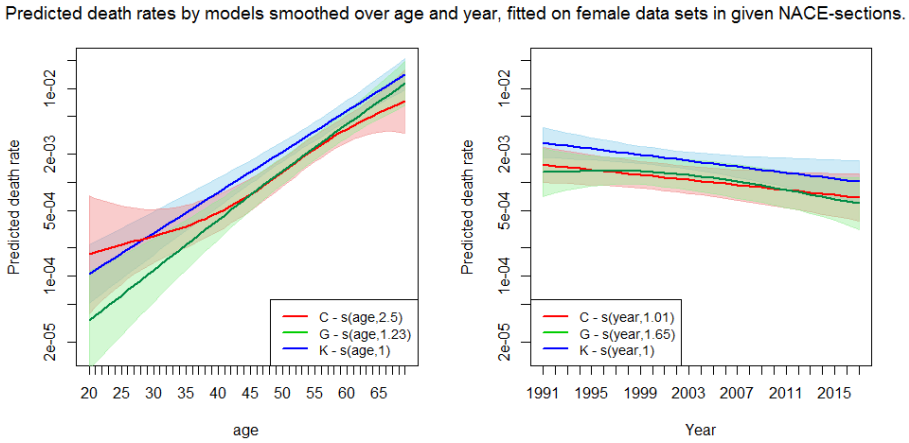


Figure 6.11: Death rate predictions and confidence intervals of models default-smoothed over age (left panel) and year (right panel). Models are fitted on three sets of female data, one set for each NACE-section. Each NACE-section is represented by a solid line with shaded 95% confidence intervals in each plot. Red: NACE-section C - manufacturing, green: NACE-section G - Wholesale and retail trade; repair of motor vehicles and motorcycles, blue: K - Financial and Insurance Activities. Predictions over age are made by fixing year at 1997 and predictions over years are bade by fixing age at 50. Both plots are on a log-scale.

Comparing the predictions in Figure 6.11 with the predictions made by the preferred GLM in Section 3.5, M1.nace.f, we again get a better representation of the observed death rates with smoothed age and year variables instead of numeric age and categorical year groups. The gain of using a smoother is most clear over age, as we recognize the pattern of our observations easier in the predictions. Over years it is not as clear, as the functions are smoothed close to a straight line and the predictions are more similar to the predictions that were

made by the GLM.

Given the same NACE-section, the female death rate predictions are smoothed more than the male death rate predictions over both age and year. We also have a different order of which NACE-sections have the least and most smoothed death rate predictions in the female models, than what we had for the male models. NACE-section K, which had the least smoothed functions in the male models has the highest degree of smoothing for the female models with both $df_{\lambda_{age}} = 1$ and $df_{\lambda_{year}} = 1$. Over age, in the left panel of Figure 6.11, we see that NACE-section C has the least smoothed function across age with $df_{\lambda_{age}} = 2.5$ followed by NACE-section G with $df_{\lambda_{age}} = 1.23$. Over years, in the right panel of Figure 6.11, we see that the least smoothed function belongs to NACE-section G with $df_{\lambda_{year}} = 1.65$ followed by NACE-section C with $df_{\lambda_{year}} = 1.01$. Hence the two NACE-sections swap places in having the least or second least smoothed functions depending on which smoothed variable we look at.

Model fitting - Interaction between categorical and numerical variables in GAM

To see if the observed differences in smoothing effects are significantly different across NACE-sections we fit GAMs in the same way as done earlier in this chapter, using the `gam` function from the `mgcv` package in `R`. We start with the simplest model with main effects only, here using the male model as example:

```
> gam.M1.nace.m = gam(deaths ~ s(age) + s(year) + NaceMain,
  offset=log(personYears), family=poisson, data=subset for males)
```

This model assumes that all NACE-sections have the same death-age-year relationship, as there are no interactions between either of the covariates. The only possible difference we may see across NACE-sections in this model are therefore constant values. From what we saw in the previous section however, the NACE-sections had smoothed curves of different shapes across the age and year span. This was especially clear for NACE-section K in the male data set.

Interaction between numeric- and categorical variables in GAM is different from the interaction we know from GLM (Zuur et al. 2009, p. 60). Adding an "interaction" between NACE-sections and a numeric variable, say age, in GAM can be specified in three different ways;

I): `s(age,by=NaceMain)`
II): `s(age) + s(age, by=as.numeric(NaceMain=="K"))`
III): `s(age) + s(age, by=NaceMain)`

as references for these methods Zuur et al. (2009, pp. 60–63) and Wood (2018c) are used. Option I) fits an individual smoothing curve over age for each NACE-section. Option II) fits a smoothing curve over age for all our data, hence it gives the overall effect of age across all NACE-sections. Another smoother across age is then fitted for NACE-section K using the `by` argument. It adjusts the pattern over age for NACE-section K through `as.numeric(NaceMain=="K")` which returns one if an observation is from NACE-section K and zero if its

not[9]. Option III) does the same as option II) but adjusts the pattern of all the NACE-sections, not only NACE-section K, adding interactions effects in the order of the factor levels. Through option III) we can check if there is a significant interaction between a numeric variable and NACE-sections in general, the option also allow us to check if the interaction effect is the same across NACE-sections, e.g. $\theta_K = \theta_C = \theta_G$. Option II) allow us to check if we have one NACE-sections which have an interaction effect significantly different from the others, eg. $\theta_K \neq \theta_C$.

Male - Model fitting and model selection

Starting with the simplest model, only including main effects, as specified above, we choose which interaction term to add next using AIC. We fit three different models, each with one of the possible first order interaction effects;

```
> test.mod1 = gam(deaths ~ s(age) + s(year) + NaceMain + ti(age,year),
  offset=log(personYears), family=poisson, data=subset for males)
> test.mod2 = gam(deaths ~ s(age) + s(year) + NaceMain + s(age,by=NaceMain),
  offset=log(personYears), family=poisson, data=subset for males)
> test.mod3 = gam(deaths ~ s(age) + s(year) + NaceMain + s(year,by=NaceMain),
  offset=log(personYears), family=poisson, data=subset for males)
```

The model which returns the lowest AIC value, is test.mod1. We then check if the interaction term is significant, using `anova(test.mod1)`. The test returns a p-value of 0.006, we therefore add the smoothed interaction between age and year to our model and call the new model gam.M2.nace.m.

We decide which interaction term to add next following a similar procedure. We make two test models adding `s(age,by=NaceMain)` and `s(year,by=NaceMain)` in turn to model gam.M2.nace.m. The test model that gets the lowest AIC is the model with interaction between year and NACE-section. The adjustment over year, may however not be significant for all NACE-sections. An anova test of the model as it is now may be misleadning, in terms of which NACE-section interactions are significant and not, as it fits the interactions in the order of the NACE-section factor levels[10]. We therefore fit test models adding adjustments for one NACE-section at the time:

```
> test.mod1 = gam(deaths ~ s(age) + s(year) + NaceMain + ti(age,year) +
  s(year, by=as.numeric(NaceMain=="C")), offset=log(personYears),
  family=poisson, data=subset for males)
> test.mod2 = gam(deaths ~ s(age) + s(year) + NaceMain + ti(age,year) +
  s(year, by=as.numeric(NaceMain=="G")), offset=log(personYears),
  family=poisson, data=subset for males)
> test.mod3 = gam(deaths ~ s(age) + s(year) + NaceMain + ti(age,year) +
  s(year, by=as.numeric(NaceMain=="K")), offset=log(personYears),
  family=poisson, data=subset for males)
```

[9]The function makes an indicator variable for observations from NACE-section K.
[10]Order of the NACE-sections as factor levels: C, G, K

Testing the significance of the interaction effects by using anova on each of the test models show that the most significant interaction is between year and NACE-section K. The test model with this interaction is also the model with the lowest AIC. We then test if the interactions between the two other NACE-sections and year also should be added by adding `s(age, by=as.numeric(NaceMain=="C"))` and `s(age, by=as.numeric(NaceMain=="G"))` in turn to test.mod3. Neither of the interaction effects improves the AIC of the model and neither of them get an anova-test indicating that they are significantly different from zero. Our third male model, gam.M3.nace.m, therefore end up being;

```
> gam.M3.nace.m = gam(deaths ~ s(age) + s(year) + NaceMain + ti(age,year) +
  s(year, by=as.numeric(NaceMain=="K")), offset=log(personYears),
  family=poisson, data=subset for males)
```

We then try to add the last first order interaction effect `s(age,by=NaceMain)`, this gives a model with a lower AIC than gam.M3.nace.m. We check if the interaction between age and NACE-section is significant in all NACE-sections following the same procedure as we did for the year and NACE-section interactions. The only age and NACE-section interaction effect that is marked as significant is with age and NACE-section K. Adding this interaction to gam.M3.nace.m, yields our fourth model, gam.M4.nace.m.

With all significant first order interaction effects added, it only remains to check if we have a second order interaction which will improve the fit of our model. We only add a second order interaction effect between age, year and NACE-section K, as we do not have first order interaction effects for the other NACE-sections. Doing so, we get the model gam.M5.nace.m. The second order interaction is however not significant and the AIC of gam.M5.nace.m is higher than the AIC of gam.M4.nace.m. An overview and summary of the models fitted are given in Table 6.4.

Table 6.4: Table showing model summaries of GAMs fitted on male data with age, year and NACE-section as available variables. Age and Year are smoothed numerical variables and called s(A) and s(Y) respectivelly. NACE-section is a categorical variable of three levels and called N. The functions ti(multiple variables) means that the ti() function explained in the text is used on the variables. p-val is the returned p-value from anova for the added covariate. edf is the effective degrees of freedom in the model. UBRE, AIC and BIC columns give the corresponding test-scores for each model. Models are fitted on a data set with a total of 4017 observations.

model	variables	p-val	edf	UBRE	AIC	BIC
gam.M1.nace.m	s(A)+s(Y)+N	-	12.23	-0.404	4297	4374
gam.M2.nace.m	s(A)+s(Y)+N+ti(A,Y)	0.00686	19.94	-0.408	4283	4408
gam.M3.nace.m	s(A)+s(Y)+N+ti(A,Y)+s(Y):(N=K)	0.00180	23.78	-0.411	4271	4420
gam.M4.nace.m	s(A)+s(Y)+N+ti(A,Y)+s(Y):(N=K)+ s(A):(N=K)	0.03765	24.23	-0.411	4267	4419
gam.M5.nace.m	s(A)+s(Y)+N+ti(A,Y)+s(Y):(N=K)+ s(A):(N=K)+ti(A,Y,N=K)	0.9543	25.21	-0.411	4269	4428

Comparing the fitted models, given in Table 6.4 the model with the lowest

AIC is, as mentioned earlier in this section, gam.M4.nace.m. The last interaction effect added does however have a p-value close to the 0.05 threshold. We have earlier addressed the uncertainty regarding the p-values of GAM. An alternative could therefore be to choose gam.M3.nace.m. The p-value of the last interaction term added in gam.M3.nace.m is smaller and the difference in AIC between model gam.M3.nace.m and gam.M4.nace.m is also small. The residuals of the models look almost the same, basing model choice on residual diagsnostics is therefore difficult[11].

If we compare predictions of gam.M3.nace.m and gam.M4.nace.m, it is easier to see the difference between the two. For NACE-sections C and G the predictions are almost the same, with predictions of gam.M3.nace.m slightly below the predictions of gam.M4.nace.m for most covariate values [12]. For NACE-section K however, Figure 6.12, we see a clear difference in predictions over age. Model gam.M4.nace.m has a steeper curve over age than model gam.M3.nace.m. Remembering what we observed in Figure 3.7 (Section 3.5) the predictions of model gam.M4.nace.m give a better representation of of the death-rate pattern over age, than the predictions of model gam.M3.nace.m. Considering this, the fact that model gam.M4.nace.m has the lowest AIC-score and the fact that the BIC-score also is in favour of gam.M4.nace.m, comparing gam.M3.nace.m and gam.M4.nace.m, we chose gam.M4.nace.m as our preferred model.

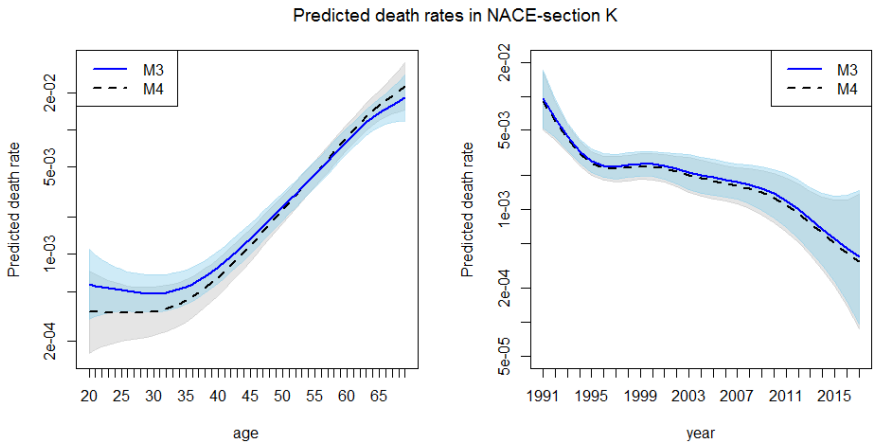


Figure 6.12: Death rate predictions with confidence intervals of default smoothed models, gam.M3.nace.m (M3: Blue line and confidence interval) and gam.M4.nace.m (M4: Gray stippled line and confidence interval), in NACE-section K - Financial and Insurance Activities. Predicitons over age (left panel) are made by fixing year at 1997. Predictions over year are made by fixing age at 50. The predictions are plotted on a log-scale.

[11] Residual diagnostic plots for model gam.M3.nace.m and gam.M4.nace.m are attached in Appendix B, page 110

[12] Plot of predictions across age and year for both models in both NACE-sections are attached in Appendix B, page 112

Female - Model fitting and model selection

We fit models to the female data set in the same way as we did for the males, starting with main effects of age, year and NACE-section. Fitting three test models, each with one of the possible first order interaction effects, the model with interaction between age and year is the only model with a better AIC-score than the simplest model. We call this model gam.M2.nace.f. When we check if the interaction effect between age and year is significant however, we get a p-value of 0.1047, which indicates that the interaction effect is not significantly different from zero. This is not surprising, considering what we saw in the death observations of Figure 3.8 (Section 3.5) and the close to parallell death rate predictions in Figure 6.11, seen earlier in this section. Table 6.5 gives a summary of two of the models fitted to the female data.

Table 6.5: Table showing model summaries of GAMs fitted on female data with age, year and NACE-section as available variables. Age and Year are smoothed numerical variables and called s(A) and s(Y) respectivelly. NACE-section is a categorical variable of three levels and called N. The functions ti(multiple variables) means that the ti() function explained in the text is used on the variables. p-val is the returned p-value from anova for the added covariate. edf is the effective degrees of freedom in the model. UBRE, AIC and BIC columns give the corresponding test-scores for each model. Models are fitted on a data set with a total of 3995 observations.

model	variables	p-val	edf	UBRE	AIC	BIC
gam.M1.nace.f	s(A)+s(Y)+N	-	7.51	-0.699	1732	1780
gam.M2.nace.f	s(A)+s(Y)+N+ti(A,Y)	0.1047	11.96	-0.699	1729	1804

Model gam.M2.nace.f has a better AIC-score than the simpler model, gam.M1.nace.f. The interaction effect of model gam.M2.nace.f does however have an unsignificant p-value and the BIC-score is in favour of the simpler model. We therefore end up choosing the simplest model, gam.M1.nace.f, with main effects only as our preferred model[13].

Prediction interpretation

We have now chosen two GAMs which we believe represent the variation in death rates well, one for each gender. To say something about the development of predicted death rates over the variables in use, we will look at predictions using covariate values equal to the minimum, maximum and middle observed value of age and year within each NACE-section.

We make predictions over age for the years 1991, 2004 and 2017[14] within each NACE-section using gam.M4.nace.m for males and gam.M1.nace.f for females. Plotting these predictions yields Figure 6.13. Stippled lines in the figure gives death rate predictions for the male population and solid lines the

[13] We do not look at the residual diagnostic plots here, they are however attached in Appendix B, p.111

[14] These years are chosen as they are in the start, middle and end of the year variable span. 2004 is a special year with little data, the GAM does however smooth well over this year. 1991 and 2017 are in the boundaries of the parameter space, they do however still show the development of predictions well and is therefore chosen.

corresponding for the female population. For all NACE-sections the difference in death rates between genders are predicted to decrease with years.

The difference in predicted death rates between each of the chosen years are greater for the male population than for the female population. This is especially clear in NACE-section K, in the right panel of Figure 6.13. Death rate predictions in NACE-section K over age in the year 1991 is much higher than the corresponding predictions of the female population in the same year. In year 2004 however the death rates over age for those above the age of 35 are predicted to be close to similar for the two genders. When we look at the most recent year which were available in our data set, 2017, the male population in NACE-section K is predicted to have lower death rates than the female population in the same NACE-section. In NACE-sections C and G the death rates over age for the two genders are predicted to be closer in 2017 than they were in 1991, but the female population still have the lowest death rates in the two NACE-sections.

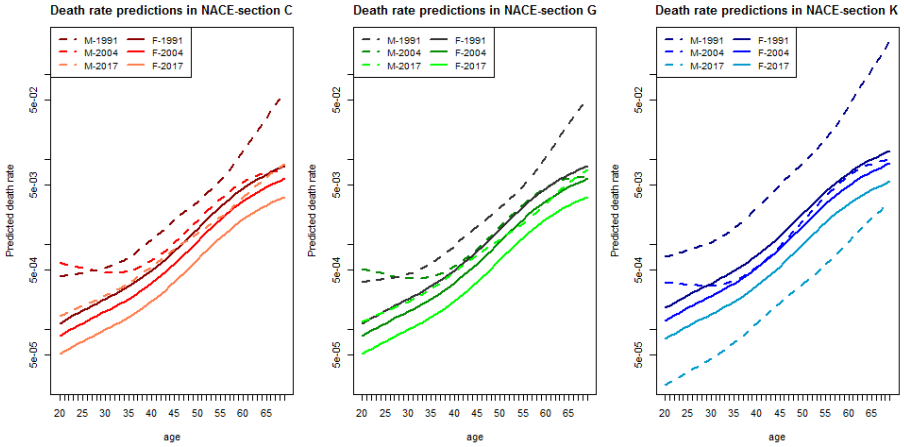


Figure 6.13: Predicted death rates for male and female over age for given years in the three biggest NACE-sections. Predictions for females (solid lines) are made by model gam.M1.nace.f. Prediction for males (stippled lines) are made by model gam.M4.nace.m. Each year is represented by a color for each gender in each NACE-section. Darkest color: 1991, medium color: 2004, lightest color: 2017. Each NACE-section is given its own panel. Left panel: C - Manufacturing. Middle panel: G - Wholesale and retail trade; repair of motor vehicles and motorcycles. Right panel: K - Financial and Insurance Activities. All predictions are plotted on a log-scale.

To get get the full picture of the death rate predictions we will also look at the death rate predictions over years for our youngest, oldest and middle-aged males and females. We make predictions over years for ages 20, 45 and 69 within each NACE-section using our chosen models for the male and female population. Plotting these predictions yields Figure 6.14, again with stippled lines for the male population and solid lines for the female population.

Comparing NACE-sections for the male population, the stippled lines of Figure 6.14, which NACE-section that has the lowest and highest death rates change differently over years for the different ages. This is of course due to

the interaction effect between age and year included in the final male model, gam.M4.nace.m. NACE-section G always has slightly lower death rates than NACE-section C, as the only difference between these sections are a constant value shifting the smoothed year curve down for NACE-section G. This is because the only interaction term we included in our final model for NACE-sections was with NACE-section K.

For the oldest men, age 69, NACE-section K is predicted to have the highest death rate in 1991, after 2014 however NACE-section K has the lowest death rate for the oldest men. For the middle-aged men, NACE-section K is predicted to have the highest death rates till around year 2009 and the lowest death rates compared to the other two NACE-sections after this. For the youngest men, age 20, the shift where NACE-sectionK gets the lowest predicted death rates happen even earlier, in year 2001.

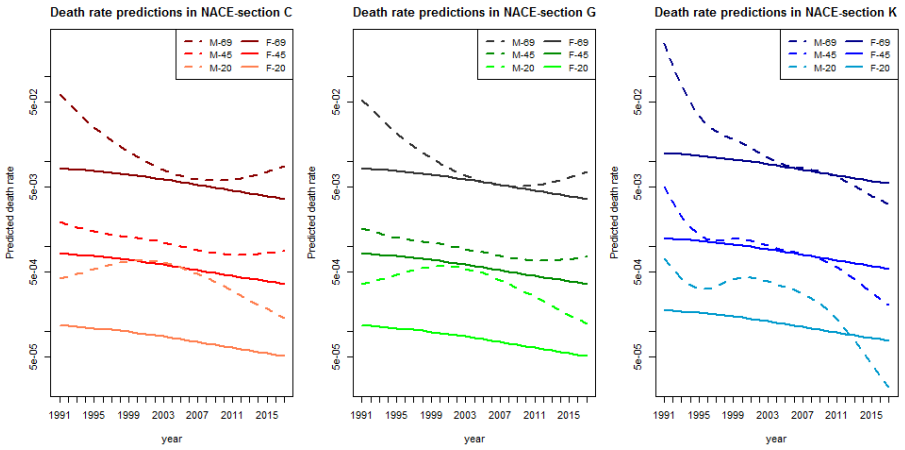


Figure 6.14: Predicted death rates for male and female over year for given ages in the three biggest NACE-sections. Predictions for females (solid lines) are made by model gam.M1.nace.f. Prediction for males (stippled lines) are made by model gam.M4.nace.m. Each age is represented by a color for each gender in each NACE-section. Darkest color: 69, medium color: 45, lightest color: 20. Each NACE-section is given its own panel. Left panel: C - Manufacturing. Middle panel: G - Wholesale and retail trade; repair of motor vehicles and motorcycles. Right panel: K - Financial and Insurance Activities. All predictions are plotted on a log-scale.

Comparing NACE-sections in the female population, solid lines of Figure 6.14, the NACE-sections keep the same order in terms of highest to lowest death rates for each year, no surprise as we only have main terms in the female model. NACE-section G and C are predicted to have close to identical death rates, NACE-section G is however predicted to have slightly lower death rates. NACE-section K is predicted to have the highest death rates for the female population. Across all NACE-sections we see a decrease in death rate over years for the female population.

In the right panel of Figure 6.13 we saw that the male population in NACE-section K was predicted to have lower death rates than the female population across all age groups in the year 2017. In the right panel of Figure 6.14, we see

that it is predicted to happen at different in different years for people of different age. Of the three ages plotted in Figure 6.13, the oldest and middle-aged men in NACE-section K is predicted to have the same death rates as women in 2006. After this they are predicted to have lower death rates than the women in NACE-section K. The youngest males in NACE-section K have lower death rates than the youngest women in the same NACE-section from year 2013 onwards.

CHAPTER 7

Summary and discussion

This last chapter gives an overview and discussion of what we have found thorugh the analysis in the earlier chapters, explains challenges met along the way and explain further work that may be done.

7.1 Differences in variables selected by GLM and GAM

When we fitted GLMs with age and year as available variables in Sections 3.3, it was easier to reduce the interaction effect for the male model, by using grouped years, than in the corresponding female model. When we included NACE-sections in Section 3.5 however, the interaction effect between age and year for the female model was no longer significant. We ended up with a model with main effects of age, year and NACE-section. This may indicate that the variation at first believed to be caused by interaction effects between age and year in the female population, actually is due to a change in the composition of NACE-sections over time.

For the male GLMs, including NACE-sections also lead to a less significant interaction between age and year groups, and it was the last first order interaction term to be added in the possible models. We ended up choosing a model without the interaction between age and year groups, but the choice was not as clear as in the female case.

When we fitted GAMs with age and year as available variables, Section 6.5, we got a somewhat different results from what we got for GLMs in terms of which gender had the most significant interaction effect between age and year. For the male population a model with interaction term was clearly preferred. For the female population however the choice of model was not as easy, and the model with an interaction effect was not as clearly preferred as in the GLM case.

When we added NACE-sections to the GAMs however, we ended up with main effects of smoothed age, smoothed year and NACE-sections for the female population. This is the same as we saw for the GLMs, the interaction between age and year is not significant for the female population when NACE-section is added. For the male population however the interaction between age and year was the most significant after NACE-sections was added. This differs from what we saw when fitting male GLMs, because, as mentioned earlier in this section, the age-year interaction was the least significant interaction effect out of the possible first order interactions.

7.2 Comparison of fitted GLMs and GAMs

In previous chapters we have discussed GLMs and GAMs separately, without much comparison of estimated death rates between the different models. In this section however we want to look at the similarities and differences between GLMs and GAMs which use the same costumer properties, for example age groups and smoothed numerical age. The variables used in the models are not necessarily the same, but are meant to capture the same observed death rate variations.

Age and year models

The first models we will compare are the GLMs $M1_m$ and $M1_f$, from Section 3.3, and the GAMs m.asas.mod.ten and f.asas.mod.ten, from Section 6.5. As a reminder the GLMs used categorical ten-year age groups and categorical single years for the main effects in the model and categorical three-year year groups for the interaction effect with age groups. In the GAMs we used smoothed age and smoothed year for the main effects with a tensor product smoother for the interaction between age and year.

We make GLM predictions over years for each age group 20-29,...,60-69 using `predict.glm()`, with model $M1_m$ for males and $M1_f$ for females. We do the same for the GAMs, with `predict.gam()`, with model m.asas.mod.ten for males and f.asas.mod.ten for females. The models do however not have age groups as a variable. We therefore use ages 25,35,...,65, as they are in the middle of each age group interval. Plotting these predictions yields Figure 7.1, where points in the figure are GLM predictions and lines are GAM predictions.

For the male population, Figure 7.1 a), it almost look like the GLM predictions twirls around the GAM prediction lines. For most years there are little differences in the death rates predicted by GLM and GAM. For year 2004 however, the GLM makes predictions close to zero, and they are therefore missing in the plot. For the two youngest age groups, 20-29 and 30-39, it is also easy to spot the difference between GLM and GAM predictions for the last 10 years. The predictions of the GLM is less stable, with greater differences in predictions from year to year than the predictions of the GAM. The GLM and GAM do however agree on less differences in death rates between the two youngest groups, compared to the other age groups, and both GLM and GAM predict higher death rates for those of age 20-29 than 30-39 at given years.

For the female population, Figure 7.1 b), it is easier to spot the differences between GLM and GAM predictions. The variation in predicted death rates are greater from year to year for the GLM than the GAM. As with the males, the GLM death rate prediction is close to zero in year 2004, whereas the GAM smooths over it, keeping a steady trend from year 2003 to 2005. It is not as big, but we see the same kind of drop in death rates for the GLM predictions in year 2009. In 2009 the GLM predicts lower death rates for age groups 20-29, 30-39 and 40-49 than the GAM predicts for age group 20-29. Both the GLM and GAM have shifting death rates over years but the GAM predictions are more stable than the GLM predictions.

If we make predictions using the same GLMs and GAMs as earlier in this section, this time over age for given years, we get Figure 7.2. As the previous figure, the points in the figure are GLM predictions and lines are GAM

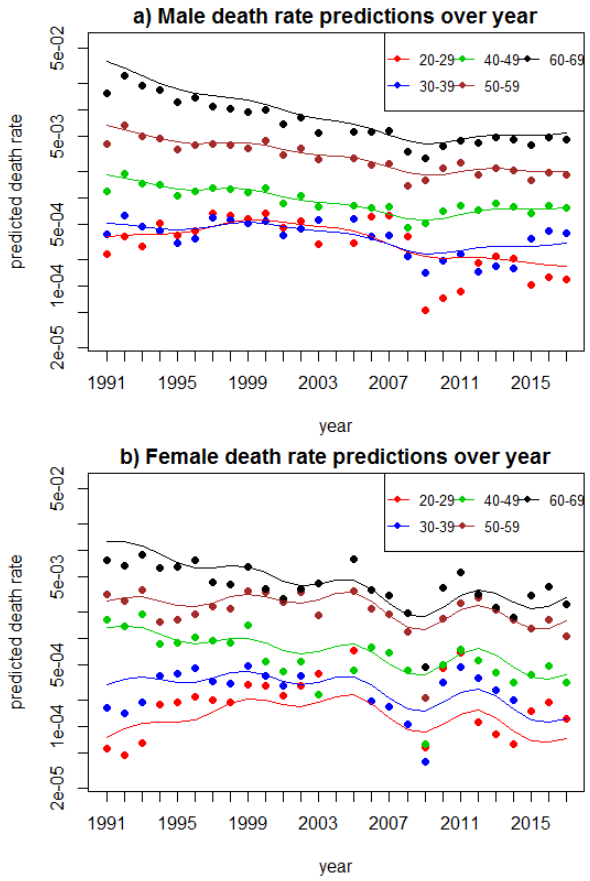


Figure 7.1: Predicted death rates for male, panel a), and female, panel b), over years by GLM (points) and GAM (lines) for given age groups. Each age group is represented by a color; red: 20-29, blue: 30-39, green: 40-49, brown: 50-59 and black: 60-69. GLM predictions are made by model $M1_m$ for males and model $M1_f$ for females. GAM predictions are made by model m.asas.mod.ten for males and f.asas.mod.ten for females. GAMs do not have age groups as a variable, age = 25,35,...,65, is therefore used for predictions within each age group. All predictions are plotted on a log-scale.

predictions. In the male predictions, Figure 7.2 a), we see that both GLM and GAM predict lowered slopes for death rates over age from year 1991 to 2017. The clearest difference in the male predictions is in year 1991, where the GAM predicts a higher death rate than the GLM across all age groups.

In the female predictions, Figure 7.2 b), GLM and GAM have the least difference between predictions in year 1991, hence the opposite of what we saw in the male predictions. For year 2005 the predictions done by GLM and GAM have greater differences, and for the age group 30-39 the GLM makes predictions close to zero. In 2017 the GLM and GAM predictions look close for most years, but for the age group 30-39 we get, as we did for 2005, predictions close to zero. The GLM and GAM do however agree on a lowered slope for death rates over age going forward in time. All inn all the GAMs seem to be less sensitive to single observations and look more stable over both age and year

than the GLMs.

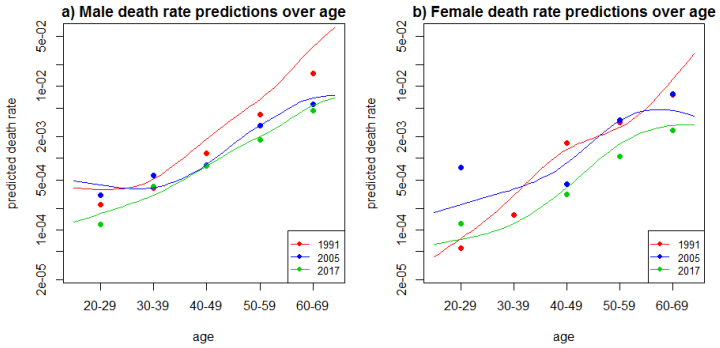


Figure 7.2: Predicted death rates for male, panel a), and female, panel b), over age by GLM (points) and GAM (lines) for given years. Each year is represented by a color; red: 1991, blue: 2005, green: 2017. GLM predictions are made by model $M1_m$ for males and model $M1_f$ for females. GAM predictions are made by model m.asas.mod.ten for males and f.asas.mod.ten for females. All predictions are plotted on a log-scale.

Age, year and NACE-section models

As with the models with age and year as variables, we will compare GLMs and GAMs with age, year and NACE-section as covariates. In Section 3.5 we chose model M1.nace.f for the female population and model M3.nace.m for the male population. Model M1.nace.f has main effects only, with linear age, categorical grouped years and NACE-sections as covariates. Model M3.nace.m has the same main effects, but in addition has two first order interactions, one between NACE-sections and year groups, and one between NACE-sections and age.

We will compare the predictions of these models with the prediction made by GAMs gam.M1.nace.f and gam.M4.nace.m, fitted to female and male data respectively in Section 6.6. Model gam.M1.nace.f is a model with smoothed age, smoothed year and NACE-sections as covariates, it has main effects only, just like the GLM M1.nace.f. Model gam.M4.nace.m has the same main effects as gam.M1.nace.f, but in addition has three interaction effects. First is an interaction effect between age and year, made by using a tensor product smoother. The two other interaction effects are between NACE-section K and smoothed year, and between NACE-section K and smoothed age.

We start with the male models and make predictions in the same way as we did for the simpler models earlier in this section. Plotting predictions over years for each of the three NACE-sections for males of given ages, yields Figure 7.3. As in earlier figures of this section the points in the plot are GLM predictions and lines are GAM predictions. The GLMs do not have single years as covariates in the models, instead they have grouped years. We have therefore put the GLM predictions in the year which is in the middle of the year group, for example, the predictions in year group 1991-1993 is drawn in year 1992 in the prediction plots.

Comparing predictions across NACE-sections, both the GLM and the GAM have higher death rates for NACE-section K than the two other NACE-sections

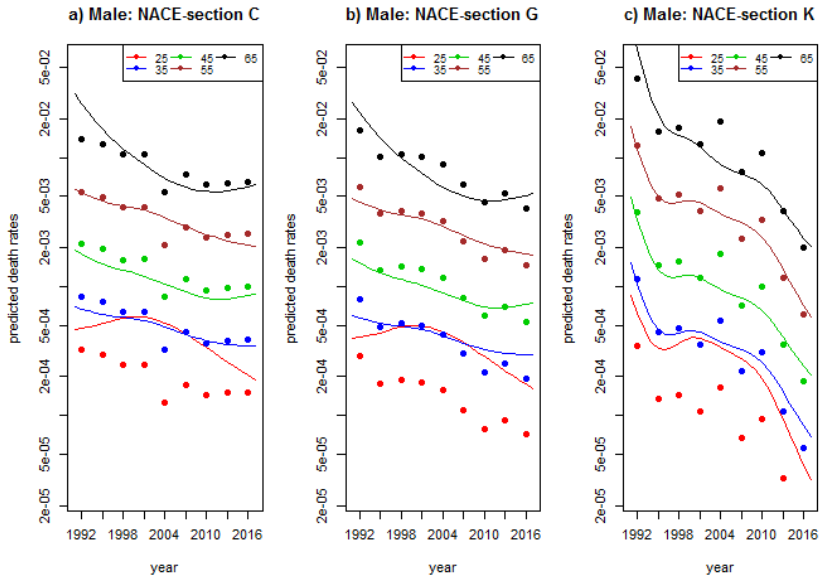


Figure 7.3: Predicted death rates for males of given age in given NACE-sections. Predictions are plotted on a log-scale over years by GLM (points) and GAM (lines). Each age is represented by a color; red: 25, blue: 35, green: 45, brown: 55 and black: 65. GLM predictions are made by model M3.nace.m and GAM predictions are made by model gam.M4.nace.m. Each NACE-section is given a panel in the figure. Panel a): NACE-section C - Manufacturing, panel b): NACE-section G - Wholesale and retail trade; repair of motor vehicles and motorcycles, panel c): NACE-section K - Financial and Insurance Activities. As the GLMs do not have single years as covariates in the models but instead have grouped years, we have put the predictions in the year which is in the middle of the group, eg. 1991-1993 $\rightarrow$ 1992.

throughout most of the year span. They also agree on a faster decreasing death rate trend over years for NACE-section K than the two other NACE-sections. As for the simpler models, the GLM predictions in each NACE-section have greater differences from year to year, than the GAM.

An improvement in smoothed death rate predictions over years for GLMs when using grouped years only, and not single years in combination with grouped years as in the previous section, is clear. We get death rate predictions in year 2004 which are much closer to the predictions in other years. The predictions in year 2004 does however still stick out from the other predictions, especially in NACE-sections C and K.

For all NACE-sections the GAM predicts less difference in death rates between males of age 25 and males of age 35 than the GLM. In NACE-sections C and G, from year 1998 to 2007, the GAM even predicts a higher death rate for males of age 25 than males of age 35. For most of the tested ages, in all NACE-sections, the predictions done by the GLM twirls around the prediction lines of the GAM. For the youngest males however the GAM predicts death rates for those of age 25 that are around the same level as the GLM predictions for those of age 35.

If we instead look at male GLM and GAM predictions over age for each NACE-section, Figure 7.4, we see that the higher predicted death rates for

the youngest group in Figure 7.3, may be due to the dip in death rates from age 20 to around 30 in all NACE-sections, predicted by the GAM in 2005. This dip straightens out and turns more and more to a slower ascending curve for the younger population in years moving away from 2005. We know 2005 was a special year, as it is the year after Storebrand and If split up. With a linear age variable, as in the GLM here, we do not get this dip for 2005. We do however also miss out on the slower ascending death rate curves for the youngest population in other years.

The GLM and GAM predictions in Figure 7.4 both have greater differences in death rates from year 2005 to 2017 across the age groups in NACE-section K, than in the two other NACE-sections. The GAM predictions in NACE-sections C and G have the same shape, there is however a subtle constant difference between the two, where NACE-section C is predicted to have higher death rates than NACE-section G.

For the male GLM we had interaction effects between NACE-sections and year groups, as well as NACE-sections and age. Hence we may observe differences in the development of predictions over age and year for all NACE-sections, not just NACE-section K, which we do for the GAM predictions. The GLM predicts less difference in death rates between each of the tested years in NACE-section C than in the two other NACE-sections. In NACE-section C the death rates are predicted to stay approximately the same from 2005 to 2017, whereas in NACE-sections G and K the death rates over age are predicted to decrease.

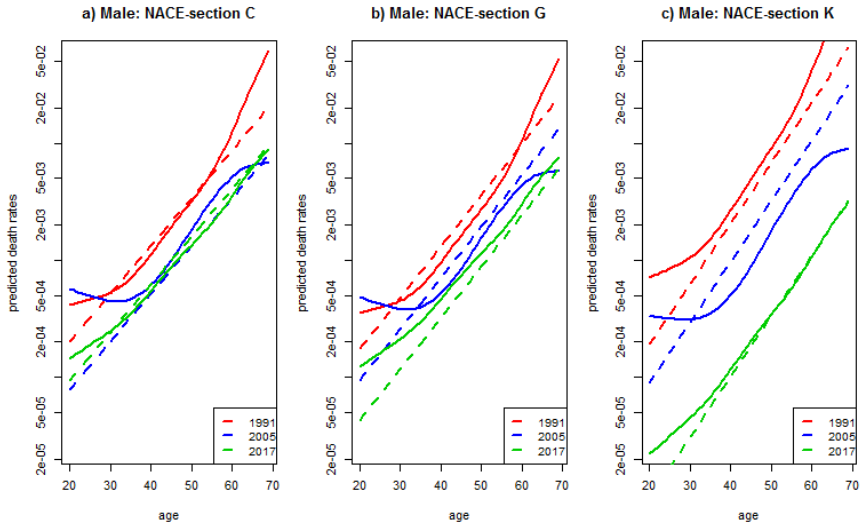


Figure 7.4: Predicted death rates for males in given years in given NACE-sections. Predictions are plotted on a log-scale over age by GLM (stippled lines) and GAM (solid lines). Each year is represented by a color; red: 1991, blue: 2005, green: 2017. GLM predictions are made by model M3.nace.m and GAM predictions are made by model gam.M4.nace.m. Each NACE-section is given a panel in the figure. Panel a): NACE-section C - Manufacturing, panel b): NACE-section G - Wholesale and retail trade; repair of motor vehicles and motorcycles, panel c): NACE-section K - Financial and Insurance Activities. As the GLMs do not have single years as covariates, but use grouped years, we have used the corresponding year group to make the predictions for each year; 1991: 1991-1993, 2005: 2003-2005, 2017: 2015-2017.

For the female GLM and GAM we have chosen to only look at the prediction plots of two NACE-sections, the reference section C and NACE-section K, the section with the biggest difference from the reference. We will not look at the prediction plots of NACE-section G, as the prediction plots look almost identical to those of NACE-section C, they are however attached in Appendix B, p.113 for those that want to take a closer look.

We make GLM and GAM predictions for the female population by using M1.nace.f and gam.M1.nace.f respectively. This is done in the same way as we did for the male population. Looking at the female predictions over both age and year, Figure 7.5, we have less differences between the GLM and GAM predictions for the female population, than what we had for males. We also have smoother GLM predictions over years. The GLM predictions in year 2004 do not stand out as clear as they did in the male predictions.

Over age, Figure 7.5 a)-b), both the GLM and GAM predict higher death rates for NACE-section K, than for the other NACE-sections[1]. For most ages the GLM predictions twirls around the GAM prediction lines. For those of age 25 however, we see, as we did in the male predictions, that the GAM predictions are higher than the GLM predictions.

If we look at the female GAM predictions over age at given years, Figure 7.5 c)-d), we have a slower ascending curve for the youngest, under 30, and the oldest females, 60+, compared to those in the ages in between. It is also for the youngest and oldest female population that we see the clearest difference in GLM and GAM predictions. The GLM predicts almost identical death rates in years 1991 and 2005. The GAM predictions however indicate a decrease in death rates across age from 1991 to 2005. Both GLM and GAM predict a decrease in death rate across age for the female population from year 2005 to 2017. Comparing NACE-sections for the female population, NACE-section K is predicted to have the highest death rate for any given age at any given year according to both GLM and GAM.

The GLMs and GAMs fitted in this thesis are overall both agreeing and disagreeing in their predictions. What they do agree on is that death rates should decrease with years, and that the fastest decreasing death rates are those of males in NACE-section K. In which pace the death rates should decrease from one year to the next however, the GLMs and GAMs disagree on. Independently of age, gender and NACE-section GLMs on average predicts a 2.7% slower decrease in death rates per year than the corresponding GAMs do[2].

All GLMs and GAMs predict an increased death rate with increasing age, and they agree that the steepest death rate curve over age is for males in NACE-section K. They also agree that for the female population the highest death rates over age are found in NACE-section K. Again the pace in which the death rates changes differs between GLMs and GAMs. Independently of year, gender and NACE-section the GLMs on average predicts a 2% higher increase in death rates for a one year increase in age than the corresponding

[1]There is a subtle constant difference between NACE-section C and NACE-section G, where NACE-section G has lower death rates. If NACE-section K has higher death rates than NACE-section C, we therefore say that NACE-section K has higher death rates than the other sections.

[2]This regards GLMs and GAMs with age, year and NACE-sections as covariates. Code used and an explanation of how we found this difference is given in Appendix A, p.98

GAM predictions do[3].

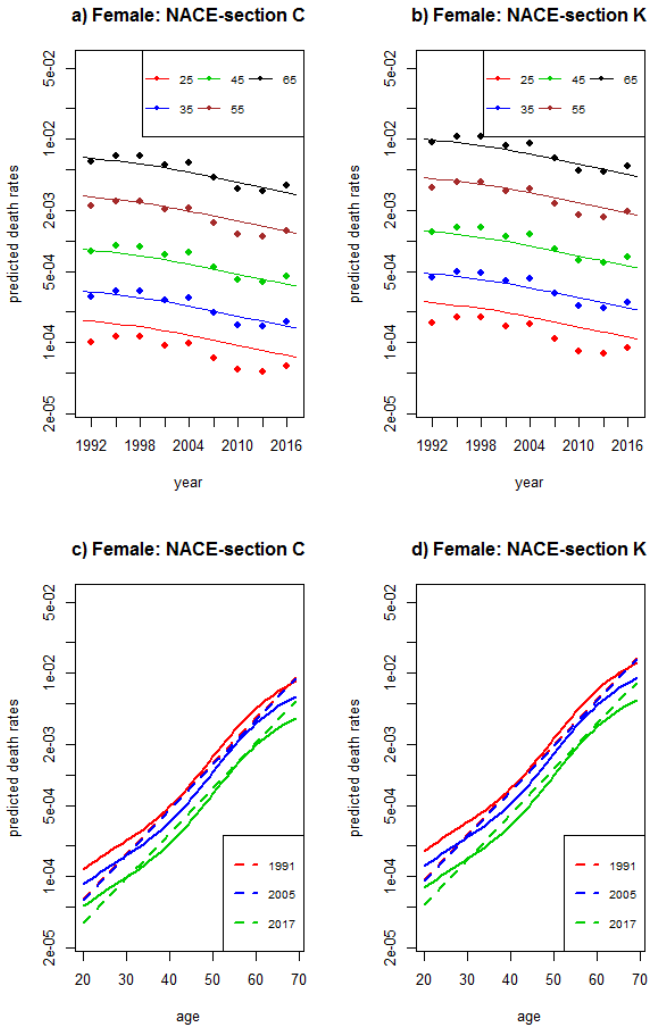


Figure 7.5: Predicted death rates in given NACE-sections for females of given age, panels a)-b), and given years, panels c)-d). Predictions are plotted on a log-scale by GLM (stippled lines) and GAM (solid lines). Each age in panels a) and b) is represented by a color; red: 25, blue: 35, green: 45, brown: 55 and black: 65. Each year in panels c) and d) is given colors; red: 1991, blue: 2005, green: 2017. GLM predictions are made by model M1.nace.f and GAM predictions are made by model gam.M1.nace.f. Each NACE-section is given a panel in the figure. Panel a) and c): NACE-section C - Manufacturing, panel b) and d): NACE-section K - Financial and Insurance Activities.

[3]We have used the same method to find the difference per age, as we did per year. See Appendix A, p.98

7.3 Overview of analysis

This thesis was intended to get a better understanding of which factors increase and decrease the risks an insurance company takes when covering death compensations for employees insured for death due to other illnesses. We also wanted to find good methods of smoothing the non-linear patterns we saw in our death rate observations, at the same time as we wanted to explain the variations we observed.

In Chapter 2 we looked at the costumer properties which were available in our data set. We split up the death rates according to different costumer properties to see if we there were any differences in death rates over given properties, which there were. We then considered some of these properties as covariates in Poisson GLM in Chapter 3.

Throughout Chapter 3 we mostly based the choice of model on likelihood ratio testing. For the bigger and more complex models however, involving NACE-sections, we also looked at AIC and BIC when the p-value of the likelihood ratio test (LRT) was close to the 0.05 threshold. If we instead would have based the model choice purely on LRT we would have ended up with a more complex model for the male population. For the female population however, the model choice would have been the same, regardless of basing the model choice on LRT, AIC or BIC. We have less data for the female population. Less data may lead to less significance, as there is more room for random variation, so a less complex model for the female population is not surprising.

In Chapter 4 we checked if we had overdispersed data and if a Quasi-Poisson model or a Negative Binomial Distribution had a better fit than the models with an assumption of Poisson distributed death. Some of the results we found in the Chapter 4 may indicate mild overdispersion compared to Poisson, but the main picture of the Poisson dispersion tests and the alternative models fitted was that there were no clear signs of overdispersion. We therefore concluded that we could continue with an assumption of Poisson distributed data.

In Chapter 5 we took a deeper look at an alternative way of smoothing nonlinear patterns through splines. In Chapter 3 we tried to smooth out non-linearities by categorisation of intervals of the numerical variables age and year. This worked to some degree, but as the GLM predictions in Section 7.2 showed, the models were still sensitive to single observations of low death rates, e.g female death rate predictions in age group 30-39 Figure 7.2. In Chapter 5 we saw that splines was a good method of smoothing nonlinearities within a parameter space, but that curves can be unstable at the edges of the parameter space.

In Chapter 6 we used splines in GAMs fitted to our data. These models gave a better representation of the observations than the corresponding GLMs. The GAMs were also less sensitive to single observations of low death rates. The problems we had with year 2004 in the GLM predictions, was not seen in the GAM predictions.

7.4 Discussion

Smoothing non-linear patterns in GLMs using categorical variables worked to some extent, but it did not smooth out the most extreme observations, like year 2004. Using numerical age for the GLMs did smooth over the most extreme observations, but we missed out on some of the observed trends for the youngest and oldest population.

Having models with splines, give a good representation of the death rate observations. We do however know that splines are unstable in the edges of the parameter space. Outside of the variable span, splines turn to linear trends. For the female population we have no interaction effects in the model, the splines are also close to being linear over years. Going beyond our boundary year 2017 will therefore not make major changes to our predictions. For the male population however the years leading up to year 2017 will have a big influence on the death rates predicted for the years ahead in time.

We have made predictions from 2005 to 2025 for NACE-sections C and K in Figure 7.6. The figure illustrates well the possible challenges with future predictions using GAM. For the female model, with close to linear splines and no interaction terms, Figure 7.6 b), there is no major changes in predictions before and after 2017. In the male predictions however, Figure 7.6 a), the small increase in death rate differencies between NACE-sections in the years leading towards 2017, leads to huge differences in 2025. In NACE-section C, the solid lines in Figure 7.6 a), we even get an increased death rate for all ages after year 2017. Using GAMs for future predictions must therefore be done with care, especially when dealing with models with interaction effects and little smoothing. Predictions for a short time ahead is therefore preferred and to get adequate predictions the models should be updated frequently.

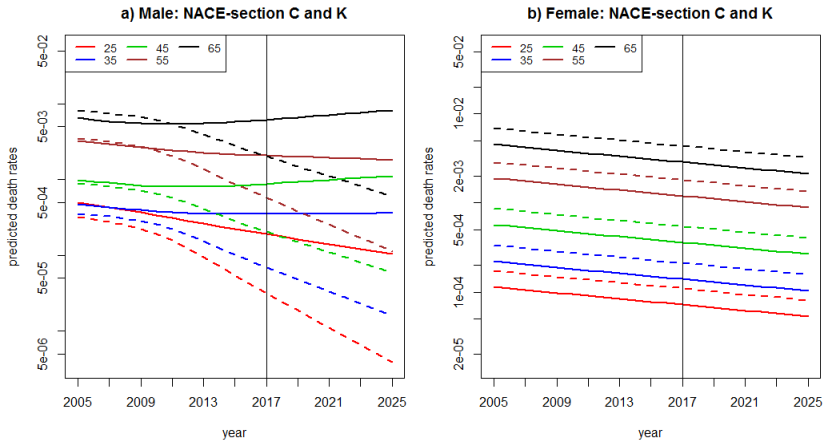


Figure 7.6: Predicted death rates in NACE-section C (solid lines) and NACE-section K (stippled lines) for males, panel a), and females, panel b), of given age over years. Predictions are plotted on a log-scale and made by GAMs gam.M4.nace.m for males and gam.M1.nace.f for females. Each age in both panels is represented by a color; red: 25, blue: 35, green: 45, brown: 55 and black: 65. A vertical black line is drawn in the boundary year, 2017, in both panels.

Both GLM and GAM agreed that NACE-section K for women had the highest death rate for a given age. This is a bit odd, considering that this NACE-section involves jobs such as banking, fund management and insurance pricing which often require higher education. Borgan (2009) states that people with higher education have the lowest mortality and that the largest distance to other occupations are found for men. We saw this for our most recent years in the GAM predictions for males. For the female population however we have higher death rates for NACE-section K than we have for the two other sections C and G. These sections mostly involve jobs that require less education, such as production and sale of food, drinks and textiles. Hypothetically we would therefore think we would have lower death rates for NACE-section K than NACE-sections C and G.

A reason that we have higher death rates for females in NACE-section K, may be due to something known as the healthy workers effect. The people that work have lower mortality than the average population due to the fact that people, such as those with major disabilities or those that are really sick, are being excluded from the job marked (Braut 2018). This may however lead to a bias when comparing different occupations. The effect is clearest for jobs where you need higher qualification and if we compare genders, the effect is more clear for women than for men (Shah 2009).

Workers with less job motivation due to reasons such as health may change jobs more frequently or retire earlier than others, they may also leave earlier if they get a deadly illness. The people that are still at work in the same jobs however then have a lower death rate compared to jobs where people do not leave as easy. Occupations with low education requirements have had a decreasing proportion of active population for every year after 1960[4], in the occupations with higher education requirements however the proportion has increased with years (Borgan 2009). Hence, the women in NACE-section K may not die more frequently than the other women in general, but they may have a higher proportion of people who die while still active in work.

7.5 Challenges and further work

Dealing with a big data set can lead you in many directions and when handling it, things take time and can easily go wrong. Our original data set had 783749 observations. Making categorical variables out of numeric intervals or making larger categorical groups, as we did for country region, therefore took a lot of time. The data had to be aggregated to a smaller number of observations, both due to running time and to get a better understanding of the data as a whole.

Despite aggregating the data, we had many residuals in our residual plots and the plots got difficult to interpret. The low number of deaths in each group, as we added more variables lead to residuals with a distribution which looked less and less normal. This also made the model choices more difficult, as we had a high number of residuals and the residual plots had little differences.

Throughout the thesis we considered missing exposures to be missing completely at random, meaning that the exposures are missing independently of variables (Hastie, Tibshirani, and Friedman 2009, pp. 332–333). As the missing exposure tables may indicate however, there are tendencies of some groups

[4]The article used as reference was published in 2009

missing more exposures than others. We have, for example, seen that there are more missing exposures for the youngest (< 25) and oldest population (60+) than rest. With more time available it would therefore be interesting to look at the assumption of data missing completely at random.

With more time I would also have liked to check the effect of adding more NACE-sections as categorical variables. I would also have liked to check the effect of fitting models to data without including the observations with estimated age, to see if the model predictions would have been different. Adding country region in which a company is stationed to see if it has a significant effect on the death rates would also have been interesting to look at.

Appendices

APPENDIX A

Calculations

A.1 Constructed data in Chapter 5

In Chapter 5 we look at smoothing of nonlinear data using splines. Throughout the chapter we use constructed data. This data was constructed in R using the following code:

```
> z = seq(0,6,0.1)
> y = sin(z) + (z-2.5)^3 + sqrt(z)/2 + cos(rnorm(length(z)))*10-
  rnorm(length(z))*5 - I(z>4.5)*rnorm(length(z))*3+
  I(z>2.5)*I(z<4.5)*(1-rnorm(length(z))*3)
```

y is used as data observations in all examples of Chapter 5.

A.2 B-spline calculation example

We will illustrate how B-splines are calculated by showing an example of B-splines with two-internal knots and two boundary knots. Throughout this example Hastie, Tibshirani, and Friedman (2009, p.186–187) is used as reference. In this thesis we use B-splines of order 4, hence cubic polynomials going between the knots. For simplicity however, we here show how to find B-splines up to order 3 where we get 2. degree polynomials going between the knots.

Given knots $(\xi_1, \xi_2, \xi_3, \xi_4) = (0, 3, 4, 6)$ we have $N = 2$ inner knots and 2 boundary knots. We want B-splines of order 3, hence polynomials of degree 2 between each of these knots. To achieve this the boundary knots must be repeated 3 times. In general, wanting splines of order K, the boundary knots must be repeated K times. In our example we get the augmented knot sequence, $(\tau_1, \tau_2, \tau_3, \tau_4, \tau_5, \tau_6, \tau_7, \tau_8) = (0, 0, 0, 3, 4, 6, 6, 6)$ which will give us 5 B-spline functions of order 3. In general, the number of B-spline functions, M, is given as $M = K + N$, where K is the order of the spline and N is the number of inner knots.

We first calculates the splines of order $k = 1$. The number of B-splines calculated for order $k < K$ is equal to $N + 2K - k$. In our example it means we end up with $2 + 6 - 1 = 7$ functions of order 1. We get 4 functions which, by equation (5.3), are always equal to 0; $B_{1,1}(z)$, $B_{2,1}(z)$, $B_{6,1}(z)$ and $B_{7,1}(z)$ as $\tau_j = \tau_{j+1}$ for these functions. We find $B_{3,1}(z)$, $B_{4,1}(z)$ and $B_{5,1}$ by using

equation (5.3):

$$B_{3,1} = \begin{cases} 1 & \text{if } 3 \leq z < 4 \\ 0 & \text{else} \end{cases}$$
$$B_{4,1} = \begin{cases} 1 & \text{if } 3 \leq z < 4 \\ 0 & \text{else} \end{cases}$$
$$B_{5,1} = \begin{cases} 1 & \text{if } 4 \leq z < 6 \\ 0 & \text{else} \end{cases}$$

We can then find the splines of order $k = 2$ by, using equation (5.4):

$$\begin{aligned}
B_{1,2} &= W_{1,2}(z)B_{1,1}(z) + (1 - W_{2,2}(z))B_{2,1}(z) \\
&= 0 \\
B_{2,2} &= W_{2,2}(z)B_{2,1}(z) + (1 - W_{3,2}(z))B_{3,1}(z) \\
&= \begin{cases} (3-z)/3 & \text{if } 0 \leq z < 3 \\ 0 & \text{else} \end{cases} \\
B_{3,2} &= W_{3,2}(z)B_{3,1}(z) + (1 - W_{4,2}(z))B_{4,1}(z) \\
&= \begin{cases} z/3 & \text{if } 0 \leq z < 3 \\ 4-z & \text{if } 3 \leq z < 4 \\ 0 & \text{else} \end{cases} \\
B_{4,2} &= W_{4,2}(z)B_{4,1}(z) + (1 - W_{5,2}(z))B_{5,1}(z) \\
&= \begin{cases} z-3 & \text{if } 3 \leq z < 4 \\ (6-z)/2 & \text{if } 4 \leq z < 6 \\ 0 & \text{else} \end{cases} \\
B_{5,2} &= W_{5,2}(z)B_{5,1}(z) + (1 - W_{6,2}(z))B_{6,1}(z) \\
&= \begin{cases} (z-4)/2 & \text{if } 4 \leq z < 6 \\ 0 & \text{else} \end{cases} \\
B_{6,2} &= W_{6,2}(z)B_{6,1}(z) + (1 - W_{7,2}(z))B_{7,1}(z) \\
&= 0
\end{aligned}$$

Then using $B_{j,2}$ for $j = 1, ..., 6$ we find our five 2. degree polynomials:

$$\begin{aligned}
B_{1,3} &= W_{1,3}(z)B_{1,2}(z) + (1 - W_{2,3}(z))B_{2,2}(z) \\
&= \begin{cases} (3-z)^2/9 & \text{if } 0 \leq z < 3 \\ 0 & \text{else} \end{cases} \\
B_{2,3} &= W_{2,3}(z)B_{2,2}(z) + (1 - W_{3,3}(z))B_{3,2}(z) \\
&= \begin{cases} \frac{-7z^2+24z}{36} & \text{if } 0 \leq z < 3 \\ (4-z)^2/4 & \text{if } 3 \leq z < 4 \\ 0 & \text{else} \end{cases} \\
B_{3,3} &= W_{3,3}(z)B_{3,2}(z) + (1 - W_{4,3}(z))B_{4,2}(z) \\
&= \begin{cases} z^2/12 & \text{if } 0 \leq z < 3 \\ \frac{-7z^2+48z-72}{12} & \text{if } 3 \leq z < 4 \\ (6-z)^2/6 & \text{if } 4 \leq z \leq 6 \\ 0 & \text{else} \end{cases}
\end{aligned}$$

$$
\begin{aligned}
B_{4,3} &= W_{4,3}(z)B_{4,2}(z) + (1 - W_{5,3}(z))B_{5,2}(z) \\
&= \begin{cases} (z-3)^2/3 & \text{if } 3 \leq z < 4 \\ \frac{-5z^2+48z-108}{12} & \text{if } 4 \leq z \leq 6 \\ 0 & \text{else} \end{cases} \\
B_{5,3} &= W_{5,3}(z)B_{5,2}(z) + (1 - W_{6,3}(z))B_{6,2}(z) \\
&= \begin{cases} (z-4)^2/4 & \text{if } 4 \leq z \leq 6 \\ 0 & \text{else} \end{cases}
\end{aligned}
$$

Plots in Figure A.1 a) and b) show the splines of lower order, used to reach the goal of B-splines of order 3, plot c) of the same figure. The B-splines of Figure A.1 are plotted over z values ranging from 0 to 6 with steps of 0.1.

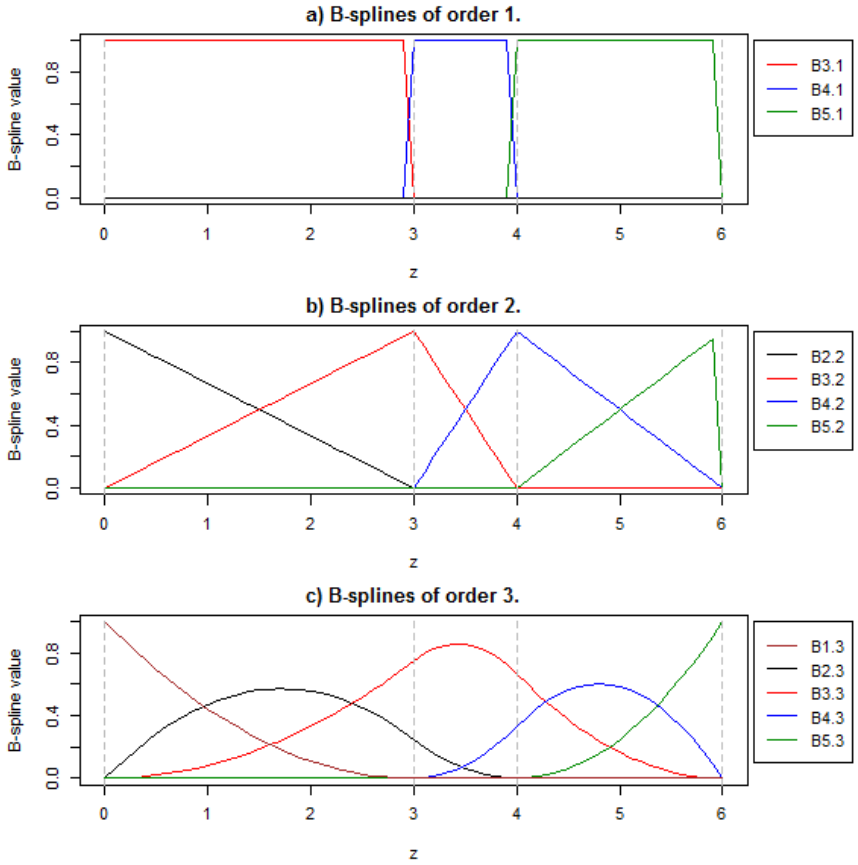


Figure A.1: B-spline exampe with boundary knots in $z = 0$ and $z = 6$, and two inner knots at $z = 3$ and $z = 4$. a) B-splines of order 1, b) B-splines of order 2, linear and c) B-splines of order 3, polynomial of degree 2. z has been given values from 0 to 6 with steps of 0.1

Running;

```
> z= seq(0,6,0.1)
> BS = bs(z, knots=(3,4), degree =2, intercept = T)
> plot(z,BS[,1])
> lines(z,BS[,2])
> lines(z,BS[,3])
> lines(z,BS[,4])
> lines(z,BS[,5])
```

in R yields the same results as in Figure A.1 c), returning splines $B_{j,3}$ for $j = 1, ..., 5$ in form of a matrix, where each spline corresponds to a column and each row correspond to the value of the spline at a given value of z. Results from R running `BS = bs(z, knots=(3,4), degree =2, intercept = T)` along with results found by hand, can be seen in table A.1. When fitting models it is normal to have the intercept outside the B-spline matrix. In these cases the first column is removed, and the other columns stays the same. In other words we end up with a B-spline matrix with $K + N - 1$ columns which including the intercept gives $K + N$ model parameters.

A.3 Trace of smoothing matrix equals degrees of freedom in fitted curve.

In Section 5.4 we look at the effect of changing λ in natural cubic smoothing splines. In this section we show how the curves in Section 5.4 was defined in R. We also show two different methods that can be used to find the effective degrees of freedom (edf) of a curve. We can find the edf by extracting curve information from the data frame of a given curve, or we can take the trace of the smoother matrix. Both methods should yield the same result.

We use the same values for z as earlier, hence $z = 0.0, 0.1, ..., 5.9, 6.0$ and y is the constructed data given in Section A.1. Code below is used in R to check that the trace of the smoother matrix is the same as degrees of freedom in a curve fit. The function smoother.matrix is originally defined by Wood (2006).

```
smoother.matrix = function(z=z,y=y,lambda=lambda) {
   S = matrix(nrow = length(z), ncol = length(z))
   for(i in 1:length(z)) {
       ym = rep_len(0, length(z)) ; ym[i] = 1
       S[,i] = predict(smooth.spline(z, ym, lambda = lambda, cv=TRUE,
               all.knots = TRUE), z)$y
    }
    return(S)
}

> lamb001 = smooth.spline(y, lambda =0.001 ,all.knots = T)
> lamb001$df
> 6.619974
> sum(diag(smoother.matrix(z,y,0.001)))
> 6.619974
```

Table A.1: Results of B-spine calculations done by R using the `bs( )` function and results found by doing calculations by hand. Each column, except the first in the table, correspond to a given B-spline and each row correspond to the value of the B-spline for a given value of x. Splines whose names start with R are R-results whilst those that have a name starting with H are results found by hand.

x-value	R.B1.3	H.B1.3	R.B2.3	H.B2.3	R.B3.3	H.B3.3	R.B4.3	H.B4.3	R.B5.3	H.B5.3
0.00	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.10	0.934444	0.934444	0.064722	0.064722	0.000833	0.000833	0.000000	0.000000	0.000000	0.000000
0.20	0.871111	0.871111	0.125556	0.125556	0.003333	0.003333	0.000000	0.000000	0.000000	0.000000
0.30	0.810000	0.810000	0.182500	0.182500	0.007500	0.007500	0.000000	0.000000	0.000000	0.000000
0.40	0.751111	0.751111	0.235556	0.235556	0.013333	0.013333	0.000000	0.000000	0.000000	0.000000
0.50	0.694444	0.694444	0.284722	0.284722	0.020833	0.020833	0.000000	0.000000	0.000000	0.000000
0.60	0.640000	0.640000	0.330000	0.330000	0.030000	0.030000	0.000000	0.000000	0.000000	0.000000
0.70	0.587778	0.587778	0.371389	0.371389	0.040833	0.040833	0.000000	0.000000	0.000000	0.000000
0.80	0.537778	0.537778	0.408889	0.408889	0.053333	0.053333	0.000000	0.000000	0.000000	0.000000
0.90	0.490000	0.490000	0.442500	0.442500	0.067500	0.067500	0.000000	0.000000	0.000000	0.000000
1.00	0.444444	0.444444	0.472222	0.472222	0.083333	0.083333	0.000000	0.000000	0.000000	0.000000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.50	0.250000	0.250000	0.562500	0.562500	0.187500	0.187500	0.000000	0.000000	0.000000	0.000000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2.00	0.111111	0.111111	0.555556	0.555556	0.333333	0.333333	0.000000	0.000000	0.000000	0.000000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2.50	0.027778	0.027778	0.451389	0.451389	0.520833	0.520833	0.000000	0.000000	0.000000	0.000000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.00	0.000000	0.000000	0.250000	0.250000	0.750000	0.750000	0.000000	0.000000	0.000000	0.000000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.50	0.000000	0.000000	0.062500	0.062500	0.854167	0.854167	0.083333	0.083333	0.000000	0.000000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4.00	0.000000	0.000000	0.000000	0.000000	0.666667	0.666667	0.333333	0.333333	0.000000	0.000000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4.50	0.000000	0.000000	0.000000	0.000000	0.375000	0.375000	0.562500	0.562500	0.062500	0.062500
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5.00	0.000000	0.000000	0.000000	0.000000	0.166667	0.166667	0.583333	0.583333	0.250000	0.250000
5.10	0.000000	0.000000	0.000000	0.000000	0.135000	0.135000	0.562500	0.562500	0.302500	0.302500
5.20	0.000000	0.000000	0.000000	0.000000	0.106667	0.106667	0.533333	0.533333	0.360000	0.360000
5.30	0.000000	0.000000	0.000000	0.000000	0.081667	0.081667	0.495833	0.495833	0.422500	0.422500
5.40	0.000000	0.000000	0.000000	0.000000	0.060000	0.060000	0.450000	0.450000	0.490000	0.490000
5.50	0.000000	0.000000	0.000000	0.000000	0.041667	0.041667	0.395833	0.395833	0.562500	0.562500
5.60	0.000000	0.000000	0.000000	0.000000	0.026667	0.026667	0.333333	0.333333	0.640000	0.640000
5.70	0.000000	0.000000	0.000000	0.000000	0.015000	0.015000	0.262500	0.262500	0.722500	0.722500
5.80	0.000000	0.000000	0.000000	0.000000	0.006667	0.006667	0.183333	0.183333	0.810000	0.810000
5.90	0.000000	0.000000	0.000000	0.000000	0.001667	0.001667	0.095833	0.095833	0.902500	0.902500
6.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000

```
> lamb010 = smooth.spline(y, lambda =0.01 ,all.knots = T)
> lamb010$df
> 4.161761
> sum(diag(smoother.matrix(z,y,0.01)))
> 4.161761

> lamb100 = smooth.spline(y, lambda =0.1 ,all.knots = T)
> lamb100$df
> 2.780111
> sum(diag(smoother.matrix(z,y,0.1)))
> 2.780111
```

A.4 GCV-trace

In Section 5.4 we explain how the optimal λ can be found for a natural cubic smoothing spline through generalized cross-valiadation (GCV). In Figure 5.6, p. 50, there are two plots showing the GCV-trace for corresponding λ- and edf-values . The code to get the GCV-trace and make Figure 5.6 follows below.

```
> lambda = 1e-6
> gcv = numeric(length = 90)
> df = numeric(length = 90)

# Function for calculating GCV at different lambda values
> for (i in seq_along(gcv)) {
>     m <- smooth.spline(x,y,lambda = lambda, all.knots = T)
>     gcv[i] <- m$cv.crit
>     df[i] = m$df
>     lambda <- lambda * 1.1 }

# Gather results
> par(mfrow=c(1,2))
> res <- data.frame(lambda = 1e-6 * 1.1^(0:89), gcv = gcv)

# Plot of results over lamda
> plot(res[which(res[,1] >= 0.00015),1], gcv[which(res[,1] >= 0.00015)],
  ylim=c(56,57.5), xlim=c(0.00015,1e-6 * 1.1^(89)), ylab = "GCV",
  xlab= expression(lambda), pch=21, bg="azure3", cex=1.25)
> abline(v=res[which(gcv == min(gcv)),1], lty=2,col="azure4",lwd=2)
> points(res[which(gcv == min(gcv)),1],min(gcv),pch=21,col="darkgreen",
  bg="green2",cex=1.25)

# Plot of results over effective degrees of freedom
> plot(df[which(res[,1] >= 0.00015)], gcv[which(res[,1] >= 0.00015)],
  ylim=c(56,57.5), ylab = "GCV", xlab="edf", pch=21, bg="azure3", cex=1.25)
> abline(v=df[which(gcv == min(gcv))], lty=2,col="azure4",lwd=2)
> points(df[which(gcv == min(gcv))],min(gcv),pch=21,col="darkgreen",
  bg="green2",cex=1.25)
> mtext("GCV trace over choice of lambda and edf in default-smoothed model.",
  side = 3, line = -2, outer = TRUE,cex=1.2)
```

A.5 Mean prediction differences across age and year for GLMs and GAMs

In Section 7.2 we say that GLMs, independent of age, gender and NACE-section, on average predict a 2% slower decrease in death rates from one year to the next compared with GAMs. We also say that the GLMs on average, independet of model covariates, predict a 2.7% greater death rate increase from one age to the next, compared with the corresponding GAM predictions. We here explain how we found these differences by using females in NACE-section G as an example.

For the female population, as we have main effects only in both the GLM and GAM, we will get the same result for each NACE-section.

We find the mean percentage change in death rate across all ages and years for females in in NACE-section G by using the following code, starting with the GLM predictions:

```
## GLM predictions ##
# Make a matrix for storage of predictions of each age within each year group
> female.all.glm.pred.G = matrix(0,nrow=9,ncol=50)
> dim(female.all.glm.pred.G) = c(9,50)
> for (i in 1:50){
>    female.all.glm.pred.G[,i]=exp(predict.glm(M1.nace.f, newdata=
>    data.frame(age=ages[i],yearGroup=year.groups,duration=1,NaceMain="G")))
> }
# Make vectors for storage of mean predictions across age and year groups
> mean.per.age.years.glm.G = matrix(0,nrow=50,ncol=1)
> mean.per.year.ages.glm.G = matrix(0,nrow=9,ncol=1)
# Vector of mean predictions for each age across year groups
> for (i in 1:50){
>    mean.per.age.years.glm.G[i] = mean(female.all.glm.pred.G[,i])}
# Vector of mean predictions for year group across ages
> for (i in 1:9){
>    mean.per.year.ages.glm.G[i] = mean(female.all.glm.pred.G[i,])}
# Make vectors for storage of percentage change in mean predictions of age and year groups
> diff.glm.per.age.g.f = NULL
> diff.glm.per.year.g.f = NULL
# Vector of percentage difference in mean predictions per age
> for (i in 1:49){
>    diff.glm.per.age.g.f[i] = mean.per.age.years.glm.G[i+1]/mean.per.age.years.glm.G[i]}
# Vector of percentage difference in mean predictions per year group
> for (i in 1:8){
>    diff.glm.per.year.g.f[i] = mean.per.year.ages.glm.G[i+1]/mean.per.year.ages.glm.G[i]}
```

We get the GAM predictions, using a similar code, but we use `predict.gam( )` instead `predict.glm( )`, we change the model from `M1.nace.f` to `gam.M1.nace.f` and the `yearGroup=year.groups` is replaced by `year=1991:2019`. The `for` loops for years are then changed to lengths of 27 (in the loop taking mean of predictions across ages) and 26 (in the loop taking the difference in average predictions). After running the code for both the GLM and GAM we find the difference through:

```
> female.diff.per.age = mean(diff.glm.per.age.g.f)/mean(diff.per.age.g.f)
> 1.015344
> female.diff.per.year = (1-(1-mean(diff.glm.per.year.g.f))/3)/mean(diff.per.year.g.f)
> 1.01384
```

We divide the GLM predictions by 3 for the differences over year. We do this to get the change per year, as the calculations done is for every third year. The result shows that the GLM death rate predictions on average per age increase $1.015 = 1.5\%$ faster than the predictions done by GAM. It also shows that the GLM death rate predictions on average per years decrease

$1.014 = 1.4\%$ slower than the predictions done by GAM. Doing the same for the male population we can the differences in predictions within each NACE-section, for each gender. We get the overall result by taking the average of the results of each gender, of each NACE-section.

APPENDIX B

Figures and tables.

B.1 Tables of missing exposures in Chapter 4

In Section 4.2 and 4.3 we group data according to age group, gender, year and NACE-section. This should leave us with a total of 2835 observations for each gender. This is however not the case, as we miss exposures for some of the variable combinations. We end up with a total of 2729 observations in the male data set and 2691 observations for the female data set. Overview of missing exposures are given in Table B.1 for the male population and in Table B.2 for the female population.

Table B.1: Overview of missing male exposures when grouping data by single years, NACE-section and and ten-year age groups in Section 4.2. Total number of missing exposures for the male population are 106.

Year	Number of missing exposures
1991	8
1992	5
1993	4
1994	4
1995	3
1996	1
1997	1
1999	5
2000	5
2001	5
2002	5
2003	6
2004	34
2005	5
2006	8
2007	1
2013	1
2017	5
Total	106

NACE-section	Number of missing exposures
A	1
B	1
D	5
E	6
I	1
N	1
O	53
P	6
Q	1
R	9
U	10
X	12
Total	106

Age group	Number of missing exposures
20-29	27
30-39	17
40-49	17
50-59	18
60-69	27
Total	106

Table B.2: Overview of missing female exposures when grouping data by single years, NACE-section and and ten-year age groups in Section 4.2. Total number of missing exposures for the female population are 144.

Year	Number of missing exposures
1991	10
1992	8
1993	6
1994	5
1995	3
1996	1
1997	2
1998	2
1999	5
2000	6
2001	5
2002	6
2003	7
2004	40
2005	9
2006	8
2007	1
2008	2
2009	1
2010	2
2011	2
2012	1
2013	2
2014	2
2015	2
2016	1
2017	5
Total	144

NACE-section	Number of missing exposures
A	7
B	7
D	6
E	25
F	1
I	3
L	1
N	1
O	56
P	8
R	11
U	11
X	7
Total	144

Age group	Number of missing exposures
20-29	29
30-39	20
40-49	19
50-59	27
60-69	49
Total	144

B.2 Tables of dispersion test carried out in Section 4.1

In Section 4.1 we explain how we can check for overdisepersed data by performing a poisson dispersion test on non-aggregated data. We do this by making test groups with age within a given range with approximately equal exposure ranges in the male and female data sets. The test results, not given in Section 4.1, for the male population is given in Table B.3. Test results for the female population is shown in Table B.4.

Table B.3: Summary tables of Poisson Dispersion Tests done on non-aggregated data for males in given age groups, year 1998. D = the test statistic defined in equation (4.1), page 33. df = the degrees of freedom for each test statistic. A line in the table-cell means no deaths have been observed for the belonging group.

Age group: 20-29 years					
Exposure	mean	variance	D	df	p value
Person years = 1	0.0021	0.0035	2399	1440	0.000 ***
1< Person years <2	0.0012	0.0012	863	863	0.494
Person years = 2	0.0027	0.0027	366	366	0.490
2< Person years ≤3	0.0020	0.0020	509	509	0.492
3< Person years ≤4	0.0037	0.0037	270	270	0.489
4< Person years ≤5	0.0065	0.0065	152	152	0.485
5< Person years ≤6	0.0000	0.0000	-	113	-
Age group: 30-39 years					
Exposure	mean	variance	D	df	p value
Person years = 1	0.0000	0.0000	-	2690	-
1< Person years <2	0.0010	0.0010	954	954	0.494
Person years = 2	0.0000	0.0000	-	855	-
2< Person years ≤3	0.0025	0.0025	805	806	0.503
3< Person years ≤4	0.0000	0.0000	-	393	-
4< Person years ≤5	0.0044	0.0044	226	226	0.487
5< Person years ≤6	0.0065	0.0065	153	153	0.485
Age group: 40-49 years					
Exposure	mean	variance	D	df	p value
Person years = 1	0.0000	0.0000	-	3051	-
1< Person years <2	0.0096	0.0095	621	626	0.549
Person years = 2	0.0000	0.0000	-	822	-
2< Person years ≤3	0.0017	0.0017	585	585	0.492
3< Person years ≤4	0.0099	0.0099	299	301	0.522
4< Person years ≤5	0.0098	0.0098	202	203	0.507
5< Person years ≤6	0.0472	0.0454	121	126	0.609
Age group: 60-69 years					
Exposure	mean	variance	D	df	p value
Person years = 1	0.0009	0.0009	1077	1077	0.494
1< Person years <2	0.0465	0.0447	123	128	0.608
Person years = 2	0.0126	0.0125	235	237	0.525
2< Person years ≤3	0.0263	0.0258	148	151	0.554
3< Person years ≤4	0.0400	0.0389	72	74	0.544
4< Person years ≤5	0.0444	0.0434	43	44	0.514
5< Person years ≤6	0.0000	0.0000	-	31	-

Table B.4: Summary tables of Poisson Dispersion Tests done on non-aggregated data for females in given age groups, year 1998. D = the test statistic defined in equation (4.1), page 33. df = the degrees of freedom for each test statistic. A line in the table-cell means no deaths have been observed for the belonging group. There are noe observed deaths of females in year 1998 in age group 20-29, this age group is therefore not included in the table.

Age group: 30-39 years					
Exposure	mean	variance	D	df	p value
Person years = 1	0.0005	0.0005	1998	1998	0.496
1< Person years <2	0.0023	0.0023	436	436	0.491
Person years = 2	0.0000	0.0000	-	452	-
2< Person years ≤3	0.0000	0.0000	-	345	-
3< Person years ≤4	0.0057	0.0057	173	173	0.486
4< Person years ≤5	0.0000	0.0000	-	122	-
5< Person years ≤6	0.0000	0.0000	-	66	-
Age group: 40-49 years					
Exposure	mean	variance	D	df	p value
Person years = 1	0.0005	0.0005	1887	1887	0.496
1< Person years <2	0.0000	0.0000	-	261	-
Person years = 2	0.0000	0.0000	-	351	-
2< Person years ≤3	0.0037	0.0037	270	270	0.489
3< Person years ≤4	0.0000	0.0000	-	140	-
4< Person years ≤5	0.0175	0.0175	56	56	0.475
5< Person years ≤6	0.0196	0.0196	50	50	0.473
Age group: 50-59 years					
Exposure	mean	variance	D	df	p value
Person years = 1	0.0006	0.0006	1591	1591	0.495
1< Person years <2	0.0115	0.0114	172	173	0.507
Person years = 2	0.0114	0.0113	346	349	0.535
2< Person years ≤3	0.0048	0.0048	206	206	0.487
3< Person years ≤4	0.0118	0.0118	84	84	0.479
4< Person years ≤5	0.0000	0.0000	-	54	-
5< Person years ≤6	0.0270	0.0270	36	36	0.469
Age group: 60-69 years					
Exposure	mean	variance	D	df	p value
Person years = 1	0.0000	0.0000	-	552	-
1< Person years <2	0.0208	0.0208	47	47	0.473
Person years = 2	0.0090	0.0090	110	110	0.482
2< Person years ≤3	0.0000	0.0000	-	47	-
3< Person years ≤4	0.0000	0.0000	-	17	-
4< Person years ≤5	0.0000	0.0000	-	13	-
5< Person years ≤6	0.0000	0.0000	-	3	-

B.3 Residual diagnostic plots of models fitted in Section 6.4 and 6.5

In Section 6.4 we fit GAMs with smoothed main effects for age and year. The models are fitten on datasets of 1350 observations each, one data set per gender. The resulting male model is called m.ay.mod and the resulting female model is called f.ay.mod. The residual diagnostic plots of these models are given in Figure B.1.

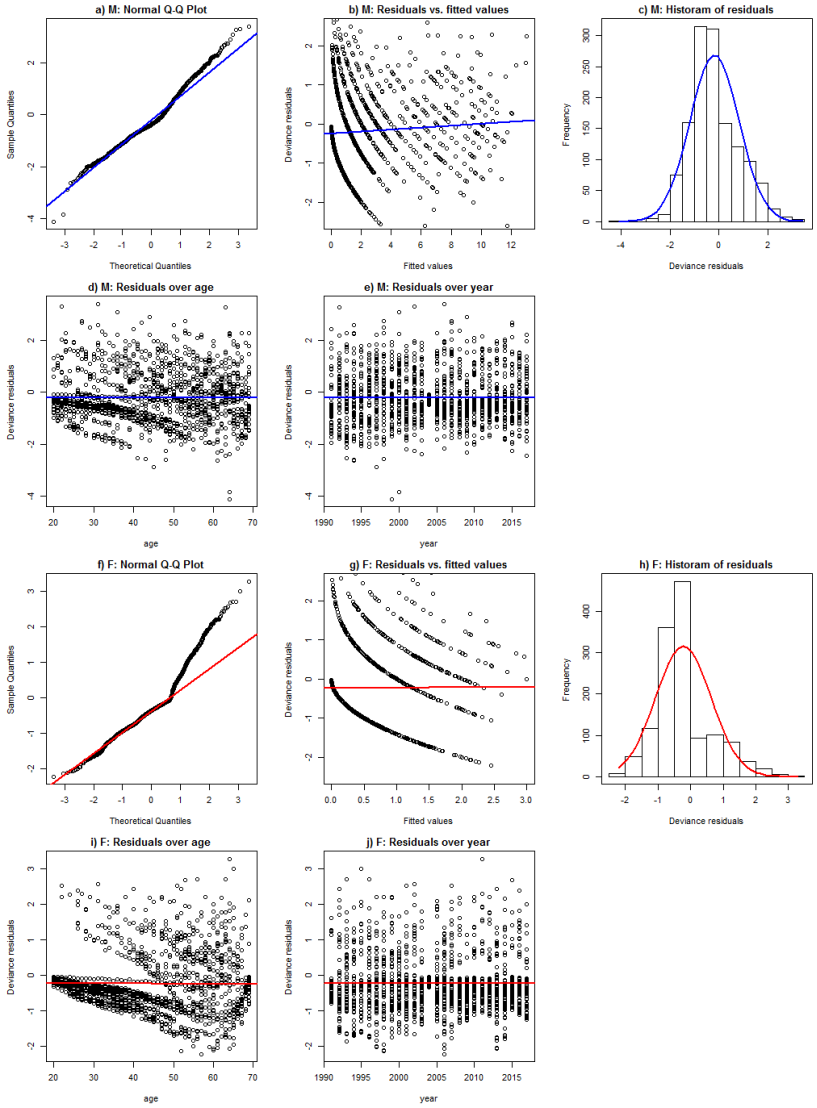


Figure B.1: Diagnostic plots of deviance residuals for the male and female models, m.ay.mod and f.ay.mod, default smoothed over age and year in chapter 6.3, page 59. a) Male QQ-plot of residuals, b) Male fitted values vs. residuals shown as points plotted with trendline, c) Male histogram of residuals with a normal line plotted on top, d) Male residuals over age, e) Male residuals over year, f)-j) same as a)-e) but for the female model.

In Section 6.5 we added a smoothed interaction term between age and year, using a tensor product smoother. The interaction effect was specified in the model by function `ti()` in R. The resulting male and female models were called m.ay.ten.mod and model f.ay.ten.mod respectively. The residual diagnostic plots of model m.ay.ten.mod is given in Figure B.2. The residual diagnostic plots of model f.ay.ten.mod is given in Figure B.3.

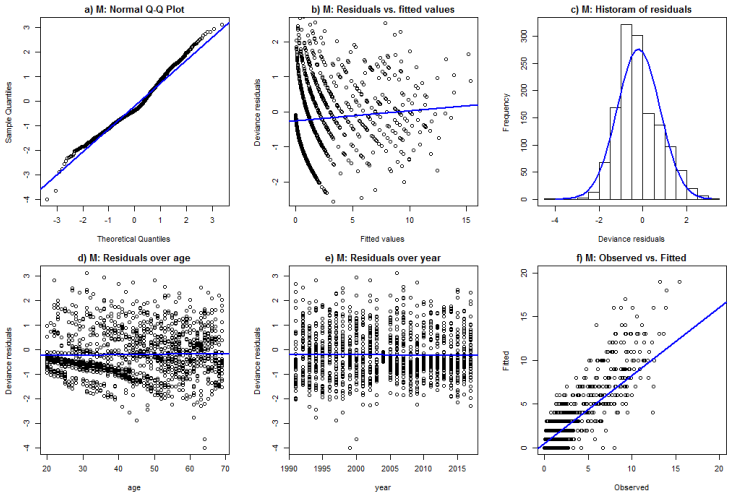


Figure B.2: Residual diagnostics and oberved vs. fitted values of GAM model m.ay.ten.mod, model with smoothed interaction terms and smooted main effects splitted. a) QQ-plot of residuals, b) Fitted values vs. residuals shown as points plotted with trendline, c) Histogram of residuals with a normal line plotted on top, d) Residuals over age with trendline, e) Residuals over year with red trendline, f) Observed deaths plotted against fitted deaths with trendline on top.

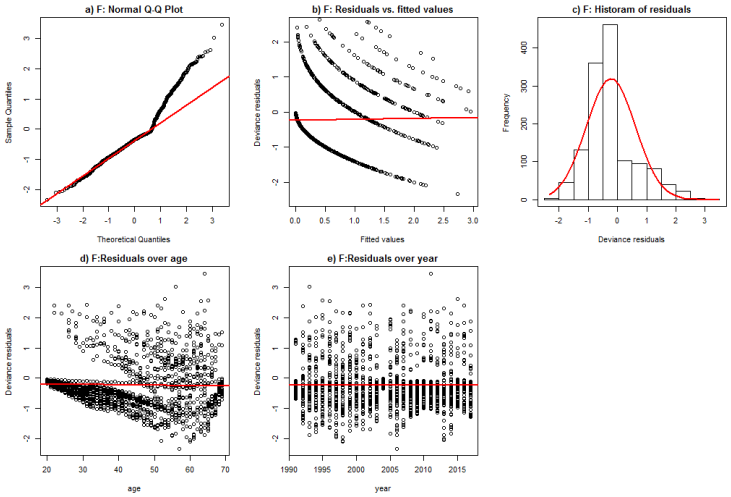


Figure B.3: Residual diagnostics for GAM model f.ay.ten.mod, model with smoothed interaction terms and smooted main effects splitted. a) QQ-plot of residuals, b) Fitted values vs. residuals shown as points plotted with trendline, c) Histogram of residuals with a normal line plotted on top, d) Residuals over age with red trendline, e) Residuals over year with red trendline

Models with smoothed interaction terms can also bw fitted through a single tensor product smooth, using `te()` in `R`. We did this in Section 6.5 and called the resulting models m.ten.mod for males and f.ten.mod for females, their residuals diagnostic plots are given in Figure B.4 and Figure B.5 respectively.

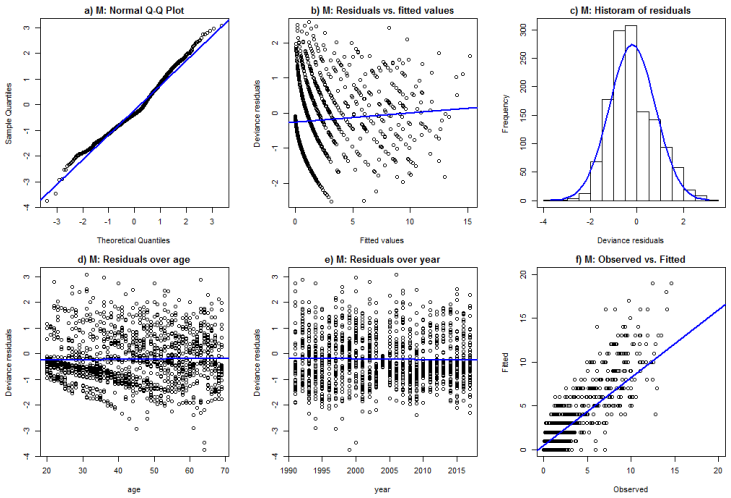


Figure B.4: Residual diagnostics and observed vs. fitted values of GAM model m.ten.mod, model with a single tensor product smooth. a) QQ-plot of residuals, b) Fitted values vs. residuals shown as points plotted with trendline, c) Histogram of residuals with a normal line plotted on top, d) Residuals over age with trendline, e) Residuals over year with red trendline, f) Observed deaths plotted against fitted deaths with trendline on top.

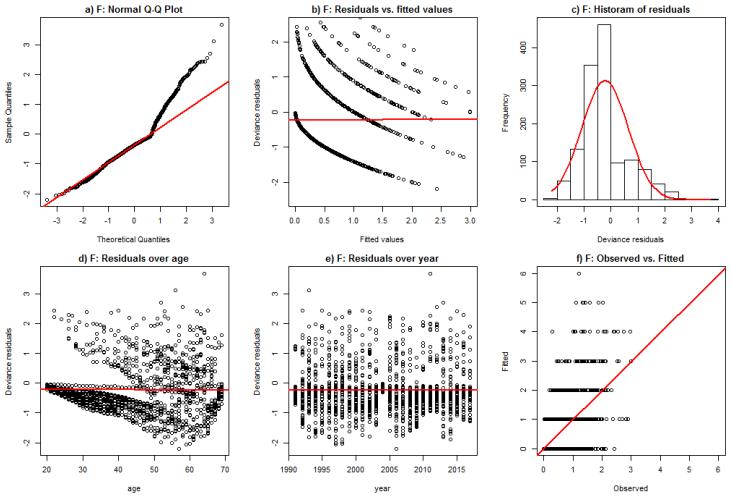


Figure B.5: Residual diagnostics and observed vs. fitted values of GAM model f.ten.mod, model with a single tensor product smooth. a) QQ-plot of residuals, b) Fitted values vs. residuals shown as points plotted with trendline, c) Histogram of residuals with a normal line plotted on top, d) Residuals over age with trendline, e) Residuals over year with red trendline, f) Observed deaths plotted against fitted deaths with trendline on top.

B.4 Tables of missing exposures in Section 6.6

In Section 6.6 we group data according to single years, single ages and the three biggest NACE-sections (C, G and K). This should leave us with two data sets with 4050 observations, one for males and one for females. We do however miss exposures for some of the variable combinations. We end up with a male data set of 4017 observations and a female data set of 3995 observations. Missing exposures for males are given in Table B.5. Missing exposures for females are given in Table B.6.

Table B.5: Missing male exposures for data used in Section 6.6

NACE-section	Insurance year	Age	NACE-section	Insurance year	Age
C	1991	20	K	1991	23
C	2003	20	K	1991	24
C	2004	20	K	2006	24
C	2007	20	K	2007	24
K	1998	20	K	1991	25
K	2004	20	K	1998	25
K	2012	20	K	2005	25
K	2013	20	K	2006	25
K	1991	22	K	2008	25
K	1997	22	K	2010	25
K	2004	22	K	2012	25
K	2013	22	K	2013	25
K	2004	23	G	2003	69
K	2005	23	G	2004	69
K	2004	27	G	2005	69
K	1991	61	G	2006	69
K	1991	22			

Table B.6: Missing female exposures for data used in Section 6.6

NACE-section	Insurance year	Age	NACE-section	Insurance year	Age
C	2004	21	K	2010	21
C	2004	25	K	1991	22
C	2004	28	K	1997	22
C	2004	63	K	1999	22
C	2004	65	K	2000	22
C	2004	66	K	2005	22
C	2004	67	K	2006	22
C	2004	68	K	2009	22
C	1995	69	K	2010	22
C	2003	69	K	2012	22
C	2004	69	K	2013	22
K	2004	20	G	2000	68
K	2005	20	G	2004	68
K	2006	20	G	2005	68
K	2007	20	G	2006	68
K	2008	20	G	1991	20
K	2013	20	G	1992	20
K	2004	21	G	1997	20
K	2013	21	G	1998	20
K	2008	22	G	1999	20
K	2005	23	G	2000	20
K	2004	25	G	2001	20
K	1991	20	G	2003	20
K	1991	21	G	2004	20
K	1996	21	G	2005	20
K	1998	21	G	2006	20
K	2005	21	G	2007	20
K	2006	21			

B.5 Residual diagnostic plots of three of the models fitted in Section 6.6

In Section 6.6 we fit multiple GAMs with smoothed age, smoothed year and categorical NACE-section as main effects. For male model gam.M3.nace.m, we also include a smoothed interaction effect between age and year, and a smoothed interaction effect between year and NACE-section K. Male model gam.M4.nace.m, has the same effects included in the model as gam.M3.nace.m, but in addition have a smoothed effect between age and NACE-section K. Choosing a male model in Section 6.6 we end up with gam.M3.nace.m and gam.M4.nace.m as top two preferred models, and these are the models we have to choose between in the end. Residual diagnostic plots of model gam.M3.nace.m are given in Figure B.6. Residual diagnostic plots of model gam.M4.nace.m are given in Figure B.7.

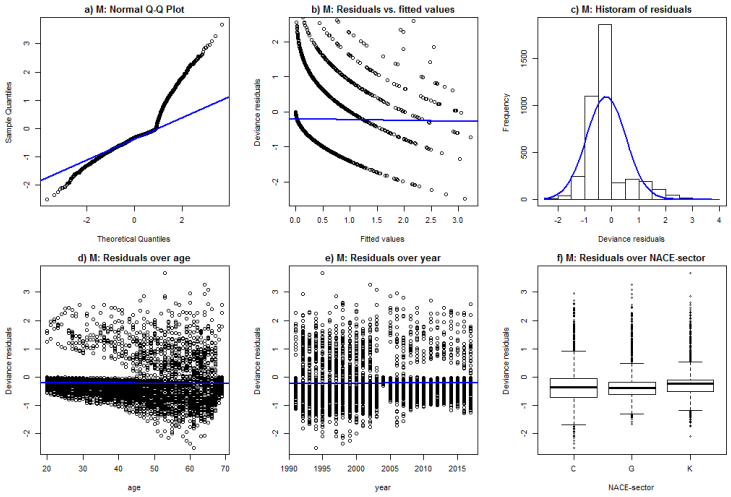


Figure B.6: Deviance residual diagnostics of GAM model gam.M3.nace.m, fitted on a male data set of 4017 observations. a) QQ-plot of residuals, b) Fitted values vs. residuals with blue trend line, c) Histogram of residuals with a normal line plotted on top, d) Residuals over age with blue trend line, e) Residuals over year with blue trendline, f) Residuals over NACE-sections.

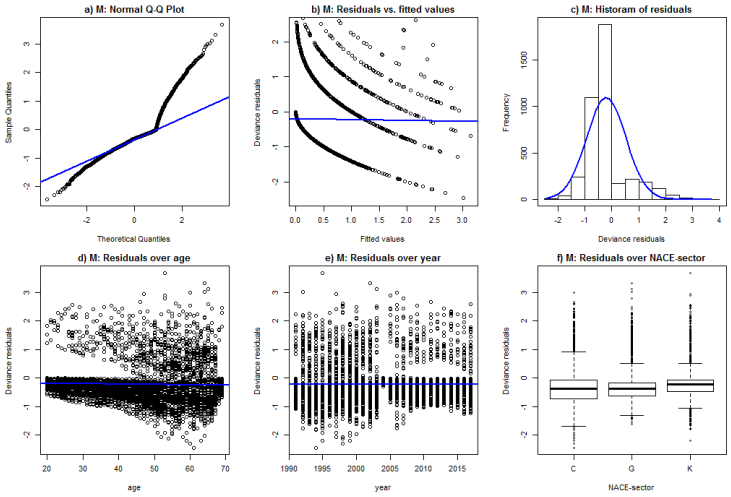


Figure B.7: Deviance residual diagnostics of GAM model gam.M4.nace.m, fitted on a male data set of 4017 observations. a) QQ-plot of residuals, b) Fitted values vs. residuals with blue trend line, c) Histogram of residuals with a normal line plotted on top, d) Residuals over age with blue trend line, e) Residuals over year with blue trendline, f) Residuals over NACE-sections.

When we fit GAMs for the female population in Section 6.6, none of the interaction terms improve the fit of the model with main effects only, model gam.M1.nace.f. This model is therefore chosen for the female population. The residual dagnostic plots of model gam.M1.nace.f are given in Figure B.8.

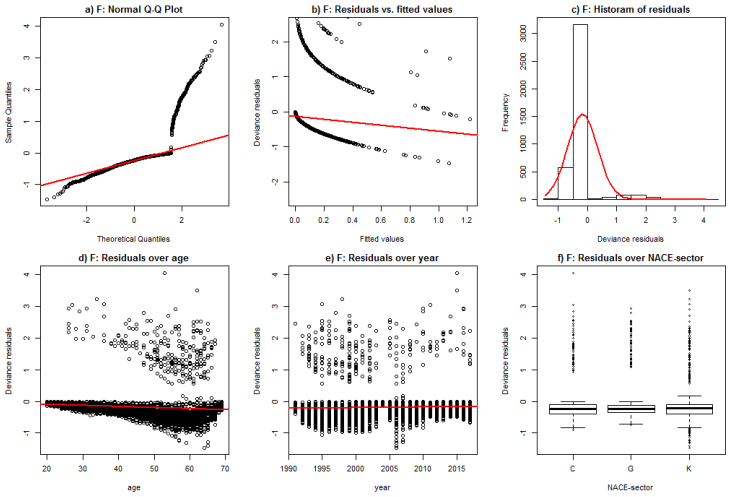


Figure B.8: Deviance residual diagnostics of GAM model gam.M1.nace.f, fitted on a female data set of 3995 observations. a) QQ-plot of residuals, b) Fitted values vs. residuals with red trend line, c) Histogram of residuals with a normal line plotted on top, d) Residuals over age with red trend line, e) Residuals over year with red trendline, f) Residuals over NACE-sections.

B.6 Comparison of death rate predictions of top two male models in Section 6.6

The death rate prediction plots of models gam.M3.nace.m and gam.M4.nace.m for NACE-section K was given in 6.6. We did however not show the death rate prediction plots for NACE-sections C and G, they were however mentioned, and are therefore given in this section. The death rate prediction plots for NACE-section C are given in Figure B.9. The death rate prediction plots for NACE-section G are given in Figure B.10.

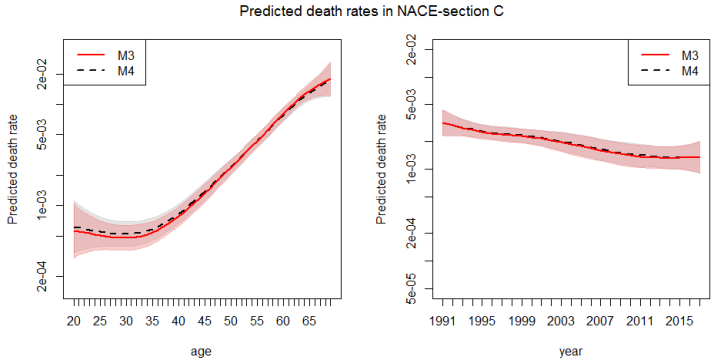


Figure B.9: Death rate predictions with confidence bands of default smoothed models, gam.M3.nace.m (M3: Red line and confidence band) and gam.M4.nace.m (M4: Gray stippled line and confidence band), in NACE-section C - manufacturing. Predicitons over age (left panel) are made by fixing year at 1997. Predictions over year are made by fixing age at 50. The models are fitted on a male data set with 4017 observations and the predictions are plotted on a log-scale.

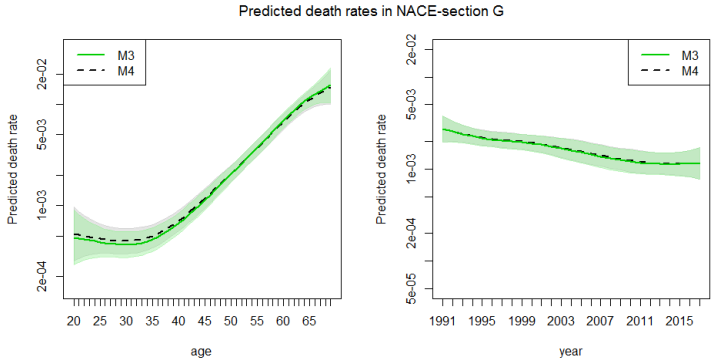


Figure B.10: Death rate predictions with confidence bands of default smoothed models, gam.M3.nace.m (M3: Green line and confidence band) and gam.M4.nace.m (M4: Gray stippled line and confidence band), in NACE-section G - Wholesale and retail trade; repair of motor vehicles and motorcycles. Predicitons over age (left panel) are made by fixing year at 1997. Predictions over year are made by fixing age at 50. The models are fitted on a male data set with 4017 observations and the predictions are plotted on a log-scale.

B.7 Comparison of death rate predictions of GLM and GAM in Section 7.2

In Section 7.2 we compare the death rate predictions made by GLM and GAM within different NACE-sections. In the section we did however not show the prediction plots for females in NACE-section G. We did this as the prediction plots look a lot like the prediction plots for NACE-section C. The prediciton plots are therefore given in this section, Figure B.11.

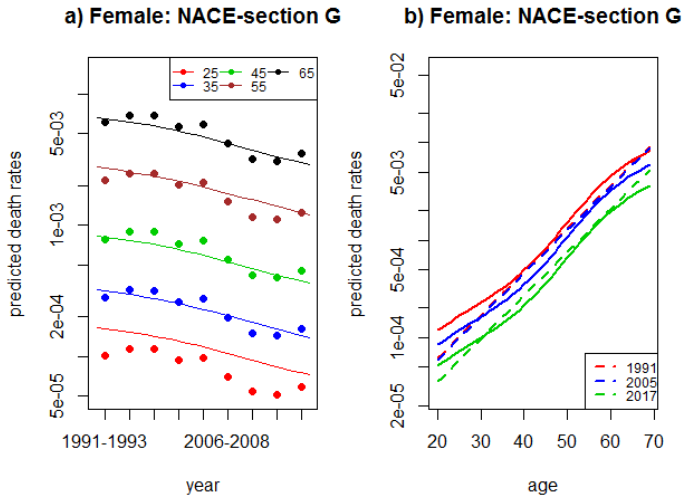


Figure B.11: Predicted death rates in NACE-section G (G - Wholesale and retail trade; repair of motor vehicles and motorcycles) for females of given age, panel a), and given years, panel b). Predictions are plotted on a log-scale by GLM (stippled lines) and GAM (solid lines). Each age in panel a) is represented by a color; red: 25, blue: 35, green: 45, brown: 55 and black: 65. Each year in panel b) is given colors; red: 1991, blue: 2005, green: 2017. GLM predictions are made by model M1.nace.f and GAM predictions are made by model gam.M1.nace.f.

Bibliography

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. 1st. John Wiley and Sons, Inc.

Borgan, J.-K. (2009). "Vedvarende ulikhet i dodelighet etter yrke". In: *Okonomiske analyser* 3, pp. 43–47.

Braut, G. S. (2018). "healthy worker effect". In: *Store medisinske leksikon*. Hentet 14. juni 2019.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. 2nd. Duxbury.

Eurostat (2008). *NACE Rev. 2. Statistical classification of economic activities in the European Community*. Methodologies and Working papers. European Union. Luxembourg.

Hastie, T, Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. 2nd. Springer Science and Buisiness Media.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. 1st. Chapman and Hall / CRC.

Jong, P. D. and Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. 1st. Cambringe University Press.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd. Chapman and Hall.

R Core Team (2017). *R: Predict Method for GLM Fits*. stats version: 3.3.3. R Foundation for Statistical Computing. Vienna, Austria.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*. 3rd. Duxbury.

Shah, D (2009). "Healthy worker effect phenomenon". In: *Indian journal of occupational and environmental medicine* 13(2). doi:10.4103/0019-5278.55123, pp. 77–79.

Wood, S. N. (2006). *R: getting the smoother matrix from smooth.spline*.

— (2017). *Generalized Additive Models, An Introduction with R*. 2nd. CRC Press, Taylor and Francis Group.

— (2018a). *R: Default GAM plotting*. version: 1.8-26.

— (2018b). *R: Define tensor product smooths or tensor product interactions in GAM formulae*. version: 1.8-26.

— (2018c). *R: Defining Smooths In GAM Formulae*. version: 1.8-26.

— (2018d). *R: Generalized additive models with integrated smoothness estimation*. version: 1.8-26.

— (2018e). *R: Prediction From Fitted GAM Model*. version: 1.8-26.

Zuur, A. et al. (2009). *Mixed effects models and extensions in ecology with R. Statistics for Biology and Health*. New York: Springer.