Modeling risk factors of mortality in Germany in the context of survival analysis using the Socio-economic Panel data

submitted by

Sarah Bekele

in partial fulfillment of the requirements
to obtain the degree
Master of Science in Volkswirtschaftslehre

at the School of Business and Economics
Humboldt University of Berlin

Supervisor
Dr. Sigbert Klinke
(Ladislaus von Bortkiewicz Chair of Statistics)

Examiner
Prof. Dr. Wolfgang Karl Härdle
(Ladislaus von Bortkiewicz Chair of Statistics)

Berlin, 4^{th} of May 2015

Contents

List of Figures

List of Tables

List of Abbreviations

CHF Cumulative Hazard Function

CPHM Cox Proportional Hazards Model

d.o.f. degrees of freedom

DIW German Institute for Economic Research

GOF Goodness of fit

HF Hazard Function

HR Hazard Ratio

KME Kaplan-Meier estimator

LR Likelihood ratio test

LT Life Table

OLS Ordinary Least Squares

SCM Stratified Cox model

SF Survivor Function

SOEP Socio-economic Panel

1 Introduction

Since the 17^{th} century researchers have been concerned with factors that influence mortality.[1] Mortality itself is defined as *"the state or condition of being subject to death"*.[2] Thus, mortality summarizes an individual's life time. Therefore, the concept of mortality is closely linked to the notion of life expectancy.

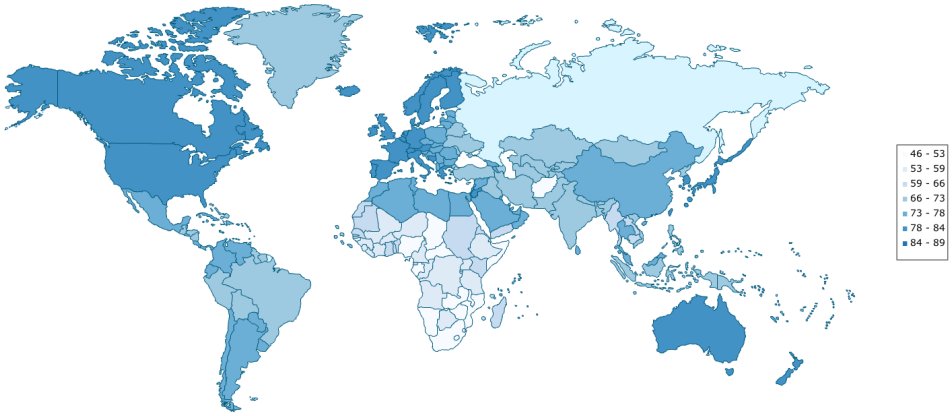

Figure 1: Average life expectancy in years at birth in 2012

Figure 1 summarizes the average life expectancy in years at birth in 2012 for each country in the world. The darker the blue on the map, the higher is the average life expectancy in the country and vice versa. Germany is ranked on the list as the 28^{th} country out of the 220 countries totally included in the ranking. With an estimated average life expectancy of 80 years in 2012, Germany's estimated life expectancy was approximately nine years lower than the estimated life expectancy of the leading country in the ranking - Monaco. As shown in figure 1, life expectancy varies greatly among the countries in the world. The continents North America, Europe, and Australia demonstrate the highest density of high estimated average life epxectancies.[3]

Research has shown that the longevity of an individual's lifetime depends on various factors, such as lifestyle[4], education, income, health care, genetics, and demographic factors.[5] Therefore, the purpose of the analysis underlying this thesis is to identify key factors that influence mortality in Germany. The emphasize here lies on analyzing

[1] Graunt (1662).
[2] Miller-Keane Encyclopedia definition: Miller-Keane (2003).
[3] Mundi (2015).
[4] This includes for example diet-related choices, smoking, frequency of exercise.
[5] Scott (2013).

general differences among certain groups. In addition, it will be discussed whether the data at hand constitutes a meaningful and useful source of information to answer the research question mentioned above. Furthermore, it will be verified whether the modeling of risk factors of mortality fits into the context of survival analysis and thus whether the Cox model is a suitable for the analysis.

Firstly, chapter 2 introduces the data set that is used in order to carry out the analysis. Furthermore, chapter 2 explains how the data set has been modified to meet the needs of the analysis and how it has been set up. Afterwards, the variables included in the analysis are introduced to the reader. Since the modeling of mortality fits into the context of the usage of survival analysis methods, the last section in chapter 2 discusses specific issues that arise from the data when modeling time-to-event linkages.

Then chapter 3 gives an overview of descriptive statistics that need to be applied when working with survival data. In particular, the life table will be introduced in detail. On the other hand, the Kaplan-Meier estimator will only be mentioned briefly. More importantly, chapter 3 then introduces in the context of the life table the concepts of the Survivor Function, the Hazard Function and the Cumulative Hazard Function. Furthermore, the log-rank test for the comparison of Survivor Functions will be explained in order to select meaningful covariates for the development of a model in chapter 4.

Chapter 4 continues then with a brief illustration on the reasoning for the selection of the Cox model to conduct the analysis. Afterwards, the basic Cox model will be introduced. Furthermore, the Cox model needs to be modified in order to allow for tied survival times. Therefore, the Breslow and the Efron approximations are described. Chapter 4 is then completed by fitting the Cox model to the data and discussing the estimation results.

Afterwards, chapter 5 describes the goodness of fit test that needs to be applied to the estimation results in order to test if the proportionality assumption underlying the Cox model is met by the data. This allows for a validation of the estimation results. If the proportionality assumption is not fulfilled, the stratified Cox model can be applied to the data to account for this issue. Therefore, the stratified Cox model is introduced.

Lastly, chapter 6 gives a brief summary of the results established along the way in the thesis and recaps possible shortfalls in the method that has been used and also in the data. In addition, chapter 6 encapsulates a short outlook into possible, further research.

2 Data Preparation

2.1 Data set

The analysis carried out in this paper is primarily using the Socio-economic Panel (SOEP) data provided annually since 1984 by the German Institute for Economic Research (DIW), this institute is based in Berlin and was founded in 1925. The SOEP holds data collected through a voluntary, representative and annually repeated household survey with the aim to portray the socio-economic status quo in Germany. This research paper uses the v29, which means that the data set includes observations from 1984 to 2012 and is equivalent to a total number of 29 waves.[6]

Data provided by the DIW to researchers can be obtained either in a long format, which holds 16 individual data files, or in a format that holds a data file for each wave and each part of the survey, which amounts to a total number of 375 individual data files.[7] For this analysis the long format has been chosen as a starting point, since this paper focuses not only on one particular wave, but on all 29 waves that have been completed so far. Within this long format data, the data sets can be divided into two major groups: the household and the personal, also called individual-specific, data sets. Given that the main aim of this analysis is to model mortality and its driving forces, the research is based on the data provided on an individual-specific level and hence the household level data, besides for the income variable as one of the covariates, will be excluded when the master data set is created. Thus the researcher faces seven data sets[8] with a total of more than 3.700 variables that contain individual level data in long format. This means each individual has multiple records in these data sets, one record for each year the individual participated in the survey.[9]

[6] Deutsches Institut für Wirtschaftsforschung (2015).

[7] In the SOEP data personal level datasets are indicated by starting with the letter p and household level datasets start with an h. The waves are marked with the letters in the alphabet that correspond to the number of the wave.

[8] This includes the following datasets: ppfadl, pbrutto, pl, pgen, pequiv, pbr_exit, and bio.

[9] Singer and Willett (2003), p. 22.

2.2 Merging and modification of the data set

Prior to the merging of the data sets the number of variables has been reduced to a minimum of possibly relevant covariates.[10] All seven separate data sets have been added to the data set ppfadl to build the master data set. In order to match the records correctly, the identification numbers of the households, the individuals and the survey year have been used as the matching variables.[11]

Table 1 on page 5 gives an overview of all variables that have been used in the analysis. Column one summarizes the name of the variable according to the coding in the software that was used to carry out the analysis. If the variable name starts with a *d* this indicates that the variable is coded as a dummy variable. Therefore it is zero if the attribute does not apply to the individual and one if it does apply. For instance, the variable dmarried is coded one if the individual has ever been married and zero otherwise. Column two indicates the coding of the variable and thus summarizes all possible values the variable can take on. In addition, column three provides a short description in verbal terms of what exactly the variable measures.

2.2.1 Missing values and inconclusive entries

The SOEP data contains large amounts of missing data because it is a study that relies on voluntary answers from the survey participants. Negative values, such as minus eight[12], minus three[13], minus two[14], and minus one[15], have been set to missing values instead due to the fact that they do not hold any useful information. There is one exception of a group of variables where this change has not been implemented. All disease variables are part of this group and their minus two values have not been recoded to missing because it was clear from investigating other variables that contained the same information that these minus two values represented a negative answer to the disease questions. Therefore, these values have been recoded as zeros instead. The category of diseases according to the Federal Bureau of Statistics contains two of the leading causes of death in Germany in 2013: cardiovascular diseases and

[10] All variables related to for instance education, income, family status, frequency of exercise, smoking habits, disease history, family background have been maintained in the data set.

[11] In the SOEP this corresponds to the variables cid and hid for the household identification number, pid as the individual identification number and syear as the survey year.

[12] Minus eight codes missing values for survey years when the variable at hand has not been asked.

[13] Values equal to minus three represent non-valid answers.

[14] Minus two values represent "Does not apply" answers.

[15] Minus one represents values when individuals chose not to answer the question.

Table 1: Overview of variable coding and description

Variable name	Coding/Range	Description
cid		Household ID
pid		Personal ID
time	[1, 109]	Survival time
censorconst	1=died, 0=censored	Censoring indicator variable
gender	1=female, 0=male	Gender
birthyear	[1882, 2012]	Birth year
educmax	1=no practical training 2=practical training 3=university	Highest education level
dmarried	1=married, 0=not married	Marriage status
ddivorce	1=divorced, 0=not divorced	Divorce status
dchildren	1=has children, 0=does not have children	Does individual have children
numchildren	[0, 12]	Number of children
dborngerm	1=yes, 0=no	Was the individual born in Germany
dgermnat	1=yes, 0=no	German citizen
immiyear	[1949, 2012]	Year of immigration to Germany
dasthma	1=yes, 0=no	Diagnosed with asthma
ddiabetes	1=yes, 0=no	Diagnosed with diabetes
dcancer	1=yes, 0=no	Diagnosed with cancer
ddepress	1=yes, 0=no	Diagnosed with depression
ddisabled	1=yes, 0=no	Ever been disabled
disabledcat	2=both 1=disabled 0=not disabled	Has individual been disabled
agestsmo	[10, 85]	Age when started to smoke
yearssmoking	[0, 66]	Number of years of smoking
smokingperday	[0, 99]	Cigars, cigarettes, and pipes smoked per day on average
dsmoking	1=yes, 0=no	Ever been a smoker
incquart	1=income$\leq$25% 2=25% <income$\leq$ 50% 3=50% <income$\leq$ 75% 4=income$\geq$ 75%	Income quartile
agefatherdeath	[17, 110]	Age of father when he passed away
agemotherdeath	[18, 110]	Age of mother when she passed away
fatheduc	2=university 1=practical training 0=no practical training	Father's highest level of education
dsiblings	1=yes, 0=no	Siblings
dmigback	1=yes, 0=no	Migration background
region	3=both 2=West 1=East	Region
dmigraine	1=yes, 0=no	Diagnosed with migraine
ddementia	1=yes, 0=no	Diagnosed with dementia
dotherdis	1=yes, 0=no	Diagnosed with unspecified disease
dnodis	1=yes, 0=no	Never diagnosed with a disease
dstroke	1=yes, 0=no	Diagnosed with stroke
dpressure	1=yes, 0=no	Diagnosed with high blood pressure
dheart	1=yes, 0=no	Diagnosed with a heart condition
quitsmoking	[11, 84]	Age when quitting smoking
agework	[5, 84]	Age when started to work
dysports	1=yes, 0=no	Did sports when young
motheduc	2=university 1=practical training 0=no practical training	Mother's highest level of education

cancer. Therefore, it is important to include these variables in the analysis as possible covariates.[16]

Generally speaking, most variables that are contained in the SOEP data do not only appear in form of one variable. For instance, the variable gender can be found in each one of the different subsets of the data. This is for example caused by the fact that for one survey year gender was coded as one variable and in the next survey year the variable gender obtained a new name. When the data sets have been merged into the master set, the values of these variables have been combined into one variable to minimize the occurence of missing values. This method has been applied to most of the variables mentioned in table 1.

The number of missing variables in the variable birth year has been minimized using a reconstruction method out of the survey year and the age variable contained in the SOEP data. This was only possible for those records that did provide an age. The survey year itself did not cause any limitations to the procedure because it does not contain any missing values.

The parents' birth and death year had to be reconstructed due to the fact that they were not all given in a four-digit format. Therefore, all years provided either in a two-digit or a three-digit format had to be recoded into the four-digit format that has been applied for the majority of observations. This mainly included a shift by 1900 years for the two-digit years and a 1000 year shift for the three-digit years. After these changes have been implemented, all unreasonable entries were excluded from the sample. This includes for example individuals who claim that their parents passed away before the individual was born. Furthermore, all values of the father's death year equal to the birth year of the individual plus a one year grace period have been kept due to the duration of a pregnancy. In addition, it has been assumed that the individual's parents were only capable of becoming parents between the age of 14 and 80. Therefore, all individuals who claim that their parents were younger than 14 or older than 80 years have also been excluded from the sample. For four individuals the provided data shows that their fathers were more than 120 years old at the time of their birth. After investigating these observations, it is clear that they have been falsely coded because when 100 years were added for three of the four observations, the father would have been in his twenties, when the individual was born. This seems to be very likely to hold true.

[16]Statistisches Bundesamt (2013).

2.2.2 Outliers

Only metric variables contained in table 1 have been investigated in regards to outliers. For all categorical variables, the analysis of outliers did not have to be performed due to their scaling level. Only a few of these variables did indeed show the existence of outliers and those who seemed unreasonable have been excluded from the sample. For example, for the variable children, one woman from Turkey claims to have 94 children which seems unlikely to be a valid answer. Therefore, this woman has been excluded from the sample. In addition, everyone who claimed in the dummy variable for children that they do not have any children has been set to zero in the variable that contains how many children an individual has.

2.3 Special problems arising when working with Survival Data

2.3.1 Kind of Events

In the context of survival analysis it is important to be aware of what kind of event one is working with in the analysis, as this will determine if further adjustments to the model have to be made in order to incorporate the specific characteristics of the event into the model. Firstly, events can be distinguished by the fact of whether they can occur repeatedly or non-repeatedly. The amount of job changes, births of a child, arrests or marriages all constitute possibly repeatable events as they can happen more than once over the course of a lifetime of an individual. Whereas on the other hand in the analysis of this thesis the main focus lies on a non-repeatable event, which is a person's death.[17]

Secondly, events can be divided into single and multiple kinds of events. An example for an event that belongs to the multiple kinds of events group is if a study examines the effectiveness of a treatment for a certain kind of disease. It is crucial for the analysis to distinguish for each individual whether the cause of death stems from the disease that the treatment was targeting or if the individual died from some other cause. When studying life expectancy, this analysis carried out in this thesis will not consider the actual causes of death and thus death will be regarded as a single and non-repeatable event.[18]

[17] Allison (1984), p. 9, 13.
[18] Allison (1984), p. 14, 42 ff.

2.3.2 Censoring

When working on longitudinal time-to-event data it most likely will contain incomplete observations. If these incomplete observations are not caused by the study design and are individual specific, this is considered censoring. Due to the fact that the most up to date SOEP data currently only contains observations from 1984 to 2012, all individuals who did not pass away during that time period are called right-censored because this censoring occurs in the right tail of the time axis. For censoring it is not of importance why an individual did not pass away yet as this can have various reasons, such as being lost for follow-up and dropping out of the study.[19]

Left censoring on the other hand causes incomplete observations, if the event of interest has already occurred previously to entering the study. When modeling life expectancy using the SOEP data, this is impossible to occur due to the design of the study and the fact that individuals cannot participate in the survey if they have already passed away.[20]

The third kind of censoring that can be observed in data is the so-called interval censoring. This type of censoring is based on the specifics in measuring the time component in the data. For instance, if subjects, who are alcohol or drug abusers, are interviewed every half a year in order to analyze the time until they relapse, the researcher will only gain information on in which half year interval the relapse occured, but not specifically on which day. Therefore, only the interval is known to the analyst and creates a discrete nature of the dependent variable.[21] Due to the fact that it remains debatable if age is considered to be continunously or discretely measured, it will be assumed for the underlying analysis that time in the SOEP data is measured in continuous units even though only the year a person passed away is reported in the data. Thus the data contains year-based reports on the individuals that died in any given period of time.[22]

[19] Hosmer and Lemeshow (1999), p. 17 ff.
[20] Hosmer and Lemeshow (1999), p. 20.
[21] Hosmer and Lemeshow (1999), p. 21.
[22] Cf. Cleves, Gould, and Gutierrez (2002), p. 140.

Generally speaking, most longitudinal data contains right-censored observations and only some of them comprise left-censored observations. It is much easier to handle left-censored observations because often it is possible to gather the needed information from past records from a retrospective point of view. Whereas when one faces right-censored observations, it is not possible to obtain information from the future and thus one cannot fill in those gaps. From a statistical point of view, however, the problems arising when facing right-censored observations are much easier to solve than the ones occurring with left-censored observations.[23]

It is of utter importance to notice, that if data includes incomplete observations caused by censoring the standard descriptive tools, such as calculating means and the standard deviation, might not be suitable to be applied in this setting. This is caused by the effect that a censored event time only suggests that the event has not occurred yet during the observation period. Therefore it provides only information about the nonoccurrence. Traditional descriptive statistical procedures cannot directly analyze censored and non-censored observations simultaneously. Thus it is necessary to use survival methods instead because these can simultaneously analyze these two types of observations.[24]

2.3.3 Truncation

The second reason to obtain incomplete information when studying time to event data is caused by truncation, which entirely arises from the study design itself. There exist two kinds of truncation, left truncation and right truncation.

Left truncation is often also referred to as delayed entry. This is the case for an individual for example in the context of life expectancy for almost all individuals because when an individual is born it is already part of the risk group of being able to experience death from birth on. In the SOEP data, however, most individuals did not participate in the survey from the point of birth onwards and entered the study during their adulthood. Thus this constitutes a delayed entry into the study. Nevertheless, due to the modification of the survey records, this issue of left truncation in the analysis has been eliminated prior to the estimation. These modifications will be discussed in chapter 3.1.[25]

[23] Andreß (1992), p. 93.
[24] Singer and Willett (2003), p. 325.
[25] Hosmer and Lemeshow (1999), p. 20.

Right truncation on the other hand occurs as a selection process when all individuals in the study must have experienced a certain event, such as a diagnosis with a specific disease, in order to participate in the study. For example if one were to analyze the risk factors of Mantel cell lymphoma, the time until the event takes place is known. Because individuals in the SOEP data are not chosen to participate in the study based on a common event that has happened to all of them and is important to the study of life expectancy, this data does not contain any right-truncated observations. Thus the analysis does not need to take this form of incomplete observations into account.[26]

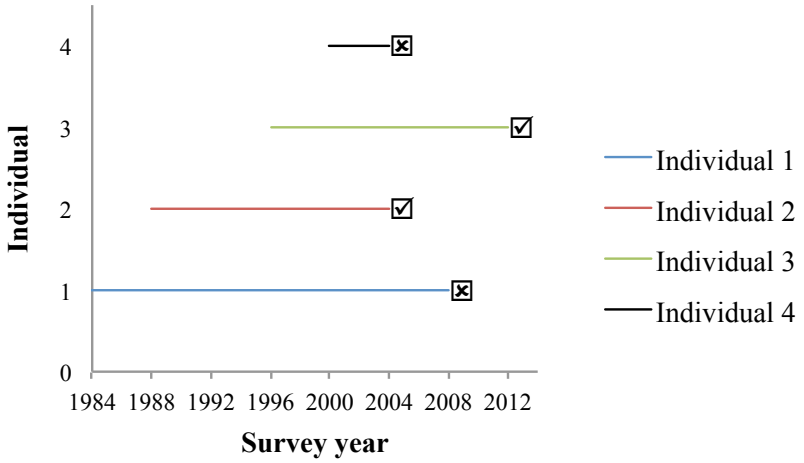

Figure 2: Censoring

Figure 2 graphically gives an overview of the censoring problematic. The x-axis contains the value of the survey years according to the SOEP and holds values from 1984 to 2012. The y-axis consists of four hypothetical identification numbers of four separate individuals. Individuals one and four are marked with an X because they both passed away during their survey participation. On the other hand, individual two and three are both right-censored. Individual two chose not to continue his or her participation in the survey and did not experience the event of interest. Moreover, individual three does not only experience right-censoring, but also right-truncation. This is caused by the fact that individual three did not die yet as of 2012, nevertheless the SOEP data artificially only included observations including 2012. Individuals two, three, and four also experience left-truncation as a result of their delayed entry into the study.[27]

[26] Hosmer and Lemeshow (1999), p. 21.
[27] Cf. Hosmer and Lemeshow (1999), p. 19.

3 Descriptive Analysis

3.1 Setup of the data

In addition to the modifications carried out and described in chapter 2, the data set has been reshaped from a long format with a total number of 697,111 observations into a wide format including 75,574 observations. The long format contains multiple observations per individual; in particular it holds one observation for each survey year as long as the individual participates in the survey. On the other hand, the wide format, also known as single episode data, excludes the survey year component and thus for each individual there is only one observation kept. Even though the SOEP was originally given in a long format, it is possible to reshape the final data set because the covariates that will be discussed in this analysis have been created in such a fashion that they are not varying with time, but are time-constant, which means that they hold the same value for each survey year and thus the survey year component is not needed for the calculations. It also implicitly assumes that all observations are independent of each other.[28] Generally speaking, survival analysis procedures are capable of working with time-varying covariates and therefore it does not belong to the group of possible drawbacks of the method. Nevertheless, adjustments have to be made to the basic approach.[29]

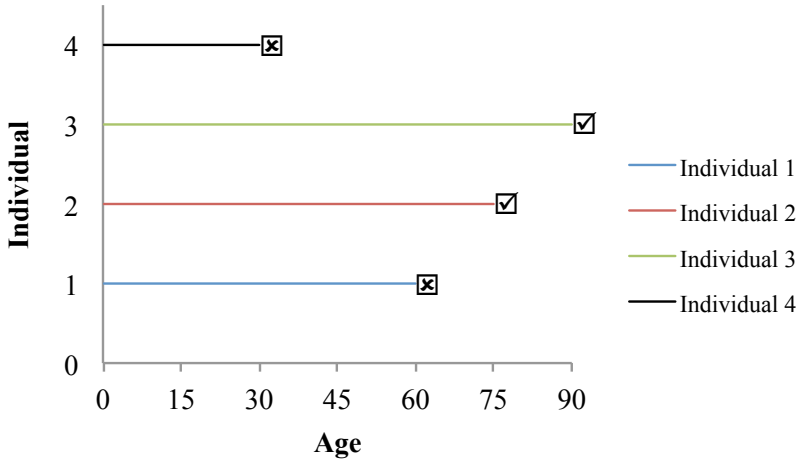

Figure 3: Desired setup of the data

[28] Blossfeld, Golsch, and Rohwer (2007), p. 49.
[29] Schneider (1991), p. 163-173.

Figure 3 describes graphically the desired setup of the data that has been achieved through the modifications to the data that have been explained above. In contrast to figure 2, figure 3 contains the age of the individual instead of the survey year on the x-axis. This allows for each individual that the origin state is defined as the birth year of the individual. Therefore, the analysis does not apply a direct cohort-based approach.[30] The y-axis remains unchanged. This setup is desirable because it removes the survey year component from the data, which is not needed to carry out the analysis.[31]

A time variable describing the survival time per individual has been created. This variable stems either from a combination of the year the individual was born and the year the individual passed away. Alternatively for everyone who is right-censored the time variable has been calculated as the difference between the year the person was born and the last survey year the individual participated in the study. For 77 individuals, stemming from the survey years 2010, 2011 and 2012 only, the survival time was equal to zero. This means that they have only been observed once and this was in the year they were born. Therefore, it is likely but not certain that these individuals survived until they completed their first year, but because these individuals are right-censored and there exists no certainty that they did survive until they reached the second year of participation, they have been excluded from the sample. This is also a wise decision to make in regards to when the data is declared to be survival data in STATA[32], which is necessary in order to use the software's tools specifically provided for this type of data. When one carries out this step, STATA automatically excludes these entries because the origin and destination state are identical. The origin state can be defined as the starting point. For this analysis, the origin state is identical for all individuals in the sample, equal to zero and purely based on the year the individual was born. The destination state in this example is given as the event of interest, which means that an individual passed away. This can be defined by using the fact that an individual provided the year in which he or she died. Thus the failure variable, which is equivalent to the censoring variable, determines the destination state.[33]

[30] A cohort-based approach means that individuals are grouped into cohorts based on their birth year. Nevertheless, in order to carry out the log-rank test for the covariate birthyear, the variable had to be grouped artificially because the log-rank test can only be applied to categorical and indicator variables.
[31] Cf. Hosmer and Lemeshow (1999), p. 20.
[32] For the analysis, the software STATA has been used in the version 12.0.
[33] Blossfeld, Golsch, and Rohwer (2007), p. 49-53.

Furthermore, a censoring dummy variable also needed to be generated to distinguish the individuals according to whether they experienced the event of interest during the observation period. Everyone who belongs to the individuals that are right-censored have been coded with a zero value for the censoring variable, and all individuals that did pass away obtained a one value. A total amount of 5,295 individuals died during the participation in the survey, and thus 70,279 people were right-censored and did not experience the event of interest during the observation period. This means that seven percent of the study participants passed away during the observation period.

3.2 Life Table

The first record of a Life Table (LT) goes back to John Graunt who published the first LT in 1662.[34] Thus this is a descriptive statistical tool that has already been in practice for over three hundred years and belongs to the class of non-parametric descriptive methods. The method is considered to be non-parametric due to the fact that it does not make any assumptions regarding the underlying distribution of the process. Therefore, the LT is an excellent tool to get a first impression of the data at hand. With the current standard of technology and increased computational power of computers, the LT cannot be considered preferable to the Kaplan-Meier estimator (KME)[35] anymore. In fact, the LT method holds a disadvantage because the researcher has to define discrete time intervals prior to performing the analysis. Therefore, the method forces the researcher to group the time until the event or alternatively censoring occurs into fixed intervals. Consequently, the results depend on this arbitrary choice of intervals. It is not necessary for the intervals to be equidistant.[36] In addition to this limitation, it is also important to notice that reliable approximations can only be achieved, when a relatively large number of observations exists, due to the fact that a conditional calculation concept is applied.[37]

[34] Graunt (1662), p. 206 ff.

[35] The KME, also known as the Product-Limit estimator, will not be discussed in this paper, because the results of the life table method and the KME are very similar. They only differ by the fact that the KME does not require any arbitrary grouping of the data prior to carrying out the analysis and is thus used for continuous-time data. Cf. Luy (2002), p. 64 ff, Singer and Willett (2003), p. 483 ff.

[36] Blossfeld, Hamerle, and Mayer (1986), p. 43.

[37] Blossfeld, Golsch, and Rohwer (2007), p. 58.

The LT method allows for the calculation of non-parametric estimates of the Survivor Function (SF), the Hazard Function (HF) and the Cumulative Hazard Function (CHF). These three functions give the researcher a first impression of the distribution of risk that relates to the shape of the HF and also makes it possible to identify especially risky intervals.[38] For single transitions, which applies to the case of modeling mortality, the LT for each interval $j, j = 1, \ldots, p$ is purely relying on three main components:

$$\begin{aligned} R_j &= \textit{the number of individuals who enter the interval} \\ E_j &= \textit{the number of individuals who die in the interval} \\ C_j &= \textit{the amount of censored individuals at the end of the interval.} \end{aligned}$$

R_j often refers to the risk set in the literature. This term describes the number of people who enter a particular interval j, because they have neither previously been exposed to the occurrence of the event of interest nor have they been subject to censoring in any of the intervals preceding interval j. This means that everyone who did not die prior to reaching this particular interval j and has not been censored is considered to be part of the risk set. Therefore, this method constitutes a conditional concept.[39] The intervals are created in such a fashion that the lower bound is closed and the upper bound is open and therefore not included in the interval. The first risk set is equivalent to the initial total number of individuals participating in the study and thus it holds that:

$$n = R_1. \tag{1}$$

All following risk sets can be derived using the values of the risk set of the previous interval, the number of individuals who experienced the event of interest in the previous interval and the number of individuals who were exposed to censoring in the previous interval. They can be calculated using the following equation:[40]

$$R_j = R_{j-1} - E_{j-1} - C_{j-1}. \tag{2}$$

[38] Singer and Willett (2003), p. 332.
[39] Cf. Singer and Willett (2003), p. 326; Blossfeld, Golsch, and Rohwer (2007), p. 59.
[40] Blossfeld, Golsch, and Rohwer (2007), p. 63.

The LT also contains the survivor and the HF, which are both intertwined with each other. The hazard rate describes the risk that an individual will experience the event of interest in a given interval j. Thus the hazard $h(t_{ij})$ is defined as the conditional probability that an individual i will pass away in interval j, given that the person has not passed away in a previous period and is thus part of the risk set. Therefore, this discrete-time hazard summarizes whether and when an event occurs. The time variable in the data set describes a discrete random variable that can be denoted as T_i and summarizes the interval j in which individual i dies. The distribution of T_i can be easily illustrated using the concept of either the density or the cumulative density function. The first method defines the probability that individual i will die in time period j and can be written as:

$$Pr[T_i = t_j]. \tag{3}$$

On the other hand, the cumulative density function describes the probability that individual i will die prior to reaching interval j and can be denoted as:

$$Pr[T_i < t_j]. \tag{4}$$

In a general context, it is important to notice that each individual i has a unique discrete-time HF that describes the relationship between parameters and covariates. But for now the HF will only be used as a descriptive tool and thus will not be considered to be individual-specific throughout the analysis. Consequently, the index i will be dropped for now. The estimated discrete-time hazard for each interval j can be denoted as:

$$\hat{h}(t_j) = \frac{E_j}{R_j}. \tag{5}$$

The hazard rate is given as a probability bounded by zero and one. For the initial interval, the hazard rate $h(t_0)$ is not defined because the upper bound is not included in the first interval and it has been assumed that individuals are excluded from the analysis if their observation time is equivalent to the initial point in time and equal to zero. Therefore, no one can die during the first interval and also no one can be censored. [41]

[41] Singer and Willett (2003), p. 330 ff.

The SF on the other hand can be described as a cumulative tool that relates to the probability that an individual has not experienced the event of interest prior or during the interval j. Thus the probability of surviving at least until interval j has passed can be written as:

$$S(t_{ij}) = Pr[T_i > j]. \tag{6}$$

For simplicity, the index for the individual will be dropped again as it has been done previously with the HF because when looking at LTs, the overall trend and the rates for each interval are of importance for the researcher. Even though there exist multiple options in estimating the population SF, the most commonly used approach is capable of dealing with difficulties arising when censoring occurs in the data. It is calculated as the product of the estimated survivor probability of the preceding year and one minus the estimated hazard probability of the current interval and is given in the following form:

$$\hat{S}(t_j) = \hat{S}(t_{j-1})[1 - \hat{h}(t_j)]. \tag{7}$$

Thus the estimated SF gives a maximum likelihood estimate of the probability for each interval that an individual randomly selected from the sample will not pass away. Equation 7 thus depends on the estimated SF of the previous year, but it is also possible to express the estimated SF for time interval j solely based on the estimated hazard probability by using repeated substitution in the sense that:

$$\hat{S}(t_{j-1}) = \hat{S}(t_{j-2})[1 - \hat{h}(t_{j-1})]. \tag{8}$$

And then inserting this formula in equation 7 leads to:

$$\hat{S}(t_j) = \hat{S}(t_{j-2})[1 - \hat{h}(t_{j-1})][1 - \hat{h}(t_j)]. \tag{9}$$

This process is repeated until the estimated SF solely depends on the estimated hazard probabilities:

$$\hat{S}(t_j) = [1 - \hat{h}(t_j)][1 - \hat{h}(t_{j-1})][1 - \hat{h}(t_{j-2})] \dots [1 - \hat{h}(t_1)]. \tag{10}$$

Therefore, if censoring occurs it is possible to derive the estimated SF directly from the estimated HF, but not vice versa.

Generally speaking, at the initial point in time no one can have passed away previously and thus the SF is set to one per definition for the first interval. The lower bound of the SF is zero. It is important to notice that even though the HF can decrease, increase or remain unchanged, the SF will never increase, even though it does respond to the movements in the HF. The SF only remains unchanged, if no individual experiences the event of interest in a given interval.[42]

From the SF the estimated median lifetime can be derived directly. This value, represented by the SF being equal to 0.5, describes the interval in which 50 percent of the individuals have already experienced the event of interest and the other half did not. Due to the fact that this analysis is based on discrete-time data, most likely the median lifetime estimate will not be exact, but only approximated between two intervals. A method applied to handle this issue is interpolation, which is especially useful to compare estimated median lifetimes among different groups or subsamples.[43]

In addition, the CHF $H(t_{ij})$ can also be directly derived from the SF in the following form:

$$H(t_j) = 1 - S(t_j). \tag{11}$$

The individual-specific index has been omitted due to practicality concerns as it has been done previously with the survivor and the HF. The CHF is used in the context of analyzing continuous-time data and therefore mentiond for completeness and because it remains to be debatable whether time measured in year intervals is to be considered a continuous or discrete-time measured variable.[44] For the underlying analysis it will be assumed that survival time is indeed measured continuously.[45]

In table 2 on the following page, a LT has been created using the SOEP data set.[46]

[42] Singer and Willett (2003), p. 334 ff.
[43] Singer and Willett (2003), p. 337 ff.
[44] Singer and Willett (2003), p. 488 ff.
[45] Cleves, Gould, and Gutierrez (2002), p. 140.
[46] The interval [104,108) has been excluded from the LT because no one died in this interval and no one was subject to censoring. Thus all variables in the table will remain unchanged for this interval.

Table 2: Life Table for the SOEP data with n=75,574

Interval	Risk Set R_j	Died E_j	Censored C_j	Survival	Cum. Failure	Hazard
[0, 2)	75,574	1	321	1.0000	0.0000	0.0000
[2, 4)	75,252	6	1,236	0.9999	0.0001	0.0000
[4, 6)	74,010	2	1,415	0.9999	0.0001	0.0000
[6, 8)	72,593	4	1,473	0.9998	0.0002	0.0000
[8, 10)	71,116	3	1,514	0.9998	0.0002	0.0000
[10, 12)	69,599	0	1,526	0.9998	0.0002	0.0000
[12, 14)	68,073	2	1,635	0.9998	0.0002	0.0000
[14, 16)	66,436	3	1,693	0.9997	0.0003	0.0000
[16, 18)	64,740	4	1,752	0.9996	0.0004	0.0000
[18, 20)	62,984	9	1,930	0.9995	0.0005	0.0001
[20, 22)	61,045	15	2,292	0.9992	0.0008	0.0001
[22, 24)	58,738	9	2,372	0.9991	0.0009	0.0001
[24, 26)	56,357	10	2,288	0.9989	0.0011	0.0001
[26, 28)	54,059	5	2,154	0.9988	0.0012	0.0000
[28, 30)	51,900	15	1,924	0.9985	0.0015	0.0001
[30, 32)	49,961	13	1,893	0.9983	0.0017	0.0001
[32, 34)	48,055	15	1,877	0.9979	0.0021	0.0002
[34, 36)	46,163	19	1,753	0.9975	0.0025	0.0002
[36, 38)	44,391	16	1,767	0.9972	0.0028	0.0002
[38, 40)	42,608	39	1,878	0.9962	0.0038	0.0005
[40, 42)	40,691	26	2,046	0.9956	0.0044	0.0003
[42, 44)	38,619	29	2,165	0.9948	0.0052	0.0004
[44, 46)	36,425	53	2,197	0.9933	0.0067	0.0008
[46, 48)	34,175	61	2,247	0.9915	0.0085	0.0009
[48, 50)	31,867	57	2,194	0.9896	0.0104	0.0009
[50, 52)	29,616	87	2,132	0.9866	0.0134	0.0015
[52, 54)	27,397	105	2,013	0.9827	0.0173	0.0020
[54, 56)	25,279	98	1,969	0.9787	0.0213	0.0020
[56, 58)	23,212	144	1,805	0.9724	0.0276	0.0032
[58, 60)	21,263	158	1,664	0.9649	0.0351	0.0039
[60, 62)	19,441	147	1,662	0.9573	0.0427	0.0040
[62, 64)	17,632	202	1,658	0.9458	0.0542	0.0060
[64, 66)	15,772	222	1,442	0.9318	0.0682	0.0074
[66, 68)	14,108	226	1,203	0.9162	0.0838	0.0084
[68, 70)	12,679	223	1,368	0.8992	0.1008	0.0094
[70, 72)	11,088	244	1,393	0.8781	0.1219	0.0119
[72, 74)	9,451	291	1,341	0.8490	0.1510	0.0169
[74, 76)	7,819	327	1,149	0.8107	0.1893	0.0231
[76, 78)	6,343	338	942	0.7640	0.2360	0.0296
[78, 80)	5,063	342	723	0.7084	0.2916	0.0377
[80, 82)	3,998	325	601	0.6461	0.3539	0.0460
[82, 84)	3,072	303	530	0.5764	0.4236	0.0571
[84, 86)	2,239	255	354	0.5051	0.4949	0.0659
[86, 88)	1,630	252	311	0.4188	0.5812	0.0934
[88, 90)	1,067	190	215	0.3359	0.6641	0.1099
[90, 92)	662	163	129	0.2442	0.7558	0.1579
[92, 94)	370	96	67	0.1746	0.8254	0.1664
[94, 96)	207	72	34	0.1084	0.8916	0.2338
[96, 98)	101	40	16	0.0618	0.9382	0.2740
[98, 100)	45	19	11	0.0321	0.9679	0.3167
[100, 102)	15	8	3	0.0131	0.9869	0.4211
[102, 104)	4	2	1	0.0056	0.9944	0.4000
[108, 110)	1	0	1	0.0056	0.9944	0.0000

The first column of table 2 contains information regarding the intervals. For functionality reasons, an interval length of two has been arbitrarily selected. This column corresponds to the time variable in the data set. The interval length remains constant throughout all intervals, but it is important to notice that because no

individual holds a survival time $T_i \in [104, 108]$ independent of whether an individual passed away at that age or was subject to censoring, those two intervals have not been included in forming the LT.[47]

The second column comprises the risk set and thus holds the number of individuals who are at the beginning of the interval at risk of experiencing the event of interest. For the first interval the risk set R_j is equivalent to the total number of individuals participating in the SOEP study. The risk set of the following year R_{j+1} is described in equation 2. It is calculated by subtracting the number of individuals who passed away in the previous interval and the amount of people who were subject to censoring in the interval preceding the current interval from the risk set of the previous year. For example, for the third interval the risk set can be calculated as:

$$\begin{aligned} R_3 &= R_2 - E_2 - C_2 \\ R_3 &= 75,252 - 6 - 1,236 \\ R_3 &= 74,010. \end{aligned}$$

Therefore, the risk set can only decrease over time, but will never increase for the data at hand.

The third column of the table summarizes how many individuals passed away in a given interval. In this column the LT shows the feature that is often recognizable when looking at infant mortality. The absolute number of individuals who passed away increases before the age four and then decreases from age four until the sixth interval. Beyond the age of ten besides small fluctuations the overall trend is increasing in the number of individuals who passed away in the intervals. It is already from the LT visible that between the survival times of 62 and 88 years, the number reaches above 200 cases per interval and reaches the highest absolute values above 300 individuals between the age of 74 and 84 years. From the age of 84 on, the number of individuals who pass away per interval declines, which is caused by the fact that the risk set is much smaller for these high-age intervals.

[47] STATA reports the standard errors and the confidence intervals for the results in the LT. These are not of primary interest in obtaining a first impression of the shape of the survivor or CHF and thus these have not been reported in the table.

The fourth column describes the number of individuals who are exposed to censoring in the given interval. This number is increasing just below the mid-twenty years and then declines until the mid-thirties. Afterwards the number of censored individuals rises until the mid-forties are reached and then the number declines continuously to one in the last interval. Thus it looks like there can be two peak areas identified in the censored observations. The peaks occur in the mid-twenties and mid-forties, where panel attrition solely stemming from censoring is at its highest point.[48]

Column five then describes the SF, which is one for the first interval and then slowly approaches zero in the last interval. The overall trend of the SF is decreasing as expected and will never be increasing. The median lifetime according to the SOEP lies between the intervals [84,86) and [86,88). The sixth column describes the CHF and increases continuously over the intervals. The seventh column contains the HF.

From the LT described above, graph 4 has been derived.

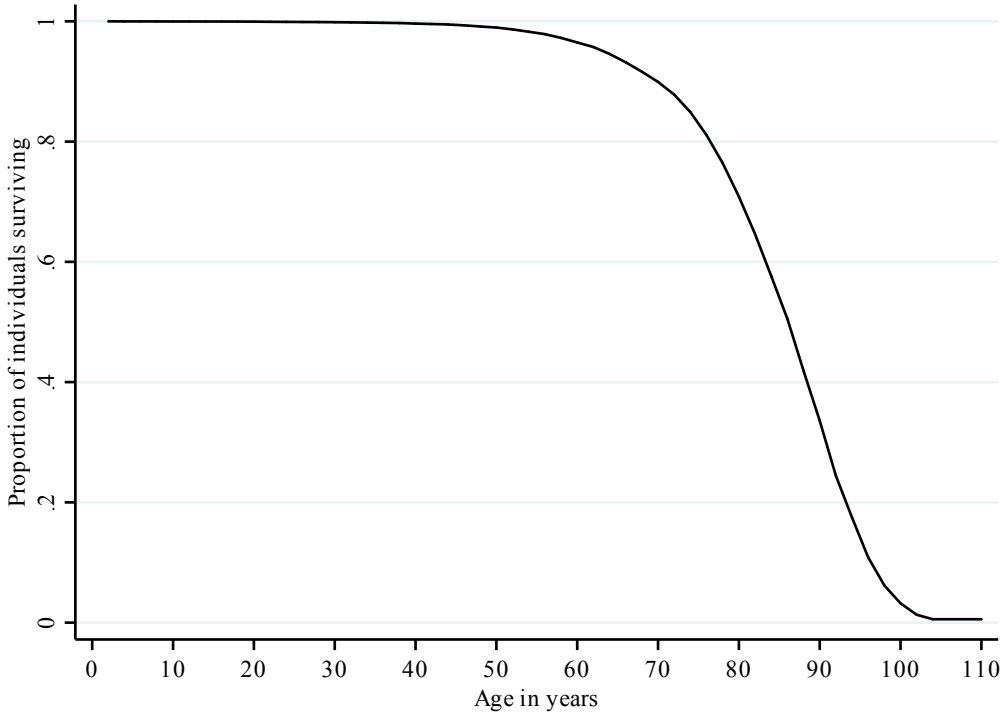

Figure 4: Survivor Function derived from the Life Table

[48]Panel attrition is defined as individuals who do not continue their participation in the survey either temporarily or permanently due to various reasons, for a detailed definition cf. for example Engel and Reinecke (1994), p. 255 ff.

The graph depicts the SF for the SOEP data and is also based upon two-year intervals, just like the LT. As it has been visible and observed in the LT already, the function is monotonically decreasing. The function remains fairly steady at a very high level close to one until the age of 60 and then declines rapidly beyond that point.

The CHF on the other hand, has been visualized in graph 5 and portrays a mirror image of the SF.

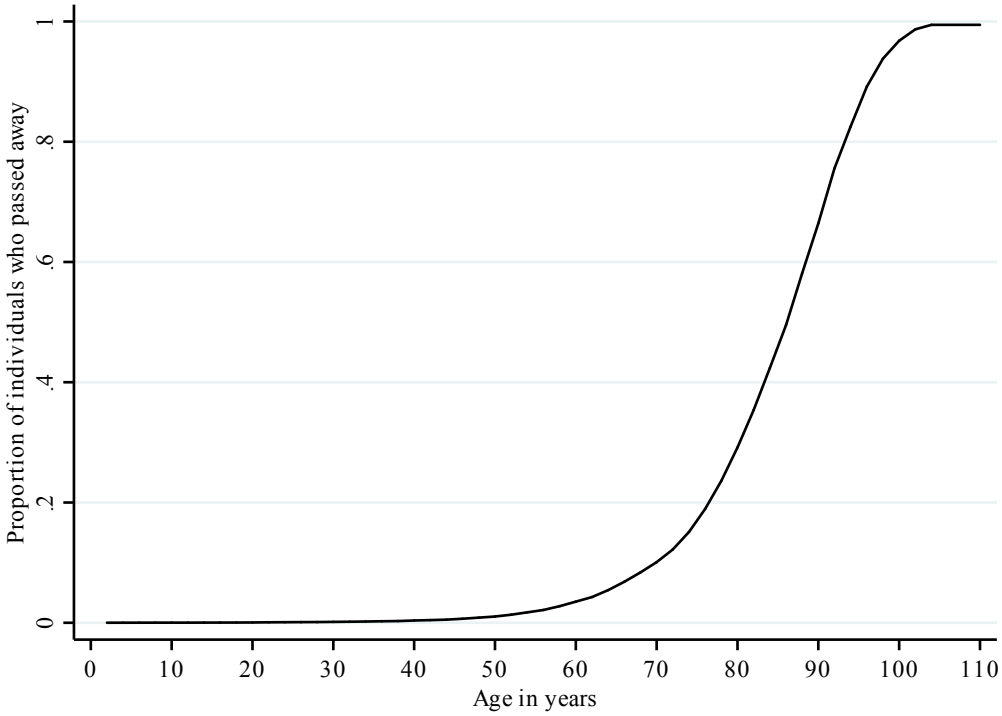

Figure 5: Cumulative Hazard Function derived from the Life Table

Therefore, the pattern in the survivor and CHF graphs as well as the data in the LT itself so far look as expected. In addition, it is widely known that the male mortality lies above the mortality of females for all age groups and is especially visible beyond the age of 40 years.[49]

[49]Luy (2002), p. 3 ff; Rößger (2015), p. 21.

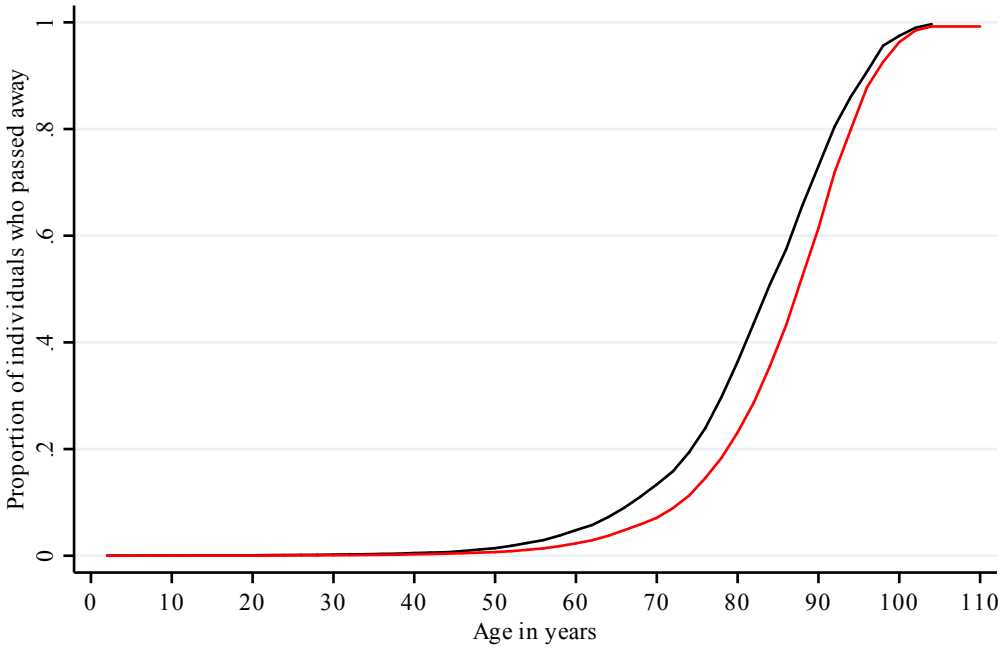

Figure 6: Gender-specific Cumulative Hazard Function with the black curve representing men and the red curve women

Graph 6 shows the CHF for both male (black curve) and female (red curve) survey participants. Because this graph contains all age groups, the fact that infant mortality for boys lies above the infant mortality rate of girls cannot be seen in this particular graph. However, when the graph is modified in such a way that only individuals below the age of 15 are to be considered, the difference in child mortality is also visible in the SOEP data. The graph also illustrates that men die on average at a younger age than females do with the males' curve ending before the females' curve.

To sum this chapter up, the SOEP data behaves in terms of the LT, the survivor and CHF as expectd. This can be seen in the graphs because the overall structure of the probability of passing away and also the gender and age differences in mortality meet the expectations.

3.3 Comparison of Survivor Functions using the Log-rank test

After obtaining a first impression of the data in the previous section by analyzing the graphs of the SF and CHF, the next step in the analysis is to carry out non-parametric tests based on subgroups in the data to compare the SFs for these subgroups. This

allows the researcher to obtain an idea about possible covariates that could be used as possible predictors about the influence on mortality. Standard statistical procedures such as the two-sample t-test, the non-parametric rank sum test and the one-way analysis of the variance can only be applied, if the data does not contain any censored observations. But the SOEP does consist of censored observations and thus adaptations to these tests have to be used in order to compare the SFs. The procedure on how to compare SFs graphically has been demonstrated using the example of gender differences in chapter 3.2. The purpose now is to identify whether these differences are significant.[50]

Generally speaking, the methods to compare the SFs can be split into two major groups. The first one is based on the calculation of confidence intervals and identifying whether these confidence intervals overlap for the distinct groups. This will not be discussed as the confidence intervals are very narrow for the SOEP data and thus graphically difficult to interpret and prone to misinterpretation. The second method relies on the calculation of test statistics. The most popular tests in this area are: the log-rank test and the Wilcoxon test, which can be found in the literature in various adaptations such as the Breslow, Taron-Ware and the Peto-Prentice version. These tests only differ in the weight that they put on the difference between observed and expected failure times. The latter methods can be carried out using the KME as a basis for the analysis.[51]

It is important to notice that the second method relies on global tests. Thus the tests do not test for equality of the SFs at a specific point in time, but compare the overall SFs seen as an entity. They analyze the equality of the expected and observed number of events in each specific group at each failure time and then combine these results over all failure times observed.[52]

In addition, the tests can only be applied to subgroups. This poses a limitation on the method in such a way that metric variables cannot be analyzed using this method. However, it is indeed possible to group the metric variables into subgroups and then apply the method to the subgroups.[53]

[50] Hosmer and Lemeshow (1999), p. 57 ff.
[51] Blossfeld, Golsch, and Rohwer (2007), p. 76 ff.
[52] Cleves, Gould, and Gutierrez (2002), p. 106.
[53] Andreß (1992), p. 155.

The log-rank test is mostly used as a test to compare SFs in the context of survival analysis and can be described as a large-sample χ^2 test. The test contrasts observed and expected failure times for the outcome categories and is based on the ordered failure times. The notation from chapter 3.2 will be retained to explain the working mechanisms of the log-rank test for an example containing two exclusive and non-intercepting groups. The log-rank test can be extended and adapted for more than two groups. The index k, $k = male, female$ separates the two groups into two subgroups, this could for example be the gender difference example that has been explained in the previous section. As a result, the expected counts of failures per group Q_{kj} can be given as:

$$Q_{mj} = \left(\frac{R_{mj}}{R_{mj} + R_{fj}}\right) \times (E_{mj} + E_{fj}) \tag{12}$$

$$Q_{fj} = \left(\frac{R_{fj}}{R_{mj} + R_{fj}}\right) \times (E_{mj} + E_{fj}). \tag{13}$$

The first equation is the calculation of the expected failures for the male group and the second equation is the equivalent for the female group. The first term in parenthesis for both equations describes the proportion of each group in the risk set in interval j. Whereas the second term in parenthesis is the total number of experienced events in interval j summed up for both groups. Then the observed failure times O_k minus the expected failure times X_k per group can be defined as:

$$O_k - X_k = \sum_{j=1}^{p} (E_{kj} - Q_{kj}). \tag{14}$$

The log-rank test statistic itself is given in the following form:

$$Log - rank\ test\ statistic = \frac{(O_k - X_k)^2}{Var(O_k - X_k)} \tag{15}$$

with

$$Var(O_k - X_k) = \sum_{j=1}^{p} \frac{R_{mj} R_{fj} (E_{mj} + E_{fj})(R_{mj} + R_{fj} - E_{mj} - E_{fj})}{(R_{mj} + R_{fj})^2 (R_{mj} + R_{fj} - 1)}. \tag{16}$$

The null hypothesis for this test consists of the statement that there are no differences in the SFs in both groups and can be written as follows:

$$H_0 : h_m(t) = h_f(t). \tag{17}$$

Under the null hypothesis it then holds that:

$$Log - rank\ test\ statistic \sim \chi^2(1). \tag{18}$$

The number of degrees of freedom (d.o.f.) depends on how many categories the variable of interest contains. Thus for the two-category gender example the number of degrees of freedom is equal to one and in general terms it can be calculated as:[54]

$$\text{d.o.f.} = \#\ of\ categories\ the\ variable\ contains - 1.$$

The powerfulness of the results of the log-rank test depends on how the null hypothesis is violated exactly. If the SFs are not equal among the groups but instead proportional to each other, the log-rank test holds the most power.[55]

The theory behind the log-rank test can now be applied to the SOEP data in order to identify possible predictors for mortality. The log-rank test has been carried out for numerous variables. Table 3 summarizes the results of the test for the variables that did not lead to significant results.

Table 3: Results of the Log-rank test for non-significant results at the five percent level

Variable	χ^2	d.o.f.	Prob> χ^2
dsiblings	0.29	1	0.5894
dmigback	2.42	1	0.1197
region	1.55	2	0.4604
dmigraine	3.05	1	0.0807
ddementia	0.63	1	0.4282
dotherdis	0.56	1	0.4556
dnodis	1.33	1	0.2494
dstroke	1.91	1	0.1668
dpressure	2.63	1	0.1046
dheart	0.01	1	0.9053
quitsmoking	4.98	6	0.5466
agework	4.40	2	0.1108
dysports	0.49	1	0.4838
motheduc	1.77	2	0.4120

[54]Kleinbaum and Klein (1996), p. 58 ff.
[55]Cleves, Gould, and Gutierrez (2002), p. 108.

Table 4 on the other hand holds the results of the Log-rank test for results that are at least significant at the five percent level.

Table 4: Results of the Log-rank test for significant results at the five percent level

Variable	χ^2	d.o.f.	Prob> χ^2
gender	293.09	1	0.0000
educmax	167.61	3	0.0000
dmarried	55.30	1	0.0000
ddivorce	5.96	1	0.0146
dchildren	7.85	1	0.0051
numchildren	51.42	12	0.0000
dborngerm	8.54	1	0.0035
dgermnat	53.95	1	0.0000
immiyear	25.79	5	0.0001
dasthma	8.75	1	0.0031
ddiabetes	9.04	1	0.0026
dcancer	61.65	1	0.0000
ddepress	5.06	1	0.0245
ddisabled	147.24	1	0.0000
disabledcat	155.97	2	0.0000
agestsmo	38.22	5	0.0000
yearssmoking	14.55	6	0.0241
smokingperday	690.19	5	0.0000
dsmoking	224.12	1	0.0000
incquart	271.25	3	0.0000
agefatherdeath	53.99	7	0.0000
agemotherdeath	115.39	7	0.0000
fatheduc	6.63	2	0.0363
birthyear	230.82	10	0.0000

Both tables are set up identically. The first column contains the name of the variable, the second column the χ^2 value, the third column the corresponding degrees of freedom and the p-value in the last column. The variables have been separated according to whether the results of the log-rank test are significant at the five percent level. The lower table 4 contains variables that passed the test and the upper table 3 holds all variables that did not pass and thus do not have significant differences in the survivor functions for the groups in the variables.

As the results show whether an individual has siblings or not does not seem to cause a difference in the SFs, neither does it make a difference if the individual has some

sort of migration background. In addition, the region an individual lived in, specified as living in East or West Germany or having lived in both parts of the country, does not cause any differences in the SFs. This might be surprising because the average life expectancy among the 16 states in Germany does differ according to the Federal Bureau of Statistics in Wiesbaden.[56]

Most non-deadly diseases such as migraines, dementia, strokes, high blood pressure, heart conditions do not show significant differences in the SFs according to the log-rank test. Furthermore, even though whether an individual smokes or not does demonstrate a significant difference in the SFs, the age when an individual quit smoking does not seem to cause a difference. Additionally, the age at which an individual started to work as well as whether an individual performed active sports in his or her teenage years also does not result in significant differences in the SFs. Surprisingly, the highest level of education the individual's mother achieved does not cause significant differences in the SFs, whereas the individual's father's highest level of education does.

In conclusion, all variables that did not demonstrate significant differences in the SFs will be excluded from the possible set of covariates or predictors from this point on. Only the variables that were significant at least at the five percent level will be considered as possible predictors in the next chapter that focuses on model development.

4 Model Development

When analyzing survival data it is of utmost importance to notice that the simple Ordinary Least Squares (OLS) method cannot be applied to such data. This is mainly caused by the fact that the OLS relies on the assumption that the residuals are distributed normally. However, for most time-to-event data this assumption is not justifiable. In addition, the normal distribution is theoretically defined on the entire real number line whereas failure times can only take on positive values. The latter problem can be dealt with by choosing the variance σ^2 appropriately. Furthermore, right-censoring does not cause a problem with using OLS because linear regression can be modified in such a way that it can handle censoring, for instance by applying a censored normal regression. Thus, only the normality assumption for the distribution of residuals causes a problem. Even though linear regression is quite robust to deviations

[56] Rößger (2015), p. 33.

from normality, it is not robust if the distribution of failure times is not symmetrical, which is often the case with survival data. Thus a more suitable assumption than the normality assumption must be made and would lead to a parametric model.[57]

In the context of survival analysis, non-parametric, parametric and semi-parametric models can be applied for estimations. Just like the log-rank test discussed in chapter 3.3, non-parametric methods do not make any assumptions regarding the shape of the HF or the SFs, depending on which specification is used. Examples for these methods are the KME and the LT that has already been discussed in chapter 3.2. Nevertheless, these methods only allow for a very limited set up of possible covariates, such as for example gender differences and thus are not suitable for a complete analysis of risk factors affecting mortality.[58] The most commonly used semi-parametric model is the Cox Proportional Hazards Model (CPHM). For parametric approaches the Weibull, Gompertz, exponential, log-normal or log-logistic model are widely known.[59]

4.1 Model selection

In order to choose the appropriate model to estimate the influence of covariates on the failure time, it is important to gain an understanding of the shape of the mortality rate distribution among age groups. The age-specific mortality contains a non-monotonic shape, which means that it is neither strictly increasing nor strictly decreasing over all ages. This is easily validated by the fact that due to higher infant mortality, the mortality rate starts off at a given positive point and then decreases for individuals until the age of ten. Between the ages of ten and 30 the mortality rate curves upwards and has a more concave form. The overall shape of the age-specific mortality rate, however, is more comparable to a convex shape. From the age of 30 onwards, the mortality rate increases steadily over time. The increase when using a log-transformed scale of the mortality rate matches an approximately linear distribution, which is equivalent to an exponential distribution when the log-transformation is not used. [60]

Due to the fact that the exponential, the Weibull and the Gompertz model all rely on the assumption that the relationship between the mortality rate and the survival

[57] Cleves, Gould, and Gutierrez (2002), p. 2.
[58] Diekmann and Mitter (1984), p. 58 ff.
[59] Blossfeld, Hamerle, and Mayer (1986), p. 48 ff.
[60] Rößger (2015), p. 25.

time or process duration is some sort of a monotonous relationship, these three models do not suffice in measuring the effect of covariates on mortality. Only the exponential model assumes a time independence of the mortality rate and the duration, whereas the Weibull, Gompertz, log-logistic and the log-normal model all allow for time-dependence between mortality rate and duration, which is in the context of mortality equivalent to the age of the individual. The time-dependence assumption seems more realistic when it comes to modeling mortality. On the other hand, the remaining two most commonly used models in the class of parametric models, the log-normal and the log-logistic model, are more flexible in regard to the fact that they allow for a non-monotonous relationship between the mortality rate and the age of the individual. These parametric models only describe parts of the distribution appropriately at a time, for instance the Gompertz model does well in modeling mortality for individuals who are over the age of 30, but does not perform sufficiently for age groups below the age of 30. Because these distributional assumptions do not describe the entire relationship between mortality rate and age, the semi-parametric Cox model will be used for the analysis. [61]

4.2 The Cox model

The CPHM is the most popular semi-parametric model in use for continuous time data. Thus for the analysis at hand it will be assumed that the variable time until an individual passes away or is subject to censoring measured in years, is sufficiently continuous enough to apply the Cox model. Numerous debates have been carried out in order to determine whether time and consequently age are continuous variables. As the name of the model already indicates, it puts its emphasize on the hazard perspective rather than the survivor point of view, which suits the setup of modeling mortality well. Firstly, the general CPHM will be introduced and afterwards it will be modified to fit the SOEP data.

4.2.1 The basic Cox model

In general terms, the CPHM can be written in the following form:

$$h(t, x, \beta) = h_0(t)e^{x\beta} \tag{19}$$

with x being a vector of covariates and β being a vector of parameters that need to be

[61] Blossfeld, Golsch, and Rohwer (2007), p. 88 ff, p. 186 ff.

estimated.[62] The term $h_0(t)$ corresponds to the baseline hazard function, which does not dependent on the vector of predictors X, but on the time t for which the hazard is being predicted. The baseline hazard function owes its name to the fact that if all predictors are equal to zero it holds that:

$$h(t, x, \beta) = h_0(t)e^0 \tag{20}$$

$$h(t, x, \beta) = h_0(t). \tag{21}$$

Thus the baseline hazard can be seen as a Cox model without predictors. The model is referred to as a semi-parametric model because the baseline hazard does not make any assumptions regarding its distribution and is therefore unspecified. Nevertheless, it is possible to estimate the parameter β. This parameter is the only necessary part to calculate the hazard ratios that allow for an effect interpretation of the covariate, whereas the coefficients themselves do not allow for a magnitude interpretation. Thus the baseline hazard function does not need to be estimated to determine the effect of the covariates. The second component focuses on an exponential term e to the sum of the linear term $x\beta$. It is important to notice at this point that only the baseline hazard depends on time variable t while the exponential expression does not because in the basic model the assumption holds that predictors are time-independent. The CPHM allows for an extension of the basic model to include time-dependent covariates and is then called the extended Cox model.[63] Since no time-dependent variables will be considered in estimating the relationship between mortality and possible predictors, this will not be demonstrated.[64]

One of the key factors why the CPHM is so widely used is based on the fact that because the baseline hazard function is unspecified, the results are quite robust in terms of estimated regression coefficients, hazard ratios and adjusted survivor functions and the model will closely resemble the results for the parametric model. Another advantage lies in the second term of equation 19. If one were to only include the linear sum without exponentiating it as the second factor in the product, it would be possible that the hazard could take on negative values. But due to the usage of the exponential expression, the hazard is only defined as being larger or equal to zero and finite. This complies with the former definition of the hazard being non-negative.[65]

[62] Cf. Cox (1972), p. 189 ff; Kalbfleisch and Prentice (1980), p. 97.
[63] Kleinbaum and Klein (1996), p. 94 ff.
[64] For further readings on the extended Cox model cf. Kleinbaum and Klein (1996), p. 211 ff, Chapter 6.
[65] Kleinbaum and Klein (1996), p. 97 ff.

The parameters are estimated using a maximum likelihood procedure; in particular it focuses on a partial likelihood approach, which was first introduced by D.R. Cox in 1972. It is called a partial likelihood estimation because the likelihood is created in such a way that only probabilities of the individuals who experienced the event, which in the underlying analysis is death, will be included in the likelihood. While the probabilities of censored individuals will not be taken explicitly into consideration the survival times of censored individuals will be included in a sense as they are part of the risk set at each given failure time and therefore indirectly included in the likelihood.[66] The likelihood itself can be derived from the density function that is generated by the product of the HF and the SF given in the following form:

$$f(t, x, \beta) = h(t, x, \beta) \times S(t, x, \beta). \tag{22}$$

This formula relies on the fact that the following holds:

$$S(t) = e^{-H(t)}. \tag{23}$$

Because time is assumed to be continuous, it is possible to express this using the CHF as follows:

$$H(t) = \int_0^t h(u)du. \tag{24}$$

Due to equation 23 this can also be written in terms of the SF:

$$S(t) = e^{-\int_0^t h(u)du}. \tag{25}$$

Now taking the log of the equation yields:

$$h(t) = \frac{f(t)}{S(t)}. \tag{26}$$

The equation above then leads back to equation 22.[67] Due to the fact that this can be expressed as in equation 22 and because the observations are assumed to be independent

[66] Kleinbaum and Klein (1996), p. 99 ff.
[67] Hosmer and Lemeshow (1999), p. 73, 81 ff.

of each other, the likelihood function can be written in the following form:

$$l(\beta) = \prod_{i=1}^{n} \{[h(t_i, x_i, \beta) \times S(t_i, x_i, \beta)]^{c_i} \times [S(t_i, x_i, \beta)]^{1-c_i}\} \tag{27}$$

where c_i represents a dummy variable for censoring and using a simple algebraic modification leads to:

$$l(\beta) = \prod_{i=1}^{n} \{[h(t_i, x_i, \beta)]^{c_i} \times [S(t_i, x_i, \beta)]\}. \tag{28}$$

Now the log can be taken and equation 19 and the equivalent function in terms of the SF can be inserted above, which leads to the log-likelihood function:

$$L(\beta) = \sum_{i=1}^{n} \{c_i \ln[h_0(t_i)] + c_i x_i \beta + e^{x_i \beta} \ln[S_0(t_i)]\}. \tag{29}$$

Now in order to use a complete maximum likelihood approach, the parameter β, the baseline HF and the SF have to be estimated. Because the model is chosen in such a way that the error component does not have to be specified directly, the log-likelihood function in equation 29 cannot be used. Cox instead suggested using a partial likelihood function that relies solely on the estimation of parameter β and the partial likelihood can be written as follows:

$$l_p(\beta) = \prod_{i=1}^{n} \left[\frac{e^{x_i \beta}}{\sum_{j \in R(t_i)} e^{x_i \beta}} \right]^{c_i} \tag{30}$$

where $R(t_i)$ represents the number of people in the risk set at time t_i.[68]

4.2.2 The Cox model with tied survival times

The equation above does not allow for tied observations. Due to the setup of the time variable in the SOEP, an adjustment has to be made in order to allow for tied failure times, which means that at least two individuals died at the same age. Multiple methods have been introduced to handle tied failure times in the literature, the most popular methods are the exact method introduced by Kalbfleisch and Prentice (1980) and the approximations by Breslow (1974) and by Efron (1977). Especially the Breslow method assumes that the number of ties for each failure time is sufficiently small because if the number is too large a discrete model should be applied instead, but this concept remains

[68] Hosmer and Lemeshow (1999), p. 92 ff.

rather subjective. If the number of ties is too large, the Breslow method will result in biased estimates. The exact method and the Efron approximation tend not to be biased even if the number of ties is relatively large. For demonstration purposes the Breslow and Efron approximations of the partial likelihood function will be given in the case that only non-censored observations exist, which leads to a simplification of equation 30 and looks as follows:

$$l_p(\beta) = \prod_{i=1}^{m} \left[\frac{e^{x_{(i)}\beta}}{\sum_{j \in R(t_{(i)})} e^{x_j \beta}} \right] \tag{31}$$

where m are the non-tied and non-censored observations ordered according to their failure time.[69]

The Breslow approximation for a single covariate then leads to the following specification of the partial likelihood:

$$l_p^B(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)+}\beta}}{\left[\sum_{j \in R(t_{(i)})} e^{x_j \beta}\right]^{d_i}} \tag{32}$$

where d_i represents the number of individuals with survival time $t_{(i)}$ and $x_{(i)+}$ being the sum of the single covariate over the individuals with survival time $t_{(i)}$ which can be written as $x_{(i)+} = \sum_{j \in D(t_{(i)})} x_j$ and $D(t_{(i)})$ being the individuals with survival time $t_{(i)}$.[70]

The Efron approximation is known to give a slightly better approximation of the exact partial likelihood function, but its functional form is a bit more complex than the Breslow approximation and can be written as:

$$l_p^E(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)+}\beta}}{\prod_{k=1}^{d_i} \left[\sum_{j \in R(t_{(i)})} e^{x_j \beta} - \frac{k-1}{d_i} \sum_{j \in D(t_{(i)})} e^{x_j \beta}\right]}. \tag{33}$$

It is noticeable that the Breslow and Efron approximation and equation 31 are identical if $d_i = 1$.[71]

[69] Hosmer and Lemeshow (1999), p. 95 ff.
[70] Blossfeld, Golsch, and Rohwer (2007), p. 225 ff.
[71] Hosmer and Lemeshow (1999), p. 106 ff.

As an example to illustrate how minimal the differences in the estimation results are, when either the exact marginal likelihood, the exact partial likelihood, the Breslow and the Efron methods are used, the variable from the SOEP that contains the information whether an individual has been ever diagnosed with diabetes will be investigated. When the estimation of the CPHM with ties is carried out with only diabetes as the covariate coded as an indicator variable, the results of the four methods mentioned above are summarized in table 5.

Table 5: Results of the modifications of the Cox model for tied survival times

Method	Coefficient	Std. Error	Hazard Ratio	Std. Error	z	P> |z|	LR-χ^2	d.o.f.	Prob> χ^2
Breslow	0.3227409	0.1085784	1.380908	0.1499367	2.97	0.003	8.36	1	0.0038
Efron	0.3236013	0.1085700	1.382096	0.1500542	2.98	0.003	8.41	1	0.0037
Exact m.l.[72]	0.3236313	0.1085755	1.382138	0.1500663	2.98	0.003	8.41	1	0.0037
Exact p.l.[73]	0.3277114	0.1094870	1.387788	0.1519448	2.99	0.003	8.49	1	0.0036

Column one describes the method that has been applied in order to account for tied survival times in the model. The next column presents the estimated coefficients with the resulting standard errors in the column beside the coefficient. The hazard ratio is given in the fourth column and the corresponding standard errors are reported in the fifth column. Chapter 4.3 will explain how to interpret the hazard ratios themselves and the reasoning on why they need to be reported when interpreting the results of the model in the context of survival analysis, but for now this is relevant to the purpose of comparing the four methods for handling tied observations. In column six the z-value is reported and in the next level whether the estimate is significant. The last three columns contain the results of the Likelihood ratio test (LR) test: the value of the statistic, the number of degrees of freedom and whether the results are significant.

The LR test compares whether the inclusion of the variable, here the indicator variable diabetes, is desirable. Thus, if the test results are significant, the null hypothesis that the inclusion of the variable does not significantly improve the fit of the model can be rejected.[74] It is easily visible in the results of the four different methods that they only differ slightly. The values of the coefficients and the corresponding hazard ratios start differing only at the third decimal place. Both p-values and the z-value are almost identical. Only the value of the LR test statistic and the standard errors differ slightly more, but are still very similar across the four methods. In conclusion and

[72] marginal likelihood.
[73] partial likelihood.
[74] Blossfeld, Golsch, and Rohwer (2007), p. 97 ff.

because the estimates are so similar, the Efron method will be applied simply due to the fact that it works better with a larger number of ties which holds true for the SOEP data.

Then, like for all maximum likelihood approaches, the log-likelihood function with the appropriate modifications is differentiated with respect to the parameter β. Afterwards the derivative has to be set equal to zero and then solved for the unknown parameter. The estimator of the variance of the estimator of the coefficient is then derived by taking the negative of the second derivative of the partial log-likelihood function also with respect to the parameter β and computing the inverse of the expression, which is in the literature referred to as the observed information matrix.[75]

4.3 Fitting the Cox model to the SOEP data

After the theoretical concept behind the CPHM has ben explained in the previous section, the model has been estimated using different specifications in the sense that different covariates of the ones who passed the log-rank test in chapter 3.3 have been included in the model. The model that did not produce any insignificant results at the ten percent significance level incorporates ten different covariates and the results of the estimation using the Efron approximation are summarized in table 6.

Table 6: Estimation output using the Cox model in terms of the coefficients

Variable	Coefficient	Std. Error	z	P> $\|z\|$	95% Conf.	Interval
gender	-0.294847	0.0888451	-3.32	0.001	-0.4689803	-0.1207140
birthyear	0.104729	0.0085351	12.27	0.000	0.0880009	0.1214580
educmaxuni	-0.267981	0.1183096	-2.27	0.024	-0.4998638	-0.0360989
ddivorce	0.310676	0.1121316	2.77	0.006	0.0909024	0.5304501
ddisabled	0.584359	0.0799146	7.31	0.000	0.4277288	0.7409883
dsmoking	0.236009	0.0870587	2.71	0.007	0.0653775	0.4066413
averhhpgi	-0.000004	0.0000021	-1.68	0.093	-0.0000077	0.0000006
agemotherdeath	-0.008596	0.0025966	-3.31	0.001	-0.0136855	-0.0035070
fatheducprac	-0.218286	0.0938547	-2.33	0.020	-0.4022372	-0.0343337
fatheducuni	-0.557475	0.2053393	-2.71	0.007	-0.9599325	-0.1550171

The estimation was carried out for a total number of 7,347 observations of which 651 individuals passed away.[76] The total number of observations used for the estimation

[75] Hosmer and Lemeshow (1999), p. 96 ff.

[76] The results are based on 7,437 individuals with 651 failures, LR-$\chi^2(10)$=313.66, Prob> χ^2=0.000 and a log-likelihood=-4,432.1341.

differs from the total number of observations in the modified data set due to the fact that only the 7,347 individuals included in the estimation demonstrated no missing values in the variables used for the estimation. This statement is equivalent to the fact that everyone in the data set has been excluded from the estimation if the individual missed at least one value in the variables utilized in the model.

The first column of table 6 contains covariates that have been used in the model estimation. It is specific to the CPHM that no intercept or constant is reported in the estimation results, which is different from for example OLS. This is caused by the fact that the constant is included in the baseline hazard function that will not be defined in the estimation. Consequently, the constant cannot be identified from the data because the value will not matter.[77]

The second column contains the estimated coefficient for each covariate. It is important to notice here that only the sign of the coefficient can be interpreted at this point, but not the magnitude or in other words the value of the coefficient. For example, $\hat{\beta}_{gender} = -0.294847$ only represents that women ceteris paribus have a lower mortality rate than men.[78] However, the estimated coefficient does not provide any information yet on how much lower the rate is. The interpretation of the sign of the estimated coefficient is analogous for the variables educmaxuni, ddivorce, ddisabled, dsmoking, fatheducprac and fatheducuni since they are also coded as indicator variables.[79] For the continuous variable agemotherdeath the interpretation differs slightly. $\hat{\beta}_{agemotherdeath} = -0.008596$ states that for each additional year the mother of the individual lived, the mortality rate of the individual decreases. The reverse interpretation holds for the variable birthyear of the individual, which is highly counter intuitive and not likely to be true. In addition, the estimated coefficient of the average household post-government income of the individual only differs minimally from zero and is only significant at the ten percent level, whereas all other covariates in the model are significant at least at the five percent level which is summarized in column five of the table. Averhhpgi is also the only covariate in the model where the 95 percent confidence interval, which is given in the sixth and seventh column of the table, includes zero. Thus, this result seems highly questionable. For completeness purposes,

[77] Cleves, Gould, and Gutierrez (2002), p. 115.
[78] Remember, the gender variable has been coded in a way that 1 represents a female and 0 a male.
[79] It is important to keep in mind that categorical variables have to be recoded in a fashion of an indicator variable before they can be included in the estimation.

it is also important to mention that column three contains the standard errors for the estimated coefficients and that column four presents the z-values that can be calculated by dividing the estimated coefficient of the covariates by the corresponding standard error.

In order to consider the interpretation of the magnitude of the effects described above, it is now necessary to summarize the Hzard Ratios (HR)s, which are expressed in table 7.[80]

Table 7: Estimation output using the Cox model in terms of the hazard ratios

Variable	Hazard Ratio	Std. Error	z	P> $\|z\|$	95% Conf.	Interval
gender	0.744645	0.0661581	-3.32	0.001	0.6256399	0.8862874
birthyear	1.110410	0.0094775	12.27	0.000	1.0919890	1.1291420
educmaxuni	0.764922	0.0904976	-2.27	0.024	0.6066133	0.9645449
ddivorce	1.364347	0.1529864	2.77	0.006	1.0951620	1.6996970
ddisabled	1.793840	0.1433540	7.31	0.000	1.5337700	2.0980080
dsmoking	1.266186	0.1102325	2.71	0.007	1.0675620	1.5017650
averhhpgi	0.999997	0.0000021	-1.68	0.093	0.9999923	1.0000010
agemotherdeath	0.991441	0.0025744	-3.31	0.001	0.9864077	0.9964992
fatheducprac	0.803896	0.0754494	-2.33	0.020	0.6688221	0.9662490
fatheducuni	0.572653	0.1175883	-2.71	0.007	0.3829187	0.8564005

The concept of the Hazard Ratio (HR) can be described as a relative risk ratio and is comparable to the odds ratio concept in the context of logistic regression. The estimated HR can be calculated using the estimated coefficient in the following form:

$$\hat{HR}_{covariate} = e_{\hat{\beta}_{covariate}}. \tag{34}$$

Using the covariate gender as an example for the calculation of the HR, the HR can be derived as:

$$\begin{aligned} \hat{HR}_{gender} &= e^{\hat{\beta}_{covariate}} \\ \hat{HR}_{gender} &= e^{-0.294847} \\ \hat{HR}_{gender} &= 0.744645 \end{aligned}$$

which is equivalent to the estimated HR for the covariate gender given in table 7. The

[80]The results are based on 7,437 individuals with 651 failures, LR-χ^2(10)=313.66, Prob¿χ^2=0.000 and a log-likelihood=-4,432.1341.

composition of table 7 is equivalent to the structure in table 6, only column two differs because it contains now the estimated HRs instead of the estimated coefficients before. Therefore, the interpretation of the columns is equivalent as well and thus only column two will be interpreted. The estimated HR for the covariate gender states that because it holds that $\hat{HR}_{gender} < 1$ the average mortality risk for females is lower than the ones for males ceteris paribus, which is exactly what the sign of the estimated coefficient suggested before. But now the magnitude can be interpreted as percentage differences for groups defined by indicator variables and unit-changes leading to percentage changes as well for continuous variables. For the variable age the mother of the individual passed away, the interpretation of the estimated HR can be described as for each additional year the mother lived the mortality rate decreases by about 0.9 percent. It is interesting to see that the education levels of the individuals and the educational achievements of the individual's father both have positive implications on the risk and therefore lower mortality. Surprisingly, the impact of having experienced a divorce on the mortality rate is higher than the influence of smoking, but both covariates do extend the magnitude of the influence of a disability status.

The LR test statistic that is also recorded in the output of the estimation tests the null hypothesis that the covariates that have been additionally included into the model do not significantly improve the fit of the model compared to a model without any covariates. Under the null hypothesis the LR test statistic follows a χ^2 distribution with m degrees of freedom, where m represents the number of covariates that have been included in the model additionally. The null hypothesis can be rejected since the result of the test statistic is significant even at the one percent level. Consequently, at least one of the included covariates significantly improves the fit of the model.[81]

5 Goodness of fit

All results presented in the previous section have to be used with caution and this section will explain the reasoning why this is the case. Even though nine out of the ten variables used in the model above have demonstrated an impact in a direction that was expectable, the result for the birthyear was surprising, not to say illogical because the average life expectancy in Germany is steadily increasing since the first official reporting of it in 1870. This increase can clearly be seen in figure 7 below, where the women's curve

[81] Blossfeld, Golsch, and Rohwer (2007), p. 97 ff.

in red lies strictly above the men's curve in black.[82] On the contrary, the result of the CPHM suggests that for each additional year in the birthyear, which can be understood as being born a year later, the mortality rate increases.

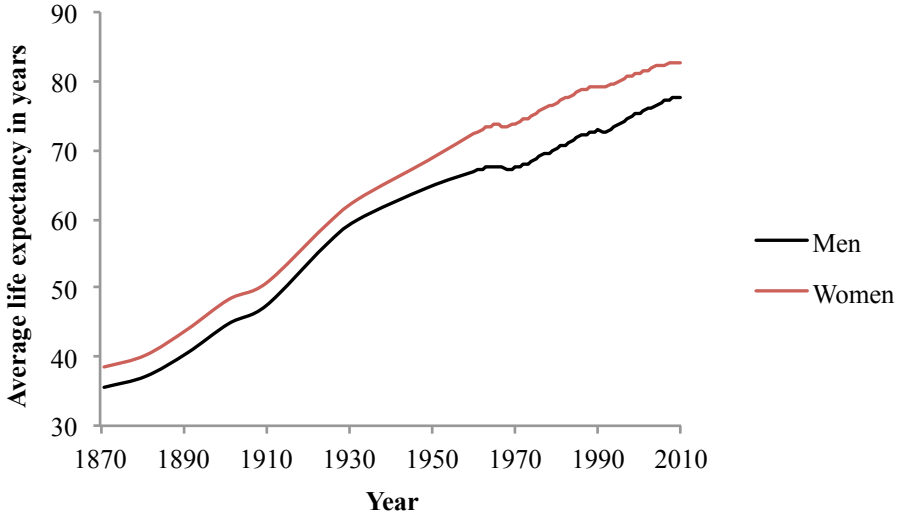

Figure 7: Average life expectancy at birth in years from 1870 to 2010 using data from the Federal Bureau of Statistics

5.1 Testing the proportional hazards assumption

The CPHM relies on the assumption that the hazard rates are in fact proportional for certain defined groups over time. For instance, the model assumes for the hazard rate for men and women separately that their group-specific hazard rates are in fact proportional over time. This assumption can be tested in the context of assessing the Goodness of fit (GOF) of the model, either graphically or by applying a test procedure. If a graphical approach is applied, in the example for differences in the hazard rate among males and females, the curves for both groups can only differ among the y-axis, which represents the hazard rate. The curves cannot be shifted among the x-axis holding the survival time, which is in the context of modeling mortality equivalent to the age of the individual.

Therefore, the curves cannot intercept in order to justify that the proportionality assumption holds and have to exhibit some form of parallelism. It is either possible to

[82] The data used to develop this graph was obtained from the website of Germany's Federal Bureau of Statistics. For the years 1949 to 1986, only data for West Germany was available, whereas all other years contain West and East Germany combined. Cf. Statistisches Bundesamt (2015).

compare predicted and observed SFs, respectively HFs, or to take the negative log of the negative log of the SFs or HFs over different categories of variables, for example for males and females. Because this graphical approach is highly subjective and cannot be used with continuous covariates unless grouped accordingly, it will not be carried out in this analysis.[83]

The second procedure in order to test the GOF of the model and thus whether the proportionality assumption holds or not relies on carrying out a GOF test, which can be done by applying a χ^2 test. If the test results indicate that there is indeed a significant interaction between a covariate and survival time, the proportionality assumption for this covariate is most likely not valid.[84]

In order to perform the test, it is necessary to define the residuals. In a linear regression setting the residuals would be calculated as the difference between observed and predicted residuals. This is not possible in the setting of dealing with survival data since the fitted model does not provide an estimate of the mean of the dependent variable and also the value of the outcome is unknown when a partial likelihood function is applied to censored data. There exist no clean-cut solution to the definition of residuals in the survival analysis context. Therefore, many different approaches have been introduced in the literature. The first method of a definition for the residuals was suggested by Schoenfeld in 1982 and it focuses on taking the log of equation 30 and then taking its derivative and leads to the following expression:

$$\frac{\partial L_p(\beta)}{\partial \beta} = \sum_{i=1}^{n} c_i \left(x_i - \frac{\sum_{j \in R(t_i)} x_j e^{x_j \beta}}{\sum_{j \in R(t_i)} e^{x_j \beta}} \right). \tag{35}$$

This expression will now be defined for the k^{th} covariate to show the individual contributions of the covariates to the partial log likelihood:

$$\frac{\partial L_p(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} c_i \left(x_{ik} - \frac{\sum_{j \in R(t_i)} x_{jk} e^{x'_j \beta}}{\sum_{j \in R(t_i)} e^{x'_j \beta}} \right). \tag{36}$$

[83] Kleinbaum and Klein (1996), p. 129 ff.
[84] Blossfeld, Golsch, and Rohwer (2007), p. 233 ff.

This expression can be simplified in the following form:

$$= \sum_{i=1}^{n} c_i \left(x_{ik} - \bar{x}_{w_i k} \right) \tag{37}$$

where

$$\bar{x}_{w_i k} = \frac{\sum_{j \in R(t_i)} x_{jk} e^{x'_j \beta}}{\sum_{j \in R(t_i)} e^{x'_j \beta}}. \tag{38}$$

The estimator of the Schoenfeld residuals $\hat{r}_{ik}$ can then be obtained for the i^{th} individual in the k^{th} covariate by plugging the partial likelihood estimator of the coefficient into equation 36:

$$\hat{r}_{ik} = c_i \left(x_{ik} - \hat{\bar{x}}_{w_i k} \right) \tag{39}$$

with

$$\hat{\bar{x}}_{w_i k} = \frac{\sum_{j \in R(t_i)} x_{jk} e^{x'_j \beta}}{\sum_{j \in R(t_i)} e^{x'_j \beta}}. \tag{40}$$

The sum of the Schoenfeld residuals is equal to zero because the estimator $\hat{\beta}$ is obtained by setting the derivative of the partial log likelihood in equation 35 also equal to zero. In addition, the test is carried out in such a way that the values of the estimates of the Schoenfeld residuals are set to missing for individuals who are subject to censoring.[85]

Since the Schoenfeld method is the most commonly used approach for the residuals in the context of survival analysis, it will be used in carrying out the test on whether the proportionality assumption in the CPHM holds. The results of the test statistic using the model with ten covariates from chapter 4.3 are summarized in table 8.

Table 8: Global test results for 10 covariates

χ^2	d.o.f.	Prob> χ^2
88.11	10	0.0000

The GOF test relies on a $\chi^2 - statistic$ with one d.o.f. The number of d.o.f. given in the table itself refers to the ten covariates that have been included in the model.

[85] Blossfeld, Golsch, and Rohwer (2007), p. 196 ff.

Given the p-value above of 0.0000, the result is significant at any possible significance level and therefore the assumption of proportional hazards is violated. Since this is a global test, it is only possible to test all ten covariates simultaneously or one covariate separately each at a time.

For example, if one were to specify a model that only includes the two covariates ddivorce and fatheducuni, the results of the test appear to be very different. These two covariates have been selected for a further investigation, because if two separate models have been estimated using the CPHM with only one of these two variables at a time, the proportionality assumption does hold. Tables 9 and 10[86] summarize the results of performing the CPHM for the two covariate example.

Table 9: Estimation output using the Cox model in terms of coefficients

Variable	Coefficient	Std. Error	z	P> |z|	95% Conf.	Interval
ddivorce	0.404902	0.0825225	4.91	0.000	0.2431608	0.5666429
fatheducuni	-0.361927	0.1401196	-2.58	0.010	-0.6365562	-0.0872974

Both covariates are still significantly different from zero and also the signs of the coefficients have not changed in comparison to the model that includes ten covariates.

Table 10: Estimation output using the Cox model in terms of hazard ratios

Variable	Coefficient	Std. Error	z	P> |z|	95% Conf.	Interval
ddivorce	1.499155	0.1237140	4.91	0.000	1.2752740	1.7623410
fatheducuni	0.696333	0.0975700	-2.58	0.010	0.5291114	0.9164045

The hazard ratio for the second model is higher for the ddivorce covariate than in the first model. On the other hand, the hazard ratio in the second model for fatheducuni has a lower magnitude than in the larger model. Both estimates remain significant at least at the five percent level in the second model as well.

The following table 11 summarizes the results of the test regarding the holding of the proportionality assumption.

[86] Both results are based on 24,712 individuals with 1,072 failures, LR-$\chi^2(10)$=27.93, Prob> χ^2=0.000 and a log-likelihood=-8,239.6185.

Table 11: Global test results for 2 covariates

χ^2	d.o.f.	Prob> χ^2
0.31	2	0.8567

The p-value for the second model is greatly larger than any commonly used significance level. Therefore, the test indicates that for the two-covariate model the assumption of proportional hazards holds. Nevertheless, the model with two covariates does not describe the best and most complete range of covariates that influence mortality.

If the proportionality assumption is violated, it is necessary to modify the Cox model accordingly. Therefore, in order to incorporate non-proportional hazard rates into the model, the Stratified Cox model (SCM) needs to be applied.

5.2 The stratified Cox model

It is important to notice, that the SCM can only accommodate categorical variables, whereas the CPHM allows for indicator variables, which can also be derived from categorical variables, and metric variables. But it is indeed possible even with the SCM to include metric covariates by grouping the data appropriately. Therefore, the hazard function in the stratified context looks as follows:

$$h_s(t, x, \beta) = h_{s0}(t)e^{x'\beta} \tag{41}$$

where $s = 1, \ldots, S$ defines the strata and thus now the hazard rates can differ for each stratum. This holds also true for the baseline hazard functions.

Because this HF must be estimated simultaneously for all groups, it is necessary to adjust the partial likelihood in such a way that it is now the product of the group-specific likelihoods. The partial likelihood for the s^{th} stratum looks as follows and is not different from the previous definition besides the index for the stratum:

$$l_{sp}(\beta) = \prod_{i=1}^{n_s} \left[\frac{e^{x'_{si}\beta}}{\sum_{j \in R(t_{si})} e^{x'_{sj}\beta}} \right]^{c_{si}} \tag{42}$$

where n_s denotes the number of individuals in the stratum s, t_{si} represents the ith observed valued in the stratum s, c_{si} is the corresponding censoring variable in the stratified model, $R(t_{si})$ are the individuals at risk in stratum s at time t_{si} and finally x_{si}

is the vector of covariates. From this stratum specific representation, one can now obtain the complete stratified partial likelihood function as mentioned before as the product of the stratum-specific likelihood functions:

$$l_p^S(\beta) = \prod_{s=1}^{S} l_{sp}(\beta). \tag{43}$$

Then after taking the log of the complete stratified partial likelihood and differentiating it with respect to the p unknown parameters and setting these derivatives equal to zero, the parameter vector β can be estimated.[87]

Due to the fact that eight out of the ten covariates used in the larger model do not fulfill the proportionality assumption, they would need to be included in the SCM as stratas. Therefore, only the two covariates ddivorce and fatheducuni would be directly included in the model. The effect of the stratified covariates on the dependent variable cannot be measured anymore if this method is applied and thus this kind of model would defeat the purpose of the analysis and is consequently not carried out, but only mentioned as a possibility to handle non-proportional hazards.

[87] Hosmer and Lemeshow (1999), p. 243 ff.

6 Conclusion

Throughout conducting the analysis, it has been clearly carved out how well the modeling of mortality fits into the context of survival analysis. This is caused on the one hand by the occurence of censored and truncated observations, which do not allow for OLS to be applied in order to use the full information available. On the other hand, the specific characteristic of the dependent variable in terms of the age of the individual blends well into the context of survival analysis because it models the time until an event occurs. This constitutes the main purpose of conducting survival analysis.

In addition, the quality of the data set independent of the estimation results remains to be questionable. As seen in figure 5 the median lifetime in the SOEP amounts to approximately 85 years, which is even higher than the average life expectancy at birth in 2010 derived from figure 7. Therefore, this difference hints towards an underestimate of mortality in the SOEP. This deviation stems likely from the low follow-up rate of the SOEP. Furthermore, the issue of inconclusive entries in the SOEP data mentioned in chapter 2.2 also contributes to the questioning of the reliability of the data.

Furthermore, the issue of missing values in the SOEP data has been tried to minimize by combining groups of variables into time-constant and mostly dummy variables to maximize the usage of information at hand. This process has proven to be very time consuming in order to prepare and modify the data prior to performing the estimation. Still, the amount of missing values remained drastically high. Therefore, in a future project imputation methods could be applied to the data to artificially increase the sample size in a reasonable matter.

The results of the Cox model suggest that indeed at least two factors that influence mortality have been identified. For all other covariates, the proportionality assumption has not been met. A strategy on how to handle violations of the proportionality assumption has been introduced in terms of the stratified Cox model. It remains to be debatable whether the violation of the proportionality assumption stems from the quality of the SOEP data or from the unapplicability of the Cox model. In order to fully understand the magnitude of this issue, the method should be applied to a different data set, for instance the data from the German Pension System could be used in order to rule this out, and then the results could be compared in a future project.

References

Allison, Paul D. (1984). *Event history analysis: Regression for longitudinal event data*. 46. Sage.

Andreß, Hans-Jürgen (1992). "Einführung in die Verlaufsdatenanalyse: statistische Grundlagen und Anwendungsbeispiele zur Längsschnittanalyse kategorialer Daten". In: *Historical Social Research/Historische Sozialforschung. Supplement*, pp. 1–323.

Blossfeld, Hans-Peter et al. (2007). "Event history analysis with Stata". In: *Lawrence Erlbaum*.

Blossfeld, Hans-Peter et al. (1986). *Ereignisanalyse: statistische Theorie und Anwendung in den Wirtschafts-und Sozialwissenschaften*. Vol. 569. Campus.

Bundesamt, Statistisches (2013). *Leading causes of death in 2013*. URL: `https://www.destatis.de/EN/FactsFigures/SocietyState/Health/CausesDeath/CausesDeath.html`.

Bundesamt, Statistisches (2015). *Durchschnittliche weitere Lebenserwartung nach Altersstufen*. URL: `https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Bevoelkerung/Sterbefaelle/Tabellen/Lebenserwartung.pdf?__blob=publicationFile`.

Cleves, Mario A. et al. (2002). *An Introduction to Survival Analysis*. Stata Corporation.

Cox, David. R (1972). "Regression models and life tables". In: *Journal of Royal Statistical Society* 34, pp. 187–220.

Diekmann, Andreas and Peter Mitter (1984). "Methoden zur Analyse von Zeitverläufen". In: *Stuttgart: Teubner*.

Engel, Uwe and Jost Reinecke (1994). *Panelanalyse: Grundlagen, Techniken, Beispiele*. Walter de Gruyter.

Graunt, John (1662). *Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality*. John Martin and James Allestry.

Hosmer, David W. Jr. and Stanley Lemeshow (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. John Wiley & Sons, Inc.

Kalbfleisch, John D. and Ross L. Prentice (1980). *The Statistical Analysis of Failure Time Data*. John Wiley.

Kleinbaum, David G. and Mitchel Klein (1996). *Survival analysis*. Springer.

Luy, Marc (2002). "Warum Frauen länger leben". In: *Erkenntnisse aus einem Vergleich von Kloster-und Allgemeinbevölkerung. Materialien zur Bevölkerungswissenschaft* 106.

Miller-Keane, Encyclopedia (2003). *Mortality*. Dictionary of Medicine, Nursing, and Allied Health. URL: http://medical-dictionary.thefreedictionary.com/mortality.

Mundi, Index (2015). *Life expectancy at birth (years)*. Index Mundi. URL: http://www.indexmundi.com/map/?t=0&v=30&r=xx&l=en.

Rößger, Felix (2015). *Allgemeine Sterbetafel: Methodische Erläuterungen und Ergebnisse*. Statistisches Bundesamt.

Schneider, Hilmar (1991). *Verweildaueranalyse mit GAUSS*. Campus Verlag.

Scott, Cameron (2013). *Life Expectancy Gains Are Slowing, Especially in the U.S.* Singularity University. URL: http://singularityhub.com/2013/09/26/life-expectancy-gains-are-slowing-especially-in-the-u-s/.

Singer, Judith D. and John B. Willett (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.

Wirtschaftsforschung, Deutsches Institut für (2015). URL: http://www.diw.de/de/diw_01.c.100293.de/ueber_uns/ueber_uns.html.

Declaration of Authorship

I hereby certify that the thesis I am submitting is entirely my own original work except where otherwise indicated. I am aware of the university's regulations concerning plagiarism, including those regulations concerning disciplinary actions that may result from plagiarism. Any use of the works of any other author, in any form, is properly acknowledged at their point of use.

Student's signature:
Name: Sarah Bekele
Date: 4^{th} of May 2015