Frailty Models in the Analysis of Socioeconomic Differences in Mortality

Application to mortality by education level in Turin

Virginia Zarulli

Dottorato in Demografia XXIV ciclo

Sapienza Università di Roma, Dipartimento di Scienze Statistiche

Contents

List of Figures

List of Tables

Acknowledgements

At the end of my PhD studies I look back and realize how long is the list of people I would like to thank for supporting me over the last couple of years of my life. Like any PhD student knows, this adventure has been both happy and hard at the same time and has brought me through ups and downs, some of which I could have never overcome without your help.

I am truly thankful to my supervisors, Graziella Caselli and Jutta Gampe. They have provided the necessary help to shape my research and analytical skills, stimulating and encouraging me to develop a scientific approach to problems. More specifically, I would like to express my gratitude to Graziella Caselli for thematizing my research interests and to Jutta Gampe for guiding me through the dangerous waters of statistics.

This dissertation would not have been possible without Giuseppe Costa and the wonderful research group of the Public Health School ASL TO3 in Turin. I would like to thank them for believing in the project and including it in their workplan, in order to use their rich dataset for the purposes of this research. A special thank you goes to Chiara Marinacci for lending her assistance on data related questions and for replying to many emails, too many to be counted.

I wish to thank Jim Vaupel for inviting me to the Max Planck Institute for Demographic Research. This is one of the most stimulating place I have ever been, where workshops, seminars, working groups and similar events never end and provide the occasion for invigorating discussions. Not to mention how friendly this environment is. Relaxing lunch and tea breaks

and necessary distractions in the company of a group of invaluable friends and colleagues make research happier and better. So thanks to Alyson van Raalte, Maren Rebke, Ester Gonzalez Prieto, Miguel Sánchez Romero, Yann Le Cunff, Svenja Weise, Saskia Hin and many others.

I am indebted to the Statistics Department of Sapienza University of Rome for many things. In particular, I would like to express my gratitude to them for giving me the possibility to attend the EDSD courses in Paris, one of the most formative experiences a first year PhD student could wish for. I'm also really grateful for giving the freedom to pursue my research interests by visiting other research institutions and groups and for providing economic support for most of my PhD studies.

I owe a lot to Giancarlo Camarda, for always finding the time to answer tricky statistical questions even when very busy and always with a smile.

I really appreciated the help of the IT group at Max Planck for providing very powerful computing facilities, of Stefania Borghi and Gunde Paetrow for their precious help with bureaucracy, of Iris Levitis for her supermegafast copy-editing and of Tigerlily Levitis for going to bed early so that the editing could be done.

I cannot find the right words to thank Adam Lenart for what he has been to me during my PhD. I could start with mentioning his assistance with R programming. I could also mention his always being available for any kind of question. But most of all, I feel an immense gratitude for his standing by me day after day, even when things were not going well and my mood was terrible (poor him).

To conclude, this dissertation is dedicated to my parents, Fiorella and Nicola (and also to my little dog, Leo). When I was younger and still lived with them I could never have realized how much I was going to miss them now that I live far away. I would like to thank them for everything they have done for me, for having supported me in every situation and for their being proud of me, while all I would like to say now is that I am proud of them, as parents and as people.

Chapter 1

Introduction

The notion of unobserved heterogeneity of frailty, from which the literature about frailty models has sprouted, was introduced in demography by Vaupel et al. (1979). The authors pointed out the fact that all populations have a higher degree of heterogeneity than the one explained by observed covariates, because individuals differ one from another for many other unobserved characteristics called *frailty*. Even when the investigator has at his disposal a very rich data set with individual level information (a situation that seldom occurs), there will always be a component of heterogeneity that is unobserved. Frailty, in fact, is a general concept that does not distinguish between acquired weakness, lifestyle factors, environmental risks and innate biological frailty. It combines in a single measure all the factors that operate to increase or decrease a given individual's mortality risk, regardless of the source of heterogeneity.

Besides affecting the individual's mortality risk, the presence of unobserved heterogeneity of frailty influences mortality at the population level, contributing to shaping the observed dynamics. A population, in fact, is a mixture of several individuals with unobserved features. Some individuals are frailer and some are more robust. As the cohort ages, frailer individuals exit the population earlier than more robust ones, gradually selecting the surviving population towards robustness. This implies that the observed dy-

namic at the population level does not necessarily correspond to the dynamic experienced at the individual level. For example, at old ages the observed population force of mortality decelerates and eventually, levels off. This phenomenon is known as mortality plateau or mortality deceleration. However, this is seen as an artifact of selection and does not mean that the individuals, as they age, see their probability of death increasing less quickly. To address the problem of unobserved heterogeneity and to improve the fit of mortality models to populations, especially at very old ages, frailty models have been suggested. Frailty models are random effect models where the random term represents the unobserved frailty component. A detailed illustration of the theoretical framework and its formalization is given in chapter 3.

This gradual selection and compositional change caused by unobserved heterogeneity represents an explanation for the phenomenon of decreasing relative mortality differences at old ages by socioeconomic status. In the analysis of differential mortality by socioeconomic group, in fact, the pattern of converging hazards at old ages is often observed. Selective effects of high early mortality experienced by the disadvantaged group would leave more robust individuals at old ages, causing convergence with the hazard of the lower mortality group that is subjected to weaker selection.

The convergence, therefore, would be an artifact of selection at the population level and would not necessarily mean that the single individuals experience a reduction of the advantage/disadvantage as they age. This is the perspective adopted, by contrast, by the age-as-leveler hypothesis, which regards the weakening association between socioeconomic status and mortality at old ages as a result of factors like governmental support for the elderly, biological frailty and disengagement from systems that contribute to social stratification, like the labor market. These factors would contribute to level off differences between individuals, that, consequently, become equally susceptible as they age. A more detailed discussion about the opposite frameworks explaining the phenomenon of convergence can be found in paragraph 2.5 in chapter 2.

Unobserved heterogeneity has another important implication. In duration dependence models, its presence poses problems for estimating the hazard function and the effect of observed covariates on the survival chances. Because of unobserved heterogeneity, in fact, the population undergoes a constant compositional change that, if neglected, can lead to biased estimates of the hazard rate and of the regression coefficients for the covariates. It is important to notice that this happens even if this omitted source of variation is uncorrelated with the predictors included in the model, because, at every age, the death rate is computed based on a population at risk whose composition is constantly changing. The aspect of the bias is treated in paragraph 3.3, chapter 3.

In the last decade, the interest in the connection between unobserved heterogeneity and socioeconomic differences in health and mortality has increased. Most of the debate is focused on whether selection processes play a role in the life course pattern of socioeconomic differentials, although the number of studies directly addressing this question is still limited. Moreover, most of these studies are rather focused on health outcomes, while few concentrate specifically on mortality. There is a wide variety of strategies used to address the issue of selection. The primary one is to keep in the sample those who die during the observation period, imputing them a value for their health status with some method. This diversity of methods poses problems because each of them can potentially affect the results. This is reflected in the literature, the findings of which are not consistent and fuel a still controversial debate (a more extensive review can be found in paragraph 3.4, chapter 3).

While studies aiming to assess if selection explains the converging health trajectories exist, there is, however, a lack of studies trying to evaluate the impact of unobserved heterogeneity on the estimation of coefficients for the mortality risk of the socioeconomic groups in survival models. This is quite surprising given the strong evidence showing that, ignoring unobserved characteristics causes the estimates to be biased towards zero, thus leading to an

underestimation of the relative risks between groups.

In this thesis I investigated the presence of selection processes in the mortality patterns, from age 50 on, by educational level. The choice of education as a proxy for socioeconomic condition is consistent with the literature on the topic. Education is known to be associated with several health outcomes, it remains stable over the adult life time (not being affected by the fluctuations that may happen in the working life), it is applicable to individuals that are not active in the labor market and is equally valid for men and women. A brief review of the characteristics of the main indicators for socioeconomic status can be found in paragraph 2.3, chapter 2.

This study analyzed the socioeconomic gradient in the mortality pattern over age from a longitudinal perspective. Its aims were 1) to investigate whether heterogeneity matters by analyzing if and how the estimates of the mortality differentials are affected by the introduction of the unobserved heterogeneity component into the models and 2) to investigate whether the theoretical framework of the frailty models can explain the observed pattern better than the standard survival analysis approach that does not control for unobserved heterogeneity. This would strengthen the validity of the selection hypothesis as an alternative explanation for the reduction of the gradient in socioeconomic mortality at old ages.

Thanks to the collaboration with the Public Health School ASL TO3 (Turin), this study used individual longitudinal histories of a 36 year mortality follow up (from 1971 to 2007) from the Turin Longitudinal Study.

The Turin Longitudinal Study (TLS) is an Italian study that links data from the last four censuses (1971, 1981, 1991, 2001), vital statistics registry and health system archives for the city of Turin. The system of integrated data sources permits the creation, via record linkage techniques, of longitudinal individual histories over a follow up period of almost 40 years. The study includes the entire population of Turin, an industrial city in the North-West of Italy of nearly one million inhabitants. Information at the individual level and at the contextual level can be connected, thus creating a very rich data

base for several analytical purposes (more information about the data are given in chapter 4).

Data were analyzed with survival analysis techniques, following the standard approach without unobserved heterogeneity and the frailty models approach, which includes the concept of frailty in the analytical methodology of duration dependence models (methods are illustrated in chapter 5).

To my knowledge this is the first Italian study that addressed the question of selection process and unobserved heterogeneity of frailty in the estimation of mortality differentials by socioeconomic status.

1.1 Outline

Chapter 2 presents a review of the literature about socioeconomic differences in mortality. The most important theories and explanations for the relation between socioeconomic status and health/mortality outcomes are illustrated, along with the main indicators used in the literature as proxy for socioeconomic status. The main studies and results about this phenomenon in Italy are also summarized. The chapter pays special attention to the change over age of socioeconomic differences in mortality, mentioning the selection hypothesis among the possible explanations, thus bridging the discussion with the next chapter.

Chapter 3 introduces the concept of unobserved frailty, the theoretical framework and the basic mathematics of the frailty models. The main effects of unobserved frailty, mortality deceleration and convergence of hazards by subgroups at old ages, are discussed and linked to the observed decreasing differences in mortality by socioeconomic group. The chapter also discusses the bias caused by neglecting this component in survival analysis models. Moreover, a review of differential mortality studies that have addressed the selection hypothesis is given.

Chapter 4 describes the Turin Longitudinal Study, the data source used for the analysis. A description of the available variables and of the study

design is provided, along with descriptive statistics and results.

Chapter 5 is about the methods. The presentation of the estimation strategy is preceded by a brief illustration of the main concept and functions in survival analysis. The likelihood functions used for the models with and without frailty are given, followed by the description of the different models implemented: exponential model , age-dependent hazard models with Gompertz and Makeham baselines, and the two approaches used to model survival improvement (cohort improvement and period improvement).

Chapter 6 reports and discusses the results of the analysis, trying to connect all the pieces to create a unique picture. It also summarizes the most important findings and draws conclusions.

The thesis also includes an Appendix where the R codes used for the analysis, programmed by the author, are given. The appendix also contains additional tables and figures.

Chapter 2

SES and mortality

2.1 Socioeconomic differences in mortality

The close relation between socioeconomic status, health and mortality conditions has been known and analyzed for a long time. Antonovsky (1967) presented a collection of studies that ranged from the early nineteenth century to the years after the Second World War, and concluded that, whatever the index used and whatever the number of classes considered, the lower classes showed higher mortality rates.

However, a collective awareness of the existence of these differentials did not exist until the late nineteenth century. It was only several decades after the Industrial Revolution had spread that, thanks to the work of people like Louis René Villermé and Rudolf Virchow, who analyzed inequality in mortality between rich districts and poor districts, that the health and mortality disadvantage of the lower social classes had become clearly visible (Mackenbach et al. 2006). As a consequence, the first public interventions took place and compulsory sickness and disability insurances were created (Vallin 1980).

After the Second World War, the creation of universal welfare states explicitly aimed to eliminate inequalities with respect to health and mortality, led to the firm belief that such inequalities would fade and, eventually, vanish.

However, Vallin (1980) pointed out the fact that recent data and studies tended to indicate that the expectation of narrowing inequality was a false belief. In the early 1980 the Black Report in England demonstrated that, although overall health had improved, widespread health inequalities were still strong and that their main cause was economic inequality (Black et al. 1980).

Since then, many studies have been conducted and socioeconomic health and mortality inequalities have been extensively analyzed in several countries. The growing body of research emerging has confirmed the existence of a substantial health and mortality gradient between social groups, independent of the indicator used as proxy of the socioeconomic status: income (Martikainen et al. 2009), occupational status (Martikainen et al. 2001), education (Elo and Preston 1996), housing tenure or a combination of these (Huisman et al. 2004; Elo et al. 2006; Koskinen et al. 2007).

Although inequalities were found in all the countries that have been examined, international comparisons have uncovered regional variations. For example, they are small in some southern European countries and larger in the Eastern and Baltic regions (Mackenbach et al. 2008). Moreover, a North-South gradient is observed and inequalities were found to be smaller in the Southern European countries (Mackenbach et al. 1997; Kunst et al. 1998b,a).

These variations challenge the widely held views that inequalities in health are smaller in countries with more egalitarian policies, such as Sweden and Norway. Hoffmann (2011b) analyzed the impact of socioeconomic status on mortality and the magnitude of the difference in Denmark and the USA. He found that the differentials were larger in Denmark than in the USA, two countries with completely different welfare systems: social democratic and liberal. Previous studies found that the introduction of egalitarian policies did not prevent inequality to widen, like with the implementation of the National Health Service in Great Britain in 1946 (Pamuk 1985) and the Medicare program in the USA in 1965 (Preston and Elo 1995).

Despite the overall improvement in survival, which has lowered the mor-

tality rates of both upper and lower classes, many studies have reported that the mortality gap between them has been widening in the last decades because mortality rates declined at a faster rate for the upper classes. Upper classes, in fact, performed better in those causes of death that are responsible for bigger improvements in survival, like cardiovascular diseases (Phillimore et al. 1994; Mackenbach et al. 2003, 2006).

It is likely that several factors contribute systematically to the growth of the gap, at adult and at old ages as well. Behavioral and life style factors play an important role, especially considering that unhealthy behaviors tend to accumulate, thus strengthening their negative effect (Laaksonen et al. 2001). Non smoking, moderate alcohol usage, physical activity and good dietary habits are related to a lower risk of mortality (Knoops et al. 2004), all factors where lower social classes are likely to perform worse than other groups of the population (House et al. 1990). Moreover, the upper classes may benefit more from the introduction of new treatments and might have more ability to take advantage of health prevention messages (Martikainen et al. 2001).

2.2 Explanations for socioeconomic mortality

The explanations for the persisting association between socioeconomic status and health outcomes are diverse. Generally they emphasize the causal direction from socioeconomic to health inequality, recalling behavioral factors, material factors and psychosocial factors. Among the first explanations, the most important role seems to be played by cigarette smoking (Mackenbach et al. 2004; Blakely and Wilson 2005; Stringhini et al. 2010) and dietary and exercise habits (Smith et al. 1997; **?**; Stringhini et al. 2010). Material factors include, among others, occupational health risks (Schrijvers et al. 1998; Rahkonen et al. 2006), access to medical care (Mackenbach et al. 1989; Van Doorslaer et al. 2006) and housing conditions (Krieger and Higgins 2002).

Psychosocial factors, instead, relate to stressful situations like feelings of deprivation, financial problems and employment insecurity (Wilkinson 1997; Aldabe et al. 2011).

2.2.1 Proximate determinants

A systematization of the factors influencing mortality is represented by the list of five categories compiled by Hoffmann (2008):

- Genetic constitution
- Natural/physical environment
- Structural and material conditions
- Behavioral and cultural factors
- Psychosocial circumstances

Genetic constitution

When it comes to mortality, *genetics* matters. Research conducted so far suggests that genetics may explain around 25% of the variability in adult lifespans (Christensen and Vaupel 1996; Vaupel et al. 1998). Morevoer, it has been found that life expectancies of monozygotic twins have higher correlation than life expectancies of dizygotic twins (McGue et al. 1993; Herskind et al. 1996). The question of whether genetics might play a role in the mortality differential by social status, in contrast, is more controversial. Obviously, any racist argument blaming genetic inferiority for the disadvantaged position of some individuals in a society, must be firmly rejected. It is different, however, to say that an interplay between genetic and social factors can take place (Hoffmann 2008). An example is sex, a genetically determined attribute, which translates into gender, a more socially influenced attribute. It is known that women experience lower mortality than men and it has been found that socioeconomic differences in mortality tend to be smaller among

females than among males (Koskinen and Martelin 1994; Elo et al. 2006). This is partly due to behaviors and social roles that are attributed by the social environment, like, for example, cigarette smoking: smoking prevalence is usually higher in lower socioeconomic groups, particularly among men, because women tended to smoke less. As a consequence, the socioeconomic gradient for lung cancer is much more pronounced among men [1](Lopez et al. 1994).

Environment

Environment and its natural and physical characteristics influence mortality. They can also affect the socioeconomic gradient in mortality. Living in an insalubrious area can accentuate the gradient because lower social classes might have less means to protect themselves from it. On the other hand, it can also be the case that living in a healthy climate or in an area where healthy food is abundant and easily accessible can partly compensate for the socioeconomic disadvantage (Hoffmann 2008).

Structural and material conditions

Structural, material conditions and their interactions are also very important determinants of mortality. Elements like income, type of occupation, housing conditions and so on are material factors. Examples of structural factors are health care and education systems. Also education itself is a structural factor because it is very likely to determine the social position of the individual. Education influences mortality in a variety of ways. It determines occupation and, therefore, the exposure to risky job environments. It determines income and, therefore financial wealth, which, in turn, influences housing conditions, access to higher quality products and services and to appropriate medical care. Education creates the conditions for a better understanding of complex

[1]This was true for older cohorts, while more recently, the so called "smoking epidemic" is spreading and more and more women are adopting this unhealthy habit and in the future, the socioeconomic patterns of lung cancer among women are expected to change

systems and processing complex information, which are very important for acquiring and receiving health knowledge. It develops the ability to self-direct toward desired goals or values, including health (Mirowsky and Ross 2003).

Behavioral and cultural factors

Closely related to the material conditions are the *behavioral and cultural factors*. The close connection is clearly visible in aspects like dietary habits or physical activity. Upper classes can afford more healthy (and expensive) food or to spend more money for sports and outdoor activities in their free time. But dietary habits are typically behavioral as is the attitude towards sports. However, other elements can act upon this sphere, like whether one's occupation is physically demanding or not. Both of these factors can lead to obesity, which is, together with smoking, one of the main risk factor for cardiovascular mortality, thus contributing to the social pattern of cardiovascular diseases (Phillimore et al. 1994; Mackenbach et al. 2003, 2006).

Psychosocial circumstances

Psychosocial circumstances may produce stress that has effects on health and, ultimately, on mortality. "Stress is the nonspecific response of the body to any demand, whether is caused by, or results in, pleasant or unpleasant conditions." (Selye 1985) The psychosocial stress refers to the stress due to acute or chronic events of psychological or social origin (Cockerham 2007). The literature has identified three major types of social stressors: life events (accumulation of several events in a person's life within a short period), chronic strains (persistent demands over a prolonged period) and daily hassles (short term events requiring small behavioral adjustments) (Pearlin 1989; Thoits 1995). Lower social classes have been found to experience more negative life events and chronic strains (Pearlin et al. 2005), to tend to be more stressed by their living environment (Browning and Cagney 2003), to have less personal resources or knowledge to cope with stress and to have more

frequent feelings of fatalism and, therefore, of no control over life (Mirowsky and Ross 2003), to have lower ability to physically respond to the stressors (lower saliva cortisol response to stress) (Kristenson et al. 2001).

2.2.2 Different levels of determinants

The aforementioned proximate determinants might not be the only causes of mortality and health inequality. Other determinants could be found at different levels of the society.

The Wilkinson hypothesis

Wilkinson (1992, 1996) suggests the hypothesis that high income inequality in a society causes higher mortality. Relative levels of income, rather than the absolute level of wealth, have an impact on health and mortality. Income disparity is an indicator for lack of and disinvestment in social capital. A highly unequal society generates feelings of relative deprivation, poor social support and hopelessness that undermine social cohesion and their psychosocial effects (stress) have health consequences, especially for the disadvantaged classes.

General susceptibility theory

An alternative approach that suggests that there might be something more than the proximate determinants, is the general susceptibility theory. It states that certain social groups always share a higher risk of death, whatever causes of death are operating. This relationship, according to this point of view, can not be explained by known risk factors and suggests the fact that there are some common factors that increase the susceptibility to disease of lower social classes. In the Withehall Study, one of the most influential reports on socioeconomic status and mortality, only one quarter of coronary heart diseases mortality gradient is explained by proximate determinants and the similarity of the risk gradient from a range of specific diseases seems to

indicate the operation of factors affecting general susceptibility (Rose and Marmot 1981; Marmot et al. 1984).

2.2.3 Life course approach

Health status in old ages depends also on previous circumstances experienced in the past. The life course approach to socioeconomic inequalities in mortality highlights the importance of the possible links between death, the final event, and previous events and circumstances experienced over the life course. Specific patterns of life-course exposure are related to specific diseases: coronary heart diseases, chronic obstructive respiratory disease, breast cancer, and suicide are mostly determined by exposures acting right across life, while stroke and stomach cancer are related to early life influences (Davey Smith et al. 2001; Galobardes et al. 2004).

Childhood conditions are considered particularly important, especially in the light of recent results that have demonstrated how they affect adult and old ages health outcomes. Early exposure to malnutrition, for example, has been found to have negative effects on blood pressure, obesity and other cardiovascular diseases related factors (Stanner et al. 1997; Painter et al. 2005; Lumey et al. 2007; Sparén et al. 2004). To the extent which the exposure to detrimental conditions early in life is socioeconomically patterned, its role in contributing to the health and mortality gradient clearly emerges. Children from lower class families, in fact, are more likely to have poor diets, to be of low birth weight, to be more exposed to passive smoking and some infectious agents and to have fewer educational opportunities (Ben-Shlomo and Kuh 2002). The importance of childhood conditions relates, indeed, also to factors that are crucial for the process of transmission of health knowledge and behaviors, like parental education and socioeconomic status and family living arrangements. Although the evidence for their influence is fragmentary, it suggests they may affect mortality later in life (Hayward and Gorman 2004) through varied possible pathways (Pensola and Valkonen 2002; Galobardes et al. 2004).

2.2.4 Reverse causation

Besides the approaches that focus on the causal direction from socioeconomic status to health, a completely opposite perspective exists, which emphasizes the reverse causality from health status to socioeconomic position. This approach, called health selection or reverse causation, states that health determines social status through several pathways. The general underlying mechanism hypothesized by this approach is that a decline in health causes higher expenses and lower income (due to limited ability to work), thus determining wealth and social status decline. This explanatory framework is mainly adopted in the economic research, where several findings support it (Ben-Shlomo and Kuh 2002; Smith 1999), while the majority of the epidemiological literature agrees with the idea that reverse causation is not of a major importance in the explanation of the existence of socioeconomic inequality in health and mortality (Goldman 2001; Cardano et al. 2004; Hoffmann 2008). However, as Goldman (2001) points out, most likely both directions are likely to take place in the causation process.

2.3 Indicators of socioeconomic status

The concept of socioeconomic status is intuitively multidimensional. It contains social, cultural and economic aspects that are related to each other in a complex system of interactions. This places its conceptualization and measuring among the more difficult and controversial subjects in social research. The main reason seems to be a "lack of clarity about the essential nature of social stratification. Although the concepts of social structure, social class and socioeconomic status are central to the social sciences, theorists have not agreed on definitions' (Oakes and Rossi 2003: 3).

This is reflected in the empirical analysis, where there is no unique practical operationalization of socioeconomic status and several indicators are used as proxy for it. The most commonly used indicators for studies of the relation between mortality and socioeconomic position are income, occupation and

education (Oakes and Rossi 2003; Hoffmann 2008). Geyer et al. (2006) note that these three are often used interchangeably and reference is often made to one of these dimensions to support research into one of the others, while this should be avoided because, although correlated, they measure different phenomena and represent different dimensions of socioeconomic status.

An indicator that is adequate and applicable to all studies and all settings does not exist, and each indicator can be more or less important for certain health outcomes or specific stages of life. Galobardes et al. (2006: 1) point out how the choice of the measure "should be informed by consideration of the specific research question and the proposed mechanisms linking socioeconomic position to the outcome. [...] However, in practice, the measures used tend to be driven by what is available or has been previously collected." Therefore, even when it is not possible to choose which measure to use, it is important for the researcher to have a clear understanding of the theoretical ground of the indicators, their advantages and limitations. The next paragraphs provide a description of the three main indicators, income, occupation and education, with a particular focus on the last one, which is the indicator used in this study.

2.3.1 Income

Income and economic indicators are less frequently used than the occupation and education indicators. They are very concrete measures that refer to the most material dimensions of socioeconomic class. However, income is not an easy measure to use and several factors can explain, at least partly, the underutilization of this measure.

Income includes several components, like wage earnings, child support, transfer payments and many others that are very difficult to measure. Income can fluctuate widely over time and, therefore, can be difficult to track. Most health inequalities research, in fact, uses income based indicator measured at a single point in time, while income volatility (and even more persistent low income) has negative repercussion on health and mortality (McDonough

et al. 1997).

Household income is the most typical income-based indicator. Often the information about the size of the household is not available, making it impossible to adjust the income level according to the number of members in the family in order to have a non-biased measure of the real available income for the individual (Krieger et al. 1997). Moreover, even when the size of the household is available, different members may be in a disadvantaged position when it comes to resource sharing and have unequal access to the household income, as in the case of women (Pahl 1990; Vogler and Pahl 1994).

Duncan et al. (2002) found that income and economic indicators were more sensitive to mortality risk than education and occupation and recommended them as a standard measure in health studies. However, the usage of this measure should be carefully considered according to the situation. For example, income as indicator of socioeconomic status at old ages can be problematic, since the elderly can frequently have low incomes, but substantial wealth.

2.3.2 Occupation

For obvious reasons, occupation is mainly used for studies about young and adult ages. However, from a life course perspective this indicator can also be an important predictor at old ages, being related to exposure to occupational risk, working conditions and environment that may have long lasting effects or consequences in the long term (Karasek and Theorell 1994; Johnson et al. 1996).

Like income, occupation can alsp change over time. Commonly, most recent or usual occupation is used as an indicator, but in very open and mobile labor markets this can be too limiting and a more comprehensive definition might be more appropriate. A definition that, for example, is able to contain and summarize the information about the individual occupation trajectory.

The classification of occupation is very complex and can differ greatly

from one country to another. Moreover, it is subject to periodic modifications. These factors hinder international comparative studies and long term analyses.

The major limitation of this measure, however, is the fact that it necessarily leaves unclassified all those categories that are less involved in the labor market or don't have one occupation. This is the case, for example, of women belonging to older cohorts, whose participation to the labor market was very limited and are usually classified by their husband's occupation (Ross and Wu 1996).

Finally, as (Duncan et al. 2002) point out, occupational status at older ages can be affected by reverse causation, since poor health can determine a decline in occupation status.

2.3.3 Education

Education measures something different from what income and occupation do, although it is closely related to them. Education marks the social status at the beginning of adulthood and functions as a bridge between the social status of two successive generations, being the main road to upward mobility. It precedes chronologically and causally occupation and income, thus affecting the potential access to material resources, freedom from financial stress and individual and household wealth (Mirowsky and Ross 2003).

Education is especially important to health and mortality because it trains the individual to acquire, evaluate and use information, to develop effectiveness and self direction toward a desired objective and provides better knowledge. All these factors have a positive impact on health knowledge and behaviors (Mirowsky and Ross 2003; Hoffmann 2008; Doblhammer et al. 2009). It is well known in the literature that education is associated with several health outcomes: people with lower level of education have a higher probability to become disabled, face a general higher risk of mortality, smoke more cigarettes and exercise less than individuals with higher educational levels. Moreover, higher education is often associated with higher levels of

self-control, efficacy and happiness (Doblhammer et al. 2009).

There are additional reasons that contribute to the popularity of this indicator. It is objective and easy to measure, unlike the financial measures it is a time constant variable that stays stable over the adult life time and is not affected by the fluctuations that may happen in working life. It is applicable to individuals that are not active in the labor market or do not have a normal income for some periods of their life and it is equally valid for both sexes (Krieger et al. 1997; Hoffmann 2008). Finally, education is regardless of changes in health status that happen at adult or old ages, thus not being subject to reverse causation because most people reach final education attainment in early adulthood.

However, education can be subject to reverse causation when considering a long life course perspective, since poor health in childhood could lower educational attainments. Moreover, being stable over the adult life, education is not able to capture either differential job training and career investments trajectories made by individuals with similar levels of formal education or income and economic volatility that have been showed to influence health (Duncan et al. 2002).

Another limitation of this measure is the fact that the meaning of educational attainment changes for different cohorts and the same numbers of schooling years cannot be compared between older cohorts, that have on average fewer number of schooling years, and younger cohorts. Studies including individuals from several birth cohorts have to consider this aspect, using a classification of education that accounts for the cohort specific relevance of the educational categories.

The literature has also identified other limitations. Using education can pose problems when the individuals have obtained their degrees in a different country from the one of residence because the two nations can have distinct systems of classifications. Finally, in general, no information about the quality of the educational experience is available, while this is likely to be important in the way that education influences health outcomes (Galobardes

et al. 2006).

2.4 Socioeconomic mortality in Italy

The first report on socioeconomic health inequalities in Italy was published by the National Institute of Statistics in 1990. It was based on the record linkage between the data from the 1981 census with the death certificates of 1981-1982 (Istat 1990). Ten years after the study was repeated with the data from the 1991 census and the death certificates of 1991-1992 (Istat 2001). This allowed researchers to monitor the trends and changes over ten years. Mamo et al. (2005a) analyzed the changes in mortality by cause and occupational category between the first and the second study and found that the cause of death specific disadvantage suffered by several occupational categories had not decreased over 10 years. Between the first and the second report, unexpected changes in the educational gradient in mortality from all causes were uncovered. In particular, the men gradient by education between ages 18-74 in the period 1991-1992 had slightly widened and become similar to that of the women gradient in 1981-1982. Surprisingly, the women gradient had substantially narrowed, in particular the advantages of a university degree over a high school diploma had disappeared.

Luy et al. (2011) hypothesize that the narrowing female gradient might be due to cigarette smoking. Smoking prevalence in Italy in the 80's and in the 90's, indeed, was inversely related to education among men but directly related to education among women (Faggiano et al. 2001) and this could explain the results of the report concerning the education mortality gradient in $1991 - 1992$. Moreover, analyses of smoking prevalence trends over time show that women of a higher educational level smoked more than those of a lower level but that the gap tended to decrease, as the less educated women started to smoke. At the beginning of the 1990s, a positive association between smoking and educational level was reported among women aged 45-74, whereas the association was reverse among younger women (Cavelaars et al.

2000), suggesting that the pattern of social differences in mortality might change in the future.

Comparisons with other nation wide population based datasets for similar periods revealed that Italy (and, more broadly the Mediterranean area) showed a lower overall level of social class inequality in both morbidity and mortality than that found in the continental and the northern countries. (Mackenbach et al. 1997). Concerning cause-specific inequality patterns, Italy presented almost no social gradient in cardiovascular mortality, the group of causes that, at the end of the 20th and in the beginning of the 21th century, is responsible for the largest share of inequality in mortality. This could explain why the overall gradient was lower than other countries, which are characterized by pronounced cardiovascular mortality inequalities. On the other hand, larger neoplasms inequalities, with the exception of lung cancer, were found in Italy compared to other countries (Kunst et al. 1998a).

Unfortunately, from the census 2001, the National Institute of Statistics stopped linking the census data with the death registration and this valuable source of information for the whole population is no longer available. Therefore, more recent investigations must find other ways to estimates mortality by social class.

Some studies have used pension and retirement registers for the the estimation of mortality by productive sector but they point out the limits of these data: they lack information about complete job histories and about jobs held (d' Errico et al. 2005). A socioeconomic status gradient in life expectancy at retirement age has been found by analyzing working histories data. In the light of the pension system, these differences cause an implicit redistribution of resources from lower social status and low earning jobs to upper social status and higher earning jobs. The mechanism for computing the pension benefits, in fact, does not take into account the different number of years that individuals who come from different working histories still have to live when they reach the retirement age (Leombruni et al. 2008). However, these studies are not generalizable to the whole population because they re-

fer only to the population that is active and officially employed in the labor market or, as in the case of the *Working Histories Panel*, only to part of the employed population (civil servants or self-employed individuals are not included in this panel).

A study conducted on data from the *Survey of Health, Aging and Retirement in Europe* has confirmed the advantage of high class individuals for several health indicators and has proved education to be a more important correlate than wealth and income among the elderly (Tsimbos 2010). The educational gap by occupation and education level and its evolution between 1980 and 1994 has been estimated using data from the multipurpose survey *Famiglia, soggetti sociali e condizione dell'infanzia (Family, welfare institutions and childhood conditions)*, waves 1998 and 2003 (Luy et al. 2011). The authors have developed a modified version of the orphanhood method to get indirect estimates of adult mortality from information about parent's survival and found that, mortality differences between the highest and the other socio-economic groups have increased among men but have decreased among women. This confirms the trend that was found in the second report published by the National Institute of Statistics (Istat 2001).

Other investigators, adopting an ecological perspective, have attempted to correlate death rates and social, environmental, economic and other indicators. Caselli and Egidi (1981) have linked the geography of various causes of death of each Italian province with a wide range of factors related to the socioeconomic and environmental situation and to the way of life of the sub-population. They found that degenerative diseases were typical of the industrialized, richer but more polluted areas, while poorer areas, which lack of efficient and adequate health structures, were mainly struck by diseases which medical progress have made treatable. The provincial level has also been used to test the hypothesis that income inequality is related to higher mortality and the results showed that total mortality and income inequality (as indicated by the Gini coefficient) in Italy are positively associated only in provinces with lower income (Materia et al. 2005). Further evidence for an

inverse association between socioeconomic level of an area and level of total mortality and mortality from the major causes of death, among men, were found in of Rome, where the census tracts of the city were examined for the period 1990-1995 (Michelozzi et al. 1999).

Other analyses are based on census-linked data available for local populations only, as in the case of the Tuscany Longitudinal Study (Merler et al. 1999) and the Turin Longitudinal Study (TLS). These are two studies that link the census data for the cities of Florence and Leghorn in the central region of Tuscany, and for the city of Turin in the North-Western region of Piedmont.

The empirical analysis of this thesis is based on TLS data. Therefore, it is worthwhile to describe this study, and to present a summary of the main results obtained from its data. For more detailed information about the data structure please see chapter 4.

Thanks to the richness and the high quality of its individual level data, the Turin Longitudinal Study is used for many analyses in epidemiology and public health, in both Italian and international studies. Since the record linkage at the national level was interrupted after the 1991 census, in almost all the most recent international comparative studies on social inequalities and health, the evidence for Italy is based on the Turin's data (Mackenbach et al. 1999; Huisman et al. 2004; Mackenbach et al. 2006).

Compared to other parts of the continent, health and mortality inequality is generally lower in Southern Europe (Mackenbach et al. 2008). This is almost certainly related to peculiar causes of death trends over time, like cardiovascular and lung cancer mortality. Mortality from cardiovascular diseases is the main driving factor for the widening of the gap between classes because it declines proportionally faster among the upper classes. But the trend in Turin at the end of the 90's was the opposite: the lower classes experienced a faster decline in mortality from these causes (Mackenbach et al. 2003). With respect to lung cancer mortality, Turin differs from what happens in other countries, like Belgium and the Northern ones, because, the

contribution of this cause of death to the overall level of mortality inequality is small, especially among women (Mackenbach et al. 2004).

Although detected socioeconomic inequalities are smaller than in other European countries, they still exist. Compared to other countries, inequality in cardiovascular mortality in Turin is small but this does not necessarily mean that there are no inequalities. Taking a look with a magnifying glass, a variegated pattern emerges, where the disparity is indeed small for ischemic heart diseases but more pronounced for cerebrovascular mortality (Costa et al. 1999). Moreover, employment status, rather than education or income, seems to be the most important factor explaining inequality in coronary heart diseases, though the disparity is bigger in mortality than in incidence, calling attention to possible differences in timely access to hospitalization and health care (Petrelli et al. 2006).

Indeed, socioeconomic barriers to opportunities for effective and better care seem to exist, even if Italy's health system offers universal coverage. Significant social class differences in cancer survival are found in both sexes, especially in sites for which effective treatments are available (Rosso et al. 1997), while the gradient in cancer incidence is less pronounced among women than among men (Spadea et al. 2009, 2010). This seems to be a consolidated pattern also for the overall mortality, where a clear social gradient among men has been documented, while the differences among women are less evident and tend to become negligible when controlling simultaneously for factors like job exposure, housing conditions and educational level (Mamo et al. 2005b). Individual level factors are not the only responsible factors for differences in mortality by social classes. Also contextual socioeconomic circumstances have an impact on mortality, both among men and women: significantly higher mortality risks have been detected among residents of deprived neighborhoods in Turin (Marinacci et al. 2004).

2.5 Socioeconomic mortality over age

Differences in mortality related to socioeconomic status are not stable at all ages. On the contrary, the differences in mortality vary with age.

2.5.1 The phenomenon of convergence

Several studies show that the relative difference between upper and lower class decreases at old ages. This was detected already by Antonovsky (1967) who analyzed about 30 studies ranging from very far back in time to more recent times and noticed that, with amazing regularity, the differentials were largest at middle adult ages and then gradually declined and tended to disappear at old ages.

More recently, evidence has been found in both morbidity (Huisman et al. 2003; Dalstra et al. 2006) and mortality (Martelin 1994; Huisman et al. 2004) in several European countries, although in the study of Huisman et al. (2004) this pattern was found not to be universal and the relative mortality differences by social status did not decline with age among women. Few studies have reported not decreasing socioeconomic mortality differences at old ages (Olausson 1991; Ross and Wu 1996; Lynch 2003), but the empirical evidence supporting this pattern remains limited.

It is important to specify that the age pattern that is discussed here refers to the relative differences in mortality and not to the absolute ones. Absolute differences, in fact, don't tend to diminish with age. As individuals age, the overall level of mortality grows so steeply for all social groups that the absolute difference, in many cases, will increase (Martelin 1994).

Figure 2.1 reports an illustrative examples with Italian data from the city of Turin. The figure shows the mortality hazard from age 50 on by 4 education groups. The follow up is the result of a record linkage between the 1971 census population of Turin and the registration of deahts, in Turin, until July 2007. The data are displayed on logarithmic scale as it helps to visualize the converging pattern of the educational hazards that, on arithmetic scale,

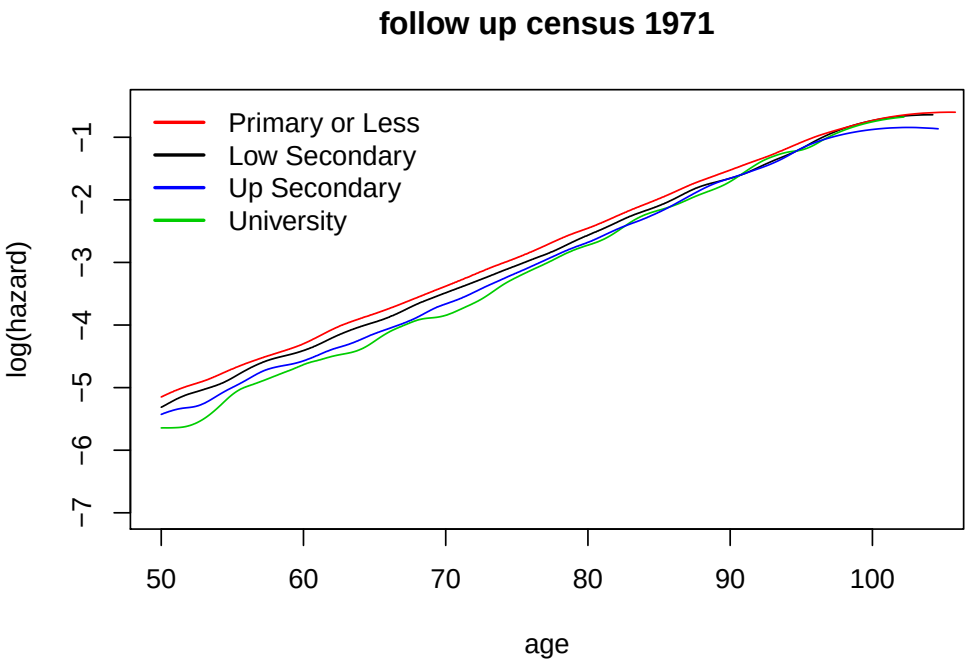


Figure 2.1: Male mortality hazard from age 50, on logarithmic scale, from the follow up of the 1971 census Turin population by four education levels.

would not be visible because of the exponential increase with age of the hazard curves.

The converging mortality trajectories at old ages of the black and white population in the USA, which has been observed from vital statistics for several decades, is a famous example of mortality convergence at old ages and has stimulated a lively debate. The age trajectories of black and white mortality rates not only converge, but produce a cross over, from a black to white ratio of 3 to 2 in young adulthood that declines to below unity at the oldest ages (Hambright 1969; Kitagawa and Hauser 1973). Although the validity of this pattern has been questioned, attributing the cross over to data problems and inaccuracy of the information about old ages (Coale and Kisker 1986; Preston et al. 1996), studies conducted on sub-populations, whose data were accurate and not affected by problems like misreporting or

death under registration, have also reported this phenomenon (Wing et al. 1985), testifying for its consistence. A possible explanation that has been advanced is that this pattern is due to selective effects of high early mortality, experienced by the black population, which leaves more robust individuals at old ages (Manton et al. 1981b).

2.5.2 Explanations for convergence

The selection process of the more robust individuals, as the frailer ones die out faster, is one of the explanations for the observed convergence of the mortality rates at old ages by subgroups of populations and is very important in the context of this research, because the theoretical framework of the frailty models relates specifically to this question. Before moving to a more detailed explanation of selection and frailty models, it is necessary to give an overview of the main arguments that address the question of the age pattern of socioeconomic differences in mortality.

Two main and opposite theoretical frameworks can currently be found in the literature that address the question of how the relation between age and socioeconomic status varies with age: the *cumulative disadvantage* framework and the *age as leveler* framework.

Cumulative disadvantage

The *cumulative disadvantage* hypothesis states that the effect of socioeconomic status on health and mortality increases with age, creating pathways of disadvantage (or advantage) over the life course. The mechanisms that are mentioned as the underlying factors refer to past experiences and events that accumulate and interact with other factors, like economic and social capital, creating a burden of accumulation of disadvantages (or advantages) that contribute to widen the inequalities (Ross and Wu 1996; Beckett et al. 2002). Possible ways for accumulation to take place are related to factors like a time lag between the experience of some deleterious event and its effects on health

(Lauderdale 2001) and increased vulnerability at old ages that makes differential exposure more dangerous for lower social classes, which are supposed to be disadvantaged in several risk factors like unhealthy work conditions, high levels of stress and economic deprivation, constantly across age groups (House et al. 1994). On an empirical basis, however, evidence supporting the arguments of the *cumulative disadvantage* hypothesis are limited (Dupre 2007).

Age as leveler

On the contrary, the *age as leveler* framework, states that health and mortality inequalities decrease at old ages. The explanations for this pattern are disparate.

Some authors stress that at old ages, regardless of social class, all people weaken and eventually die. From this perspective, at old ages, biological vulnerability takes over social determinants and, consequently, the association between social class and mortality is reduced (Elo and Preston 1996; House et al. 1990; Liang et al. 2002).

Marmot and Shipley (1996) place the accent on an alternative explanation of why the impact of social determinants loosens with age, emphasizing how, at old ages, people disengage from the systems that contribute to social stratification, like the labor market. Relative differences in mortality between those with upper class and lower class jobs have been found to narrow significantly after retirement, suggesting the importance of work in generating inequalities in health during working ages but not after.

The intervention of the welfare states and health systems is considered another potential explanatory mechanism for the observed reduction of socioeconomic differences in mortality at old ages. Many welfare systems have invested in improving the support offered to the elderly and offer facilitated access to health care and services for old people, therefore reducing poverty and deprivation and attenuating the differences by social class (House et al. 1994). This is supported by empirical findings for the United States, a coun-

try where a large portion of the working age population is not covered by health insurance and gets access to it, for the first time in life, only after age 65 thanks to the Medicare program. There is evidence that once the uninsured population obtains the insurance through Medicare, it shows an increase in the use of medical care and improvements in health (Decker and Rapaport 2002; Dor et al. 2006).

Finally, as mentioned earlier in the previous paragraphs, the convergence by social classes at old ages could be due to mortality selection. According to this explanation, the differences get smaller and smaller at old ages because the surviving population is more and more homogeneous.

2.5.3 Selection

All populations are heterogeneous because individuals differ in many characteristics, some observable and others not. Some individuals are frailer, some are more robust. The concept of unobserved frailty refers to this wide and complex set of characteristics that are not observable and make every individual different from another. Frailty is a general concept that does not distinguish between acquired weakness, lifestyle factors, environmental risks and innate biological frailty, but combines in a single measure all of the factors that operate to increase or decrease a given individual's mortality risk, regardless of the source of unobserved heterogeneity (Manton et al. 1981a).

Selection occurs as the population ages because of the presence of unobserved individual frailty. Frailer individuals, in fact, die faster than more robust individuals. This gradually selects the survivors towards robustness, thus making the population more homogeneous (Vaupel et al. 1979; Manton et al. 1981a). Heterogeneity in frailty may be a factor behind the observed declines and reversals with age of mortality differentials between populations or subgroups in the population (phenomena like mortality crossovers and convergence) (Vaupel et al. 1979; Vaupel and Yashin 1985). For example, in the case of two socioeconomic groups, the lower social group suffers from higher early mortality, which means being subjected to a stronger selective

force that leads to more compositional change as the population ages, leaving a more robust surviving population at old ages compared to the upper social group. This decreases the association between social class and health as the two subpopulations age.

The *cumulative disadvantage* hypothesis and the *age as leveler* hypothesis, especially when it relies on selective mortality, have often been seen as contrasting and mutually exclusive frameworks. However, recently it has been pointed out how the two points of view can be seen as not competing hypotheses, suggesting that the first one explains socioeconomic disparities at the individual level, while the second one describes the change at the aggregate level (Dupre 2007). The bottom line of this discourse is that what we see at the population level is the result of several individual dynamics interacting together and creating the final aggregate population dynamic. If at the population level the disadvantage of lower social classes tends to disappear, the selection hypothesis makes possible the explanation of this pattern without excluding the possibility that what the individuals experience at the individual level is a non-decreasing disadvantage (or advantage) over the life course.

The next chapter will focus on the theory of the frailty models, giving a more detailed explanation of the concept of unobserved heterogeneity of frailty and its mathematical formalization. Additionally, it will address why this component leads to selection processes, when and why this issue is crucial in the analysis of mortality dynamics and to what extent this is related to the estimation of the socioeconomic mortality differentials.

Chapter 3

Frailty Models

3.1 Unobserved heterogeneity of frailty

Many hidden differences in survival chances make individuals differ in their susceptibility to death. This source of heterogeneity is hard to fully observe, even when some covariates are available. Even if we could have at our disposal a very rich data set with several individual level information, there would always be a number of unobservable differences that operate to increase or decrease a given individual's mortality risk. This wide and complex set of characteristics is called unobserved frailty. Frailty does not distinguish between acquired weakness, life style factors, environmental risks and innate biological frailty but it indicates a general susceptibility to death (Manton et al. 1981a).

In cohort analysis, as a population ages, the frailer individuals die faster and gradually select the survivors in terms of robustness. In other words, the population undergoes a compositional change that, if neglected, can lead to biased estimates of the hazard rate or of the regression coefficients for observed covariates. This is because, at every age, the death rate is computed based on a population at risk whose composition is constantly changing. Thus, an estimate of the individual hazard rate that does not take the unobserved heterogeneity of frailty into account, as time (or age) goes by, will

underestimate the hazard function to an increasingly greater extent (Aalen et al. 2008b).

Ordinary methods in survival analysis are implicitly "based on the assumption that the study population is homogeneous up to some observed covariates." (Wienke 2010: 55) The concept of frailty provides a convenient way of introducing unobserved heterogeneity into models for survival data through the introduction of a random effect (Wienke 2010).

The problem of the effect of unobserved heterogeneity on survival data was first addressed by Beard (1959) but the term *frailty* was introduced by Vaupel et al. (1979)[1], that showed how unobserved heterogeneity of individual frailty has an impact on the dynamic of mortality at the population level.

Since then, a growing body of research has developed and widened the frailty models that are used to take the individual hidden differences into account in the analysis of population survival (Vaupel et al. 1979), to explain deviant behavior of mortality rates at old ages like cross-over and deceleration (Vaupel and Yashin 1983, 1985), to correct biased estimates of regression coefficients of proportional hazard models in duration dependence analysis (Chamberlain 1985) and to separate biological effects from compositional ones in studies about aging (Manton et al. 1986).

3.1.1 The theoretical framework

The framework of unobserved heterogeneity of frailty points out how different the mortality dynamics at the individual and at the population level can be.

According to the theory every individual has a specific level of frailty, z, that defines the individual hazard in a context of proportional hazard models as:

[1]In the same year, Lancaster (1979) independently developed the same concept, but with a different name, in econometrics. He developed the same model and applied it to the analysis of unemployment spells duration data.

$$\mu(x, z) = z\mu(x) \tag{3.1}$$

where $\mu(x)$ is the baseline hazard and z acts multiplicatively on it.

A population comprises many individuals, with different and unobservable frailty, that all contribute to the population dynamic. At any age x, what we observe is the mean mortality rate at that age, $\bar{\mu}(x)$, for the survivors of each frailty. Since frailty cannot be negative, this can be expressed as:

$$\overline{\mu}(x) = \int_0^\infty z\mu(x) f_x(z) dz \tag{3.2}$$

where $f_x(z)$ is the probability density function of frailty at age x among the survivors. $\mu(x)$ can be taken out of the integral and what is left inside is the mean frailty among survival at age x:

$$\overline{\mu}(x) = \mu(x)\overline{z}(x) \tag{3.3}$$

Equation 3.3 shows that individual and population hazards are different. In particular, the individual hazard $\mu(x)$ increases at a faster pace than the population hazard $\overline{\mu}(x)$, because the average frailty of the population, $\overline{z}(x)$, decreases with age, as the most robust individuals are selected. On a log scale, when age specific death rates are plotted against age, the points follow a straight line but they slow down at older ages, causing a mortality deceleration. In Figure 3.1 the blue lines represent the individual hazards with different frailties. As the frailer ones die out faster (blue fading lines), the population hazard, represented by the red line, decelerates. Eventually it would converge to the level of the lowest frailty. In particular, because $\overline{z}(x)$ decreases with age (given that the individuals with higher frailty tend to die earlier and the survivors who are left tend to be more robust, with lower level of frailty), the population hazard grows less fast than the individual one. The difference between the two becomes bigger and bigger at advanced age, as the selection process has had enough time to select the more robust individuals. The result at the surviving population level is a mixture model that

reflects the frailty compositional change over age and causes the mortality deceleration (Vaupel et al. 1979; Caselli et al. 2000; Vaupel 2010).

Figure 3.1 illustrates the basic setting of the frailty models: the individuals in a population have hazards that increase at the same rate but starting from different levels, where the levels are modulated by the frailty term. However, this is a simplification of the reality that is useful for modeling purposes. Heterogeneity of frailty can assume other forms, not necessarily proportional to each other, and still the resulting population hazard function must be interpreted with care.

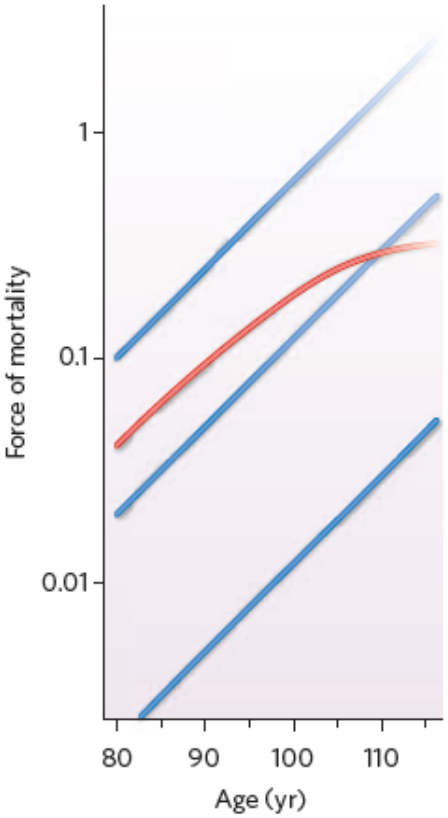


Figure 3.1: Individuals and population mortality (on log scale) - from Vaupel (2010)

With several different configurations of frailty, the differential selection of the individual at higher risks can produce patterns of mortality of the entire population that can be very different from the mortality experienced by its individuals (Vaupel and Yashin 1985). This can be represented by a simplified example of a population where heterogeneity is present in the form of two subpopulations with two different levels of frailty, represented in figure 3.2, from Duchateau and Janssen (2008).

In figure 3.2-a the two subpopulations have constant hazards but at different levels. This is a possible occurring situation in the case of the rate of

discovery of natural resources like coal, gold or oil. As the deposits that are the easiest to find are consumed (i.e., "die"), the rate of discovery of that resource declines, because it become more and more difficult to spot a new deposit.

Figure 3.2-b displays a typical reliability engineering situation, where a part of a device is defective and of low quality and the remaining part has a normal hazard rate that increases with age (i.e., usage). Initially, the failure rate of the total system decreases, but as the high mortality defective devices are eliminated, the failure rate increases again following the rate of aging of the remaining devices.

In figure 3.2-c, instead, the two subpopulations have increasing hazards but with different slopes. At first the population hazard increases. As the individuals of the group with the steepest hazard are selected out, the hazard suddenly declines toward the level of the group with a less steep slope. After some time, however, the hazard starts to increase again, following the hazard of the remaining group in the population.

Finally, figure 3.2-d represents the case, already discussed in this chapter, of two proportional (parallel) frailties that cause a deceleration of the population hazard at old ages.

3.1.2 The distribution of frailty

Explaining the relation between population and individual hazard is complicated because frailty is unobservable. A way out of this consists in assuming a distribution function for it. A convenient assumption that is usually made is to have a Gamma distributed frailty.

Gamma is a convenient distribution because it is positive (frailty can not be negative), has a very flexible shape (this is convenient because we don't know what the frailty distributions looks like) and its Laplace Transform allows to express the population survival function as a Gamma mixture of individual survival functions.

The Laplace Transform of a function $f(t)$ is a function $F(s)$ defined as

$$F(s) = \mathcal{L}f(t) = \int_0^\infty e^{-st} f(t)dt \tag{3.4}$$

If $f(x)$ is the frailty distribution $f(z)$ and s is the cumulative baseline hazard $M(x)$, the population survival turns out to be the Laplace Transform of the frailty distribution evaluated at the cumulative baseline hazard:

$$\overline{S}(x) = F(s) = \mathcal{L}f(z) = \int_0^\infty e^{-M(x)z} f(z)dz \tag{3.5}$$

The probability density function of a Gamma distribution with shape and scale parameters α and β is:

$$f(z) = z^{\alpha-1} e^{-\beta z} \beta^\alpha / \Gamma(\alpha) \tag{3.6}$$

Its mean and variance are $E(z) = \frac{\alpha}{\beta}$, $var(z) = \frac{\alpha}{\beta^2}$ and its Laplace Transform is $\left(\frac{\beta}{\beta+s}\right)^\alpha$. Therefore,

$$\overline{S}(x) = \left(\frac{\beta}{\beta + M(x)}\right)^\alpha \tag{3.7}$$

To make the model identifiable the frailty distribution can be standardized to mean 1 and variance σ^2, so that $\alpha = \beta = \frac{1}{\sigma^2}$.

It follows that, under Gamma distributed frailty, the population survival is:

$$\overline{S}(x) = \frac{1}{(1 + \sigma^2 M(x))^{\frac{1}{\sigma^2}}} \tag{3.8}$$

and the population hazard is:

$$\overline{\mu}(x) = \frac{d}{dx} - \ln(\overline{S}(x)) = \frac{\mu(x)}{1 + \sigma^2 M(x)} \tag{3.9}$$

where $\mu(x)$ is the standard individual (or baseline) hazard, $M(x)$ is the cumulative baseline hazard and σ^2 is the variance of frailty at the initial age.

Other distributions have been suggested for modeling unobserved heterogeneity of frailty (Hougaard 1986). Popular distributions are the inverse Gaussian (Hougaard 1984) and the log-normal (McGilchrist and Aisbett

1991). However, Missov and Finkelstein (2011), showed that, given the observed mortality pattern at old ages (that decelerates and, at very old ages, tends to reach a plateau (Gampe 2010)), the inverse-Gaussian and the log-normal distributions are not admissible, because they do not lead to a plateau in the mortality rate, while the Gamma distribution is among the admissible distributions.

The assumption of a Gamma distributed frailty oten yields useful mathematical results that are helpful for the development of the estimation procedures. In particular, frailty among the survivors is also Gamma distributed, but now with parameters α and $\beta + M(x)$; frailty among those who die is also Gamma distributed with parameters $\alpha + 1$ and $\beta + M(x)$.

This implies that at age x, the average frailty among survivors is somewhat lower than frailty among those who die and shows how the selection process operates eliminating the high-risk individuals:

$$\overline{z}(x)_{surv} = \frac{1}{1 + \sigma^2 M(x)} \tag{3.10}$$

$$\overline{z}(x)_{dead} = \frac{1 + \sigma^2}{1 + \sigma^2 M(x)} \tag{3.11}$$

It also implies that the variance decreases over age and the population becomes more homogeneous in absolute terms:

$$varz(x)_{surv} = \frac{\sigma^2}{(1 + \sigma^2 M(x))^2} \tag{3.12}$$

$$varz(x)_{dead} = \frac{\sigma^2(1 + \sigma^2)}{(1 + \sigma^2 M(x))^2} \tag{3.13}$$

3.2 Effects of unobserved frailty

When they introduced the concept of frailty, (Vaupel et al. 1979) pointed out that the component of unobserved heterogeneity is the cause of the deceleration of the mortality rates at old ages. Death rates, in fact, follow an

exponential low at adult and old ages (Gompertz 1825), but at older ages they tend to deviate from this pattern and increase less than exponentially, causing the phenomenon known as *mortality deceleration* and eventually reaching a mortality plateau. To address the problem of unobserved heterogeneity and to improve the fit of mortality models in populations, especially at very old ages, they suggested the use of frailty models, random effect models where the random term represents the unobserved frailty component.

Unobserved heterogeneity of frailty is also considered the cause of phenomena like hazard convergence or crossing overs between different populations, like, for example, the case of the black and white American populations. The mortality disadvantage suffered by the black population at adult ages tend to reduce and, eventually, to disappear at older and older ages (Hambright 1969; Kitagawa and Hauser 1973). Selective effects of high early mortality, experienced by the black population, could leave more robust individuals at old ages, causing the convergence with the white population, that is subjected to weaker selection (Manton et al. 1981b).

The same could apply to different socioeconomic groups. Lower educational groups have been showed to suffer from higher early mortality (van Raalte et al. 2011; Zarulli et al. 2012). This means that they are subjected to a stronger selective force that leads to more compositional change as the population ages. Thus, the surviving population at old ages is more robust compared to the upper groups and the association between socioeconomic indicator and health outcomes is weaker and weaker with age.

3.2.1 Mortality convergence by population subgroups

Converging mortality hazards by population subgroups are often observed at old ages. The death rates of the disadvantaged group (who suffers from higher mortality) tend to become more and more similar to the death rates of the advantaged group (who experiences lower morality), until the two mortality hazard functions converge. Selection processes could be the cause of such converging patterns.

For example, in the case of mortality by sex, it is well known that male mortality exceeds female mortality and that the disadvantage experienced by men often falls with age and the two hazards converge at old ages (Waldron 1985; Horiuchi 1997). Several studies try to investigate whether such a pattern has intrinsic biological determinants (Nakamura and Miyao 2008; Yang and Kozloski 2011), but the results so far are contradictory.

Such a pattern could be attributed to the stronger selection process in the male population, caused by the generally higher early mortality, that selects out frailer individuals faster than among the female population, eventually making the male group as frail as the female group, thus determining the convergence of the two hazards at old ages (Manton et al. 1995; Kohler and Kohler 2000).

Similar patterns of convergence are observed between race/ethnicity groups. A famous example of converging hazards is the black/white mortality crossover in the United States. The age trajectories of black and white mortality rates not only converge, but produce a cross over, from a black to white ratio of 3 to 2 in young adulthood that declines to below unity at the oldest ages (Hambright 1969; Kitagawa and Hauser 1973). Also in this case, selection processes due to higher early mortality, experienced by the black population, could be the cause acting behind this pattern (Manton et al. 1981b).

Similarly, the selection hypothesis can apply also to the convergence of the mortality age trajectories of different socioeconomic groups. Such convergence is a well established phenomenon in the literature about socioeconomic differential mortality (Antonovsky 1967; Huisman et al. 2003; Dalstra et al. 2006; Martelin 1994; Huisman et al. 2004). This is often explained as the effect of factors that operate as levelers of the differences at old ages, like an increased biological frailty or the intervention of welfare states and health systems (*age as leveler* framework). For a more detailed discussion about the theoretical frameworks that address the question of how the relation between age and socioeconomic status varies with age, see section 2.5.

The selection hypothesis represents an alternative explanation. The con-

vergence may be due to different paces of selection in the socioeconomic groups. Lower socioeconomic groups suffer from higher early mortality (van Raalte et al. 2011; Zarulli et al. 2012). Higher mortality groups are selected at a faster rate than the others, therefore, the frailest individuals in these groups are selected out faster than in the other groups. Consequently, at the same age (old age), what is left in these groups, is a more selected population in terms of robustness, compared to the lower mortality groups who undergo a slower selection process. The difference in the rates of mortality selection causes the mortality curves to converge and gives the impression that the effect of the covariate declines with age (Aalen 1988, 1994; Zajacova et al. 2009). This can be easily seen in Figure 3.3, which shows the case of two education groups (high education and low education). The effect of education at the individual level is stable over the life course, while the effect at the educational group level tends to decrease because of the artifact of selection.

Let $\mu_l(x)$ be the individual hazard for the low education group and $\mu_h(x)$ the individual hazard for the high education group. From equation 3.3 it follows that the ratio of the two educational population hazards is

$$\frac{\bar{\mu}_l(x)}{\bar{\mu}_h(x)} = \frac{\mu_l(x)\bar{z}_l(x)}{\mu_h(x)\bar{z}_h(x)} \tag{3.14}$$

This ratio declines because, although the ratio of the two individual hazards remains constant over age, the two groups undergo different paces of selection. In particular, the low education group is selected at a faster pace and its average frailty declines also faster

$$\frac{\bar{\mu}_l(x)}{\bar{\mu}_h(x)}\downarrow = \frac{\mu_l(x)}{\mu_h(x)}\frac{\bar{z}_l(x)\Downarrow}{\bar{z}_h(x)\downarrow}$$

More formally, considering that the frailty models framework assumes that the mean frailty at the initial age is 1, at such age, the left hand side of equation 3.14 is equal to the right hand side. This implies that, at the initial age, the ratio between the two educational groups is the same both at the individual and at the population level. Let r be this ratio at the initial age 0, the hazard for the low education group can be expressed in terms of the

hazard of the high education group: $\mu_l(0) = r\mu(0)_h$, where $r > 1$. It follows that, as age goes by, the ratio between the two population hazards is

$$\frac{\bar{\mu}_{lowEdu}(x)}{\bar{\mu}_{highEdu}(x)} = r\frac{1 + \sigma^2 M_h(x)}{1 + \sigma^2 r M_h(x)} \tag{3.15}$$

3.15 shows how the ratio between the two groups is r only at the initial age (when the cumulative hazard M is 0) and decreases as the age increases, because r is multiplied by a decreasing quantity.

It is important to clarify that frailty is a general concept. It combines all the factors operating to increase or decrease a given individual's mortality risk (acquired weakness, lifestyle factors, environmental risks and innate biological frailty), regardless of the source of unobserved heterogeneity. Most important, it does not imply any causal direction in the mechanisms connecting frailty and health susceptibility with socioeconomic status. Because of their higher mortality, the lower socioeconomic classes are considered to be frailer than the others. However, this does not necessarily imply that the higher frailty, attributed to them, has determined or determines their lower social condition. It may be the opposite as well: given that exposure to detrimental conditions (working related risks, worse housing conditions, bad health behaviors like smoking and so on) is socioeconomically patterned, lower socioeconomic groups can be considered frailer.

3.2.2 Mortality deceleration

Mortality deceleration is used to indicate the deviation of the death rates, at old ages, from the pattern of exponential growth that closely describes it at adult ages and that was identified by Gompertz (1825). In particular, at old ages, the death rates increase less fast than what would be predicted by extrapolating mortality from the previous ages.

According to the Gompertz model, at any age x, the force of mortality $\mu(x)$ is expressed by an initial mortality level, a, and by an exponential mortality increase by age, the parameter b, which is also defined as the rate

of aging (Shock 1967; Finch 1994) and represents the slope of the mortality curve on a logarithmic scale:

$$\mu(x) = ae^{bx} \tag{3.16}$$

The Gompertz model has been found to fit accurately the age mortality pattern not only for humans, but also for several other organisms. However, when large cohorts of individuals are investigated, mortality deceleration and deviation from the Gompertz pattern at old ages is commonly found. Studies have detected this phenomenon in medflies (Carey et al. 1992), Caenorhabditis elegans (Vaupel et al. 1994), mediterranean fruit flies (Carey et al. 1995), yeasts (Jazwinski et al. 1998) and several other organisms (Horiuchi and Coale 1990; Vaupel et al. 1998). Examples of studies on humans are Horiuchi and Coale (1990); Manton (1992); Horiuchi and Wilmoth (1997, 1998).

Evidence from the mentioned laboratory experiments points at heterogeneity of frailty as a possible biodemographic explanation for this phenomenon. Flies and worms kept in cages whose population density was held constant, have showed deceleration, thus demonstrating that such pattern is not an artifact of a decrease in crowding (Carey et al. 1995). When isogenic (characterized by essentially identical genes) populations are compared with genetically heterogeneous populations, significant deceleration is found in the latter ones (Brooks et al. 1994). However, leveling off of death rates occurs even when genetically homogeneous cohorts are studied (Curtsinger et al. 1992), indicating that other sources of heterogeneity are also important, like environmental heterogeneity or behavioral and physiological changes with age. Another major source of individual differences is represented by "chance", a set of random events and circumstances at different levels (molecular, cellular and so on) to which individuals are exposed from the period of development in embryo on, that affect the organisms. This explains why even genetically identical populations display phenotypical differences (Finch and Kirkwood 2000), calling attention to the fact that, ultimately, all populations are heterogeneous (Vaupel et al. 1998).

However, other studies have found evidence against heterogeneity as a major contributor to the phenomenon of late-life mortality deceleration (Drapeau et al. 2000; Rose et al. 2002; Reynolds et al. 2007). Nowadays, there seems to be no doubt about the fact that selection mechanisms of the more robust individuals play a role. Rather, what is questioned by some authors is the ability of the compositional changes to fully account for the observed mortality pattern (Wachter 2003). Selective mortality has stimulated a lively debate on the connections between hazard functions (mathematical summaries of the dependence of mortality on age), evolutionary theories of senescence and empirical research.

To the extent to which life span is considered an evolutionary trait, the discussion takes place around the general principle that operates to shape the hazard curves. The main theory relies on the evolutionary mechanisms of natural selection and random genetic drift and supposes that, at advanced ages, random genetic drift is the most importance force that drives the mortality trajectory, because, after the end of reproduction, selection becomes weaker and weaker. This implies that, from the perspective of drift and selection, all later ages are equivalent and mortality, though very high, shows no clear trend (Baeriswyl 2008).

Gavrilov and Gavrilova (2001) have developed models based on reliability theory, in which systems are composed by several components and when one or more of the major components fail, the systems also fail. They showed that reliability theory predicts both an exponential increase of the mortality rates by age and a late-life mortality deceleration. Also vitality models, in which an individual's vitality decreases during the lifetime according to a random drift process, until it reaches zero, have been used to describe mortality plateaus at old ages (Weitz and Fraser 2001; Steinsaltz and Evans 2004). In this case, the hazard function levels out because the distribution of vitalities among the survivors tends to converge to a fixed distribution.

Summarizing, the debate on mortality deceleration has risen to explain why the observed mortality pattern at old ages deviates from the exponential

growth that closely describes it at adult ages, formalized by Gompertz (1825) more than one and a half centuries ago. Although the Gompertz model remains the starting point for describing the age mortality trajectory, in the last decades several experiments have showed a consistent decelerating pattern of the death rates at old ages for varied organisms, including humans. As pointed out by Steinsaltz and Wachter (2006), the different approaches and explanations to the mortality deceleration can be grouped into three broad categories: one suggesting that the aging process changes essentially at very old ages, one seeking to find a more general low of mortality of which the Gompertz model gives only an approximation, one attributing the deceleration at selection operating in an heterogeneous population. A corollary of this third interpretation is that the hazard rate (or force of mortality) at the individual level actually follows an exponential increase, but it is masked at the population level by the selection process, that makes the population hazard increase slower than exponentially.

In practice, most discussion on heterogeneity starts from the frailty model perspective introduced by Vaupel et al. (1979), with particular focus on one specific frailty model, the "Gamma-Gompertz model that assumes a gamma distributed frailty in the population and a Gompertz baseline for the individual hazard $\mu(x)$. 3.9 leads to a logistic form for the population hazard $\overline{\mu}(x)$:

$$\overline{\mu}(x) = \frac{ae^{bx}}{1 + \sigma^2 \frac{a}{b}(e^{bx} - 1)} \tag{3.17}$$

where a and b are the Gompertz parameters of the baseline hazard and σ^2 is the variance of frailty in the population at the initial age of observation.

3.3 Unobserved frailty in PH models

When dealing with life time data, to analyze the effect of observed covariates on the survival chances, they can be included by using the framework of the proportional hazard models. This implies assuming that a covariate has the

effect of multiplying the hazard function by a constant.

Individuals differ for many characteristics and, very often, only a limited number of covariates are available. In many practical cases, it is impossible to include all the relevant covariates in the analysis. Moreover, even when the investigator has at his disposal a very rich dataset with several individual level information, a source of unobserved (because not available or not measurable) heterogeneity still persists.

3.3.1 Bias due to omitted variables

In these cases, the presence of unobserved heterogeneity and selection processes poses problems for calculating the effect of the covariates on the survival chances. Not taking this component into account, in fact, biases the parameter estimates of both the hazard rate and the regression coefficients. In linear regression models, the bias occurs only if the omitted covariates are correlated with the variables included in the model, while in hazard models, the estimates are biased even if the omitted source of variation is uncorrelated with the predictors included in the model (Gail et al. 1984; Trussell and Rodriguez 1990; Rodriguez 1994). This is caused by the most distinctive feature of survival analysis, the time dependent hazard. Selection, due to the unobserved (i.e., omitted) covariates, acts constantly on the risk set, gradually eliminating the frailest individuals and selecting the most robust ones, who remain in the risk set (Chamberlain 1985; Aalen et al. 2008a,b).

Ignoring the presence of hidden heterogeneity can seriously affect the estimates of the effect of the observed covariates on survival. When one important factor is ignored, the relative risk of the different treatments on the survival times is underestimated, leading to a loss of power for the treatment comparison (Schumacher et al. 1987; Schmoor and Schumacher 1997). More generally, when only one covariate is left in the model, whatever the number of omitted variables are, the effect of that covariate on survival is always underestimated; when several covariates remain in the model, underestimation still holds, at least up to some reasonably long time (Bretagnolle

and Huber-Carol 1988). Other econometric and biometric studies provide further substantial evidence for the bias introduced by ignoring the presence of hidden variability (Lin et al. 1998; Betensky et al. 2002) and show that the amount of bias depends on the hidden variability present in the population. The amount of the bias also depends on the extent of heterogeneity and the higher the heterogeneity of frailty, the bigger the bias is (Henderson and Oman 1999).

Therefore, drawing inferences about the individual dynamics of risk from the observed population dynamics while ignoring the selection process due to unobserved heterogeneity, can lead to biased results and wrong conclusions about the hazard rate and the effect of the covariates.

The frailty approach in the analysis of life time data allows one to account for the presence of unobserved variability, which has an impact on the survival chances, by introducing a random effect for the frailty term that is multiplicative on the hazard function. There are two broad classes of frailty models: the univariate frailty models and the multivariate frailty models.

3.3.2 Univariate frailty models

This class deals with univariate, independent and identically distributed survival data, for example the age at death of unrelated individuals. In the traditional proportional hazards regression setting, the hazard depends on the baseline hazard, assumed to be unique and common to all individuals, and on the observed covariate profiles. In addition to them, in the frailty model, the hazard depends also on the frailty, introduced in the model as a random effect term that modulates the baseline hazard multiplicatively. In this way, the variability of the data is split iton two parts, one accounted for by measurable and observed risk factors and one resulting from the hidden frailty.

The introduction of the variable frailty aims not only to account for omitted covariates, but also to describe the non proportionality of the population hazards (for example when the hazards of two subgroups, identified by a co-

variate, converge at old ages) in order to improve the fit of the model (Yashin et al. 2001).

3.3.3 Multivariate frailty models

This class is used with multivariate survival data, where independence between event times cannot be assumed. This is the case, for example, of recurrent events for the same individual (occurrence of a disease) or the age at death of related individuals, like twins or parents and children. In these situations, the event times are clustered, and each cluster (the family, the pair of twins, the individual and so on) is assumed to have a common hidden frailty that is responsible for creating dependence between the event times within the cluster. In this setting, the assumption is that the event times are independent, conditional on the latent frailty variable.

Two models have been developed in the multivariate setting. The shared frailty model assumes that individuals in a cluster share the same frailty and can be used for strongly related individuals, like monozygotic twins, or repeated measurements for the same individual. However, this model has some limitations, mainly attributable to the fact that the assumption of an identical frailty for the individuals in the cluster can be too strong and restrictive in several situations (for a detailed discussion on these limitations see, for example, Xue and Brookmeyer (1996); Wienke (2010)).

To avoid the limitations, the correlated frailty models have been developed. The individuals in each cluster are assigned to have separate frailties, in the form of random variables that are assumed to be associated and to have a joint distribution. The correlation between frailties is a much less strict assumption than sharing and it does not restrict the correlation to one type. For example, the correlation could also be negative, inducing a negative association between survival times in the cluster, while in most cases, the shared frailty model only induces a positive association within the cluster.

3.4 Frailty models and differential mortality

In the last decade, interest in the connection between unobserved heterogeneity and socioeconomic differences in health and mortality has increased, as has the possibility to use longitudinal data sets, that are increasingly available (and the only type of data that is appropriate to study this issue).

Most of the debate is focused on the role of the selection processes on the life course pattern of socioeconomic differentials, although the number of studies directly addressing this question is still limited. The majority of the existing studies try to evaluate the role of selective mortality in determining the observed converging hazards by socioeconomic groups at old ages. Decreasing health and mortality differentials with age, in fact, are the most common finding but, while the relation between socioeconomic status and health and mortality inequality is well established in research, whether this relation strengthens, weakens or is stable over age is still unclear.

Most studies are focused on the role of selective mortality on the age pattern of health outcomes (like self-perceived health, functional limitations and illness incidence), whereas studies on mortality outcomes are less numerous and often carried out in conjunction with the analysis of some health outcomes. The majority of the studies use data from longitudinal surveys. However, survey data are often affected by limitations like sample attrition, small sample size and short follow-up period. Moreover, in panel studies, the characteristics of the non respondents may be related to poor health, especially at old ages (Goldman et al. 1995).

The primary strategy used for addressing mortality selection in these studies is to keep in the sample those who die during the observation period, imputing them a value for their health status with some method. However, there are several ways to impute health, each of them potentially affecting the result. As pointed out by Herd (2006: 9), "there are numerous ways to impute health. And in fact, there is no perfect way. In essence, this is a thought exercise." Finally, other studies investigate the life course patterns of the relationship between socioeconomic status and health and mortality,

trying to disentangle age, period and cohort effects.

The literature does not report consistent results. Some studies find that mortality selection causes the convergence of the hazards at old ages, others find that such a bias does not take place. The central question in the debate about convergence of health inequalities is whether this pattern is due to differential aging processes in different groups, to changes in the impact of socioeconomic status over life or to compositional changes in the surviving population. According to the first mechanism, upper social groups are able to postpone health decline up to older ages but then they catch up with lower classes (Beckett 2000; House et al. 1994). Alternatively, what changes with age is the social context and its effect on health, which makest it possible that factors like retirement and social security slows down the health deterioration for those in the lower classes (Herd 2006). Finally, compositional change reflects not differential aging processes but different paces of selection that different groups undergo.

Ferraro and Farmer (1996), analyzing a 15-year long longitudinal survey, found that selective mortality biased downward the racial and gender differentials in the US, for two of the four health measures they used in the analysis. On the contrary Beckett (2000), analyzing the role of sample selection due to loss of follow up in panel data in evaluating the age-specific association between education and health outcomes inequalities, found no evidence that sample selection is responsible for the convergence in education differentials in health conditions or functional impairment.

Studies showing decreasing health inequalities at older ages, however, could be displaying cohort effects and not age effects. Lynch (2003) showed that educational differences in health are larger among younger cohorts than among older cohorts, and this could explain decreasing differentials with age in cross sectional analysis or in longitudinal settings with a wide range of different cohorts. Lynch (2003) also found that once controlled for cohort effect, the effect of education strengthens with age, a pattern that would stay unnoticed otherwise.

When evaluating the correlation between mortality and education, in addition to cohort, one must consider also period effects. Period events, like for example the cardiovascular revolution and the subsequent reduction of cardiovascular mortality risk, place a period constraint on the way mortality can be affected by education (Lauderdale 2001).

Herd (2006) avoids potential cohort effect following an individual cohort over time, from the Health and Retirement Study (1992-2002), and found that mortality selection did not fully explain the shrinking educational disparities in functional health at old ages. However, some data limitation could have affected the results, such as inadequate follow up that did not allow the following of individuals into their mid to late 70s and 80s. That the mortality selection bias is unlikely to be a major explanatory factor for the convergence of health trajectories was found also by Kim and Durden (2007), whose results also indicate that age-related patterns of health trajectories may differ according to the used dimensions of socioeconomic and health status.

On the contrary, in the study by McMunn et al. (2009), selective mortality appears to explain part of the decline in health inequalities at old ages, particularly in relation to heart disease. Dupre (2007) observed convergence in disease prevalence, but the rates of disease incidence and mortality were greater for the low education groups than for high education groups. This suggests that educational differences in disease incidence and mortality resulted in the convergence in disease prevalence, reflecting the bias caused by mortality selection.

These conflicting results have caused a debate where some authors question the validity of the accumulation or continuity of inequalities over life and support the Age as a leveler hypothesis, while others affirm that the Age as a leveler hypothesis does not adequately describe the real patterns, especially at the individual level.

More recently it has been suggested that the age as a leveler hypothesis needs to be revised. Studies have found stable mortality differentials across ages but converging differentials when health deteriorates, suggesting that

what acts as leveler is not the age, but the health status (Hoffmann 2005, 2011a). Even if it is reasonable to assume that increasing age is related to worsening health, for analytical purposes, these two dimensions should be kept separate (Kelley-Moore 2010) and while age increases for everyone, health can decline in varied ways between different social groups.

Research question

While studies aiming to assess if selection explains the converging health trajectories exist, there is a lack of studies trying to evaluate the impact of unobserved heterogeneity on the estimation of coefficients for the mortality risk of the socioeconomic groups in survival models. This is quite surprising given the strong evidence showing how, ignoring unobserved characteristics causes the estimates to be biased towards zero, thus leading to an underestimation of the relative risks between groups.

In this study I use longitudinal individual census linked data for the population of Turin, over a follow up period of more than 30 years. I investigate the presence of selection processes in the mortality patterns, from age 50 on, by educational level, used as proxy of the socioeconomic condition. The choice of this indicator is justified by the literature: education is notoriously associated with several health outcomes, remains stable over the adult life time not being affected by the fluctuations that may happen in the working life, is applicable to individuals that are not active in the labor market and is equally valid for men and women. For more details about advantages and disadvantages of socioeconomic indicators, please see section 2.3.

Turin is an industrial Italian city of nearly one million inhabitants, therefore the number of individuals and the length of follow up available in this study are far beyond what is usually available in studies using health surveys. The aim of the study is to investigate whether unobserved hetrogeneity matters by assessing if and how the estimates of the mortality differentials are affected by taking into account unobserved heterogeneity of frailty. Moreover, in this study I want to investigate whether the observed pattern of

mortality can be better explained by the framework of the frailty models, thus testing the hypothesis that decreasing effect of socio-economic variable with age is an artifact of selection. To my knowledge this is the first study that addresses these questions using the analytical framework of the frailty models in Italy.

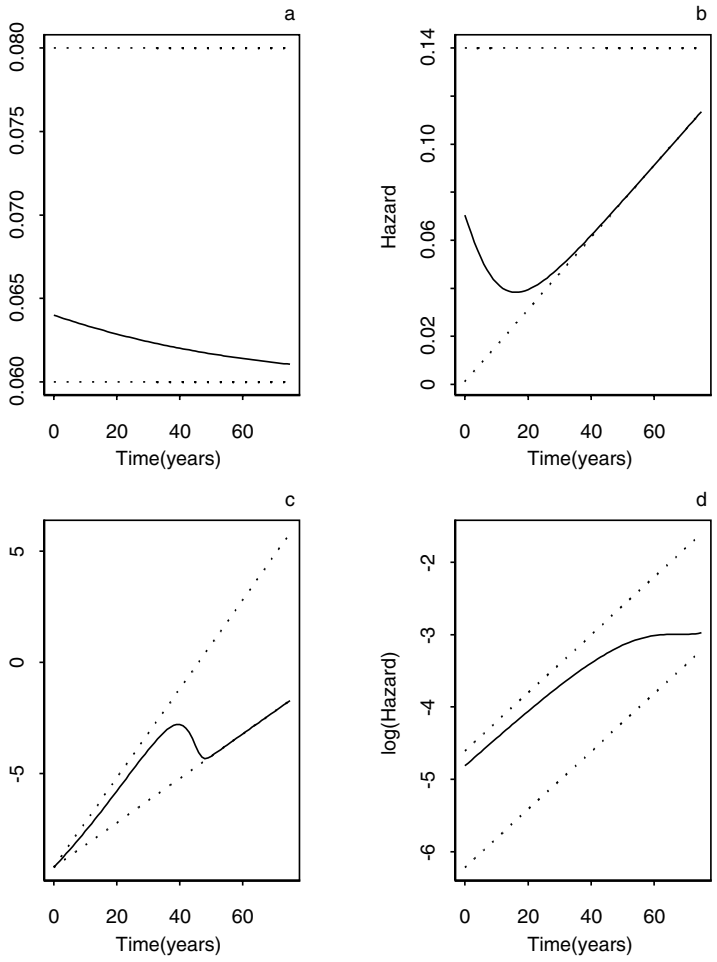


Figure 3.2: The population hazard function (solid line) and the hazard functions of the two subpopulations (dotted lines) for four different situations: (a) two subpopulations with constant but different hazard rate, (b) one subpopulation with constant hazard rate and one subpopulation with increasing hazard rate over time, (c) two subpopulations with increasing hazard rates but with different slopes, and (d) two subpopulations with hazard rates that increase at the same rate but from different levels. Note that the hazard functions in (c) and (d) are depicted on the logarithmic scale - from Duchateau and Janssen (2008: p.41)

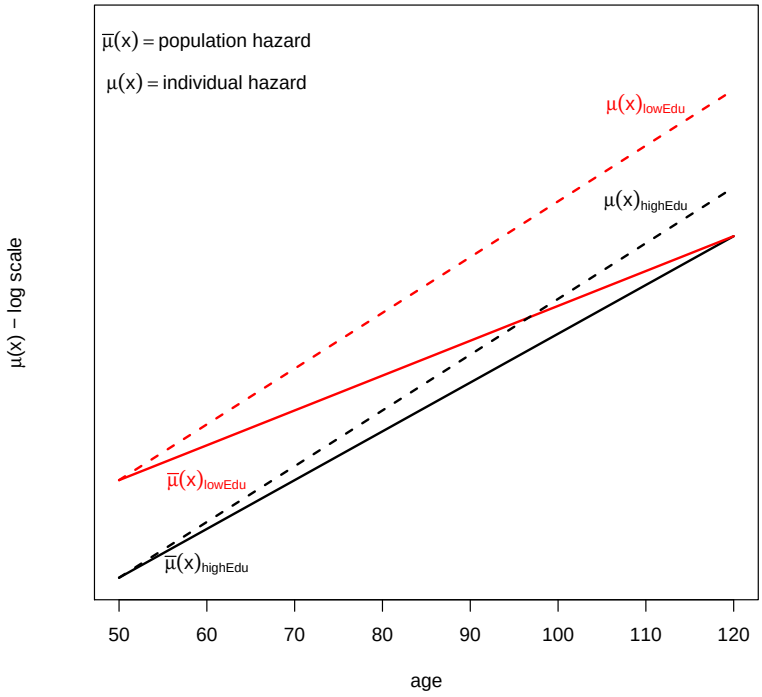


Figure 3.3: Selection process in two education groups (high and low). The higher mortality group (low education) is selected, in term of robustness, at a faster rate than the lower mortality group (high education). This differential pace of selection makes the two educational group hazards converge as the individuals age.

Chapter 4

Data

4.1 TLS: an overview

The Turin Longitudinal Study (TLS) is an Italian study that links data from the last four censuses (1971, 1981, 1991, 2001), the vital statistics registry and the health system archives for the city of Turin. The system of integrated data sources permits the creation, via record linkage techniques, of longitudinal individual histories over a follow up period of almost 40 years. The study includes the entire population of Turin, an industrial city in the North-West of Italy of nearly one million inhabitants. Information at the individual level and at the contextual level can be connected, thus creating a very rich data base for several analytical purposes (Demaria et al. 2001; Demaria 2010).

TLS is an integrated longitudinal system set up by the Piedmont Region, through its Epidemiology Unit, in collaboration with Local Health Units TO1, TO2 and TO3 and the National Institute of Statistics; data treatments are authorized by the National Guarantor Authority for Privacy (Psn 2011).

Archives in TLS

Demographic information is taken from the city population registry. Every individual and his family is assigned a unique identification number. Date

of birth, date of death, dates of emigration from the city and dates of immigration to the city are reported. Therefore, every individual that has ever passed through the city of Turin since the beginning of the study is included in the data set. In case an individual emigrates and then immigrates back, the multiple migratory spells are traced, but if the individual does not immigrate back, he exits the study and is right censored at the time of emigration. From the population registry it is also possible to obtain information on the personal relation within families, such as the parental relation, and the residential histories.

Socioeconomic and contextual variables are derived by the census form. Moreover, indirect and composite indexes can be derived by combining information pertaining to different parts of the census questionnaire, like micro-area deprivation indexes.

Finally, TLS includes various health archives. Each of these archives covers a different period of time. Up until the end of 2010, the system incorporates the hospital discharges registry (for the years 1995-2005), the medical drug prescriptions registry (for the years 2000-2006), the specialist medical consultations registry and first aid interventions registry (for the years 2002-2005), the regional cancer registry (for the years 1985-2006) and other specific pathology registries, such as diabetes, but the Study constantly works to enlarge its data base and to gradually include new records and information.

Summarizing, TLS is a study that links together several administrative, health, census and demographic data sources into an integrated system. Each archive included in the study contains specific information and covers a specific period of time. Table 4.1 reports the dimensions of the main archives.

Data quality

Data quality is very high and is ensured by a successful record linkage between the master data file (the city population registry) and the other archives.

The share of linked hospital discharge records in every year is between

Data source	**Dimension**
Population Registry	
City Residents Registry (up to 31-07-2007)	2,176,442
Migratory History (1971-2005)	3,878,626
Parents-Children connection (1971-2003)	793,448
Census	
Census 1971	1,023,578
Census 1981	1,091,033
Census 1991	929,891
Census 2001	832,106
Health archives	
Cancer Incidence Registry (1985-2006)	123,078
Hospital Discharges (1995-2005)	3,863,855
Medical drug prescriptions (1997-2006)	55,960,422
Specialist Medical Visit and First Aid Service (2002-2005)	44,258,010

Table 4.1: Dimension of the linked archives in TLS, up to 31-12-2010.

97% and 98%, with the exception of 1996, when the linkage matched slightly less than 96% of records.

The share of linked records between census and population registry is also very high, 97.8 %, 96.6 % and 96.2 % respectively for the 1981 census, 1991 census and 2001 census, while for the census of 1971, the linkage was less effective, linking 84.7% of the census records. Problems related to the informatization of the records are likely to have affected the procedure for the oldest census.

Although the percentage of unlinked records is low, these data could represent a population selected for a particular characteristic, causing potential bias in the analysis. However, an analysis of the distribution of the unmatched census records has shown that they are evenly distributed across the main socio-demographic categories, excluding potential bias problems (Demaria 2010).

4.2 TLS: the census follow-up

The focus of this thesis is mortality. Therefore, in this study I used data from the linkage between the four censuses and the city vital statistics registration.

For each census in 1971, 1981, 1991 and 2001, the census-registered population has been linked with the deaths and the emigrations which occurred in the population of Turin, resulting in an individual mortality and emigration follow up. At the time of this analysis, the record linkage between the two sources was available until the end of July 2007.

The original data came in four different linked sets, one for each census. In this way, individuals in Turin at the time of different censuses, were registered multiple times (one for each census in which they were registered). I merged the four data sets using the the variable *numind*. This variable identifies the individuals and allows the tracking of the records through different archives, thus identifying the single individuals ever registered in Turin during the period 24-10-1971 (date of the first census) and 31-07-2007 (end of the follow up). This population comprises 1,655,327 individuals, of which 847,330 are women and 807,997 are men.

4.2.1 Available variables

From the census data files I had access to the information about individual's *sex*, *macro region of birth* (according to the classification *North-East*, *North-West*, *Center*, *South and islands* and *abroad*) and *education level*. This variable is considered a good proxy for socioeconomic status and is often used. In a recent review of the literature Doblhammer et al. (2009) point out that education is a very good indicator for the socioeconomic status of a person because it presumably influences income and occupation the most, being most likely chronologically and causally prior to them. Moreover, education level is a stable individual characteristic at adult and old ages.

The educational information based on the census ensures better data quality than what can be obtained from other data sources like, for example,

the death certificate. First of all, the categorization used in the census form is more articulated than the one used in the death registration. Second, the information drawn from census linked data is known to be more reliable because it is self assigned, while in the case of the death certificate, it is assigned by the relatives or by the public officer who registers the death and this can cause precision and misclassification problems. Several studies have analyzed the validity of the education information in death certificates and found that it was biased. Shai and Rosenwaike (1989) and Sorlie and Johnson (1996) found significant misreporting problems, especially for deaths at older ages. Rosamond et al. (1997),Marmot and McDowall (1986) and Goldblatt (1989) found that the misreporting did not strongly affect the estimates of the differentials. Nevertheless, the studies all found education misreporting.

The first three censuses, 1971, 1981 and 1991, have the same educational categorization while the census of 2001 has a different one. The difference, however, is represented only by a more detailed classification used in 2001, that can be easily matched to the one used in the previous censuses. Therefore, I could adopt the same classification of 6 groups: not available, no degree (cannot read nor write), No degree (but can read or write), Primary education (italian *Scuola elementare*), Lower Secondary education (italian *Scuola media inferiore*), Upper Secondary education (italian *Scuola superiore*), University Degree. This classification scheme is reported in table 4.2.

In order to facilitate the comparison over the long period of time covered by the follow up, for the analysis I grouped them into three broad categories: High Education (Upper Secondary or higher), Medium Education (Lower Secondary) and Low Education (Primary or lower). I think that using such broad classes helps the comparability over time and over different cohorts.

From the city population register the available information are *date of birth*, *date of exit* from Turin, and *cause of exit* (whether death or emigration).

Table 4.2 shows the list of the available variables for the analysis.

4.2.2 Study design

The study focuses on socioeconomic differential mortality from late adult ages on. Notoriously selection needs some time to make population and individual hazards diverge and the distorting effects of selection are significant only from old adult ages on.

Therefore, I included in the analysis the individuals who were aged 50 and more at the beginning of the follow-up (area A in fig. 4.1) and those who were younger at that time, since the moment they reached age 50 during the period of analysis (area B in fig. 4.1). These include 391,170 men and 456,216 women observed over 36 years, from 24-10-1971 to 31-07-2007.

The focus on people older than 50 makes the chosen educational classification reasonable. Prior to the school reform which occurred in the 60s, only elementary school was compulsory and lower secondary school was considered the intermediate level. Moreover, individuals who attended high school, especially women, represented a very selected population who, almost certainly, would have continued their schooling career attending university.

4.2.3 Missing data

Only for 1,751 individuals out of the total population of 1,655,327 (0.1 %) are date of birth, date of exit from the population and cause of exit not available, so they were dropped out of the analysis because it is impossible to assess their age. Interestingly, 1,750 of these individuals were born abroad and only one in Italy.

Regarding the macro-region of birth, there are no missing data.

Educational level is not available in 201,767 cases but these refer mostly to young children that have not completed the first cycle of studies because of their age, and that are likely to be classified as *not available* rather than *no degree*. In fact, 192,659 of these cases refer to individuals whose age at first appearance in the follow up was less than 11, an age at which the first stage of education is supposed to be completed. Therefore only 9108 cases,

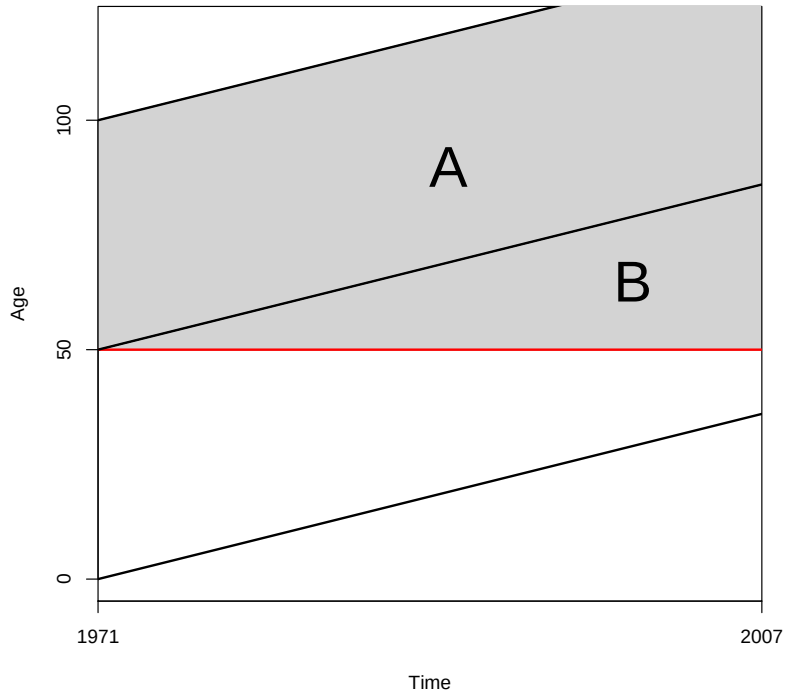


Figure 4.1: Stylized representation of the follow up data.

0.55% of the total unknown educational values, pertain to individuals who could have completed the first stage of the education system and could have earned a deggree that is missing. The share among the population selected for the analysis, that is, individuals from the age 50 on, is 0.41% (3,538 cases out of 847,386) and is similar among the two sexes: 0.43% among men and 0.40% among women, as reported in table 4.3.

4.3 Descriptive results

4.3.1 Distribution by education and region of birth

In this section descriptive results will be given in order to have an overview of the data.

Table 4.4 shows the distribution of the follow-up population by education and macro-region of birth. The majority of individuals have low education level. This percentage is especially high among women 65% while among men it is 56%. The shares of medium education are similar for both sexes while the percentage of highly educated individuals among men is higher than among women (20% and 12% respectively).

Turin is a north-western city, so individuals from the "North-West" macro-area can be considered natives (or short distance migrants). Accordingly, the share of individuals born in the macro area "North-West" is the biggest. Among the migrants, 33% of men and 28% of women were born in "South & Islands" area, around 10% for both men and women come from the "North-East" area and only small numbers from the central region and from abroad.

The distributions of the population changes across different cohorts, as shown by the figures 4.2 and 4.3.

The share of the population with low was the largest among older cohorts (between 70% and 80%), then it gradually decreases from older to younger cohorts. Meanwhile the proportion of individuals in the medium and high education categories increased. Among the youngest cohorts included in the study, the proportion falling in the medium education category represents the largest group of the population for both of the sexes, the share with a low education is 30% and 35% for men and women respectively and the share falling in the high education category is similar for men and women (around 25%-27% of the total).

For the first time, the proportion of men and women with high education is almost the same, whereas in previous cohorts women had always been the minority: for most of the time the proportion of women with high education has not gone beyond the 10% threshold, while the share of men was almost twice as much.

Concerning those in the low education group, the gender gap still persists. Although the difference tends to lessen, even in the youngest cohorts the proportion of women with low education is bigger than the proportion of

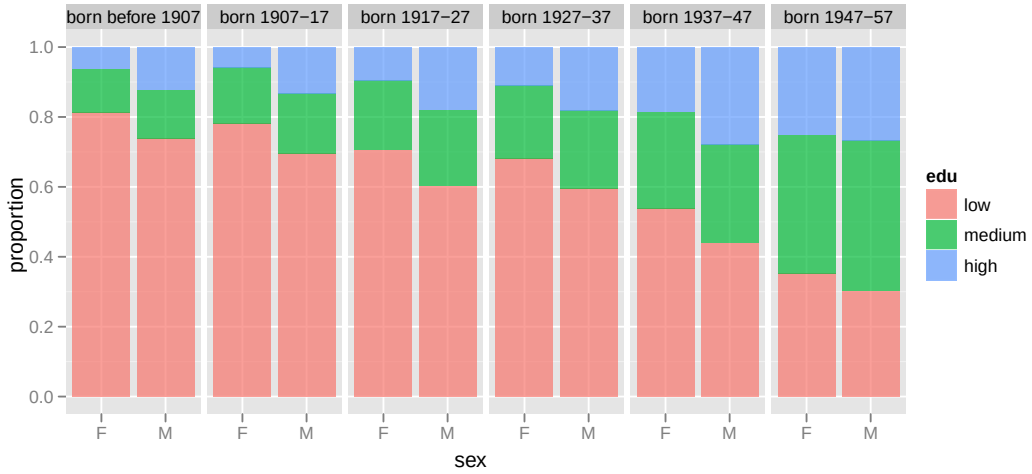


Figure 4.2: Distribution of the Turin population by education level for men (M) and women (F) in different 10 year-birth cohorts.

men.

Turin is an industrial city with a long history of immigration from the poorer southern regions of Italy, particularly after WWII. The share of migrants from the southern regions constantly grows from older cohorts to younger ones and among the immediate postwar generations, almost 50% were born in a southern region and only slightly more than 40% were natives of the North-West.

Also the "North-East" area is an important source of migrants to the city, especially among older cohorts. However, its contribution decreases with younger cohorts, following the more recent economic development of that area in the decades after the war.

The proportion coming from the "Center" and from "Abroad" areas are small and negligible.

Migrants were often individuals of low education employed in the car industries which characterize the economy of the city. Figure 4.4 shows that among men with low education, slightly less than 40% are natives and slightly more than 40% were born in the South. In the other educational groups, by

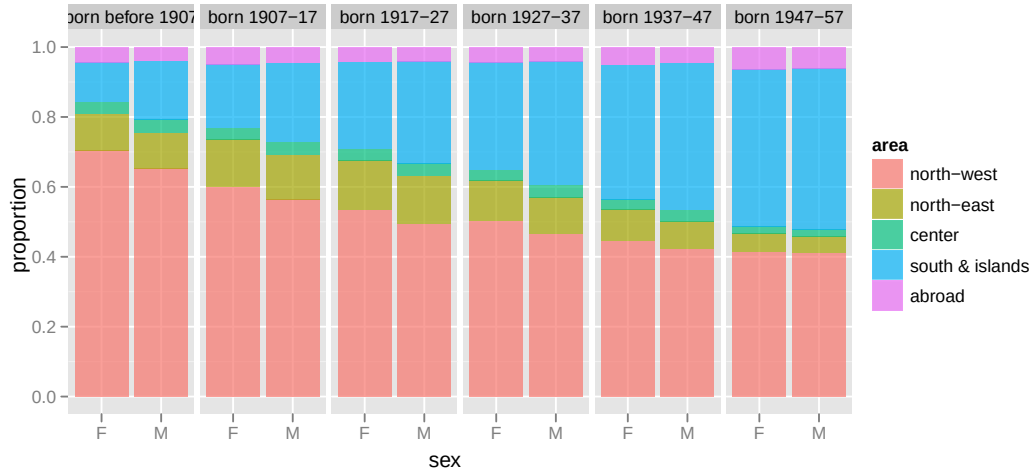


Figure 4.3: Distribution of the Turin population by macro-region of birth for men (M) and women (F) in different 10 year-birth cohorts.

comparison, the share of individuals from the South is smaller than that of the natives and it decreases as the level of education increases. Almost 70% of the highly educated men were born in the "North-West" area and only around 20% in the "South & Islands" area. Women present a similar pattern.

Also in this case the distribution changes across different cohorts. From older to younger cohorts the number of women with high education increases, independently of the macro-region of birth. There are, however, regional differences: the increase is more conspicuous among those women born in the North-West than those born in the South (see Figure A.1 in A).

4.3.2 Total Population Mortality

Before moving into the analysis based on the individual level data, it is necessary to adopt a less fine longitudinal perspective and aggregate the data in cohorts in order to catch the underlying mortality trends.

The follow up starts at the 1971 census. Different cohorts are observed from different ages.

The figure 4.5 shows the death rates, on logarithmic scale, of four different

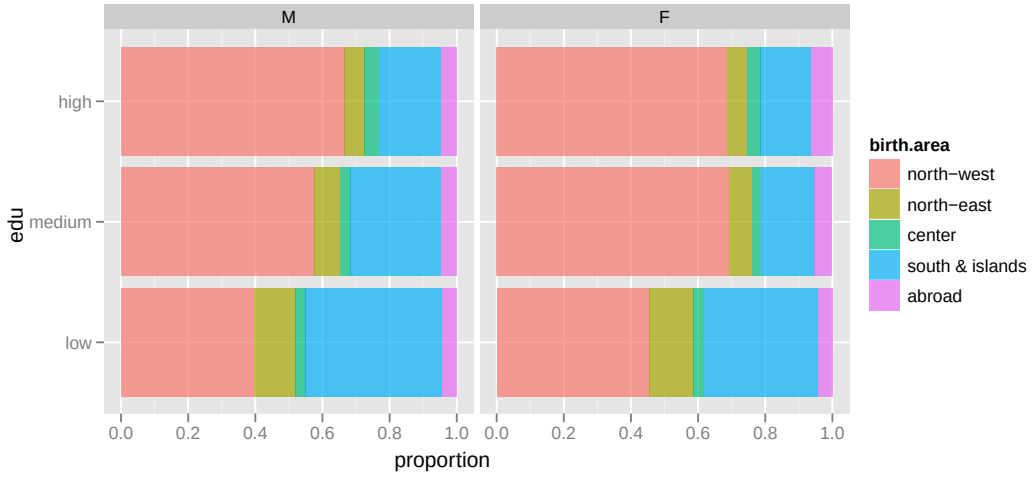


Figure 4.4: Distribution of the education groups in the Turin population by macro-region of birth of the individuals.

5 year birth cohorts, followed up from the 1971 census. For younger cohorts, like the cohorts aged 65-69 and 70-74 in 1971, the mortality curve goes on for all the 36 years of follow up, reaching the ages 100 and 105. For older cohorts, in contrast, the mortality curve is "shorter" because there are no observations after the ages 105-110, especially among men.

A first glance at the data shows that the death rates tend to deviate from a straight line and decelerate at old ages. According to the theory of the Frailty Models, mortality deceleration indicates the presence of selection process: as frailer individuals die out, the observed population hazard is dragged down because the survivors at older and older ages are more and more robust. Moreover, selection needs time to take place and, the deceleration is mostly visible for older cohorts, for which the follow up reaches advanced ages.

The process is clearly visible among women. Among men, though, the phenomenon is less visible. One possible explanation is that the male population is less heterogeneous than the female population and, therefore, undergoes a weaker selection. This seems to be consistent with what has been found in previous studies on different developed countries and in Italy (Barbi

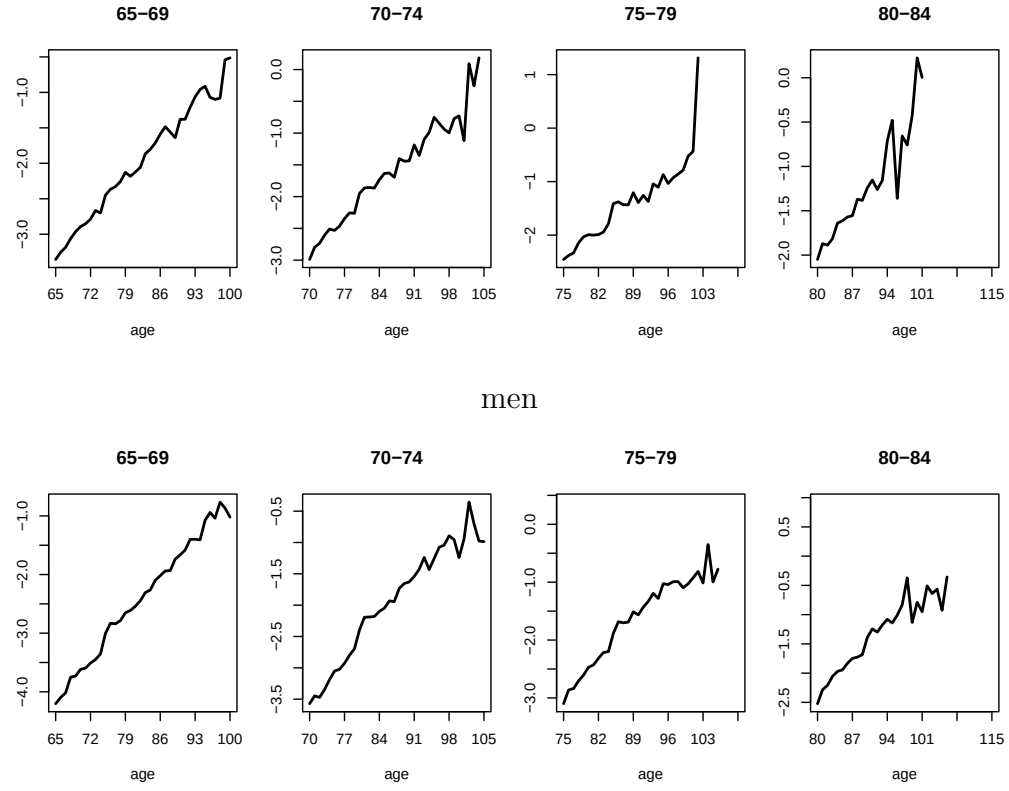


Figure 4.5: Death rates on logarithmic scale of four different 5 year birth cohorts, followed for 36 years from the census 1971.

2003).

4.3.3 Mortality by Education level

The empirical cohort death rates by education (on logarithmic scale) in figure 4.6 reveal the presence of the usual pattern of converging hazards by educational groups at old ages. The converging pattern seems slightly more pronounced among women than among men. Men, in fact, quite surprisingly show less marked educational differences, while the literature on the topic usually finds wider mortality gradients among men. Women's education level, in fact, is often less reflected in their socio-economic status and more likely

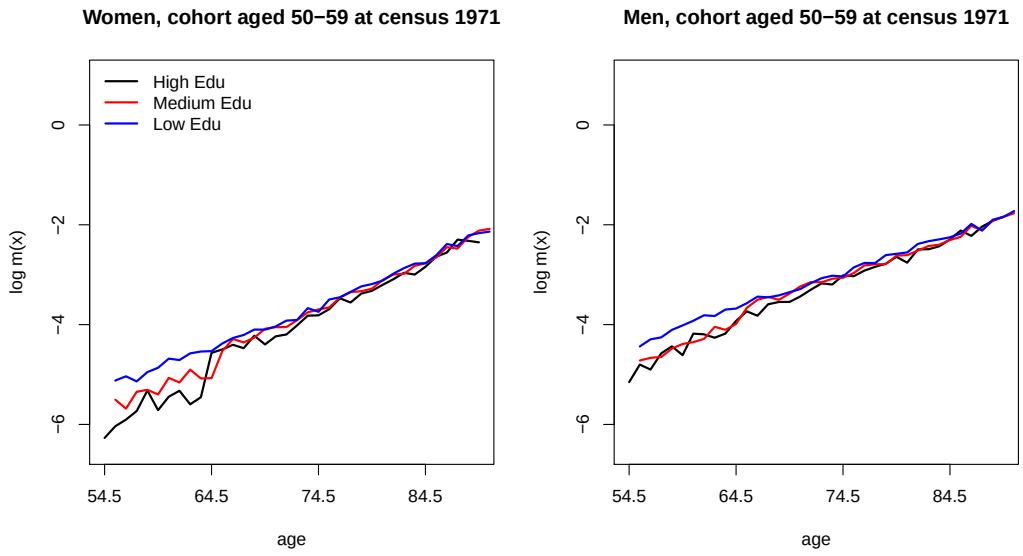


Figure 4.6: Death rates, on logarithmic scale, for the birth cohort aged 50-59 at the beginning of the follow-up (1971 census) by education level: high, medium and low.

to be influenced by the partner's education (Koskinen and Martelin 1994; Elo et al. 2006).

It is likely that the migratory history of the city may play a key role. Turin is one of the most important industrial cities in Italy and has experienced vast immigration flows in the past decades. Most of the immigrants came from the poorer southern regions of the country. Migrants are supposed to be healthier. Moreover, men (but not women) from the southern regions have been shown to have a survival advantage (Biggeri et al. 2011). In the case of Turin the big majority of immigrants were manual laborers with low education seeking a job in the car factories of the city. Therefore, the massive migratory flow could have made the male population more homogeneous in terms of frailty and caused a reduction of the male educational gradient. Selection among instead, instead, might have been less pronounced because they were likely to be passive actors in the migratory decision. Many of them often followed the husband only after some time.

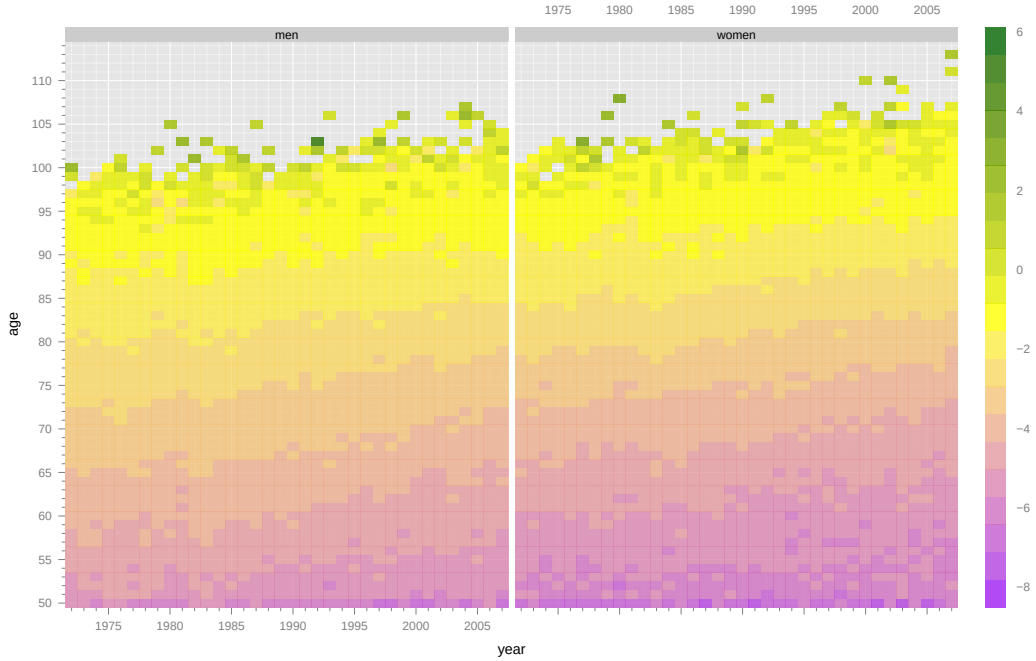


Figure 4.7: Death rates, on logarithmic scale, by 1 year age group and 1 calendar year interval.

4.3.4 Mortality Improvement over time

When dealing with a long follow up, such as the 36 year longitudinal observation available in this study, it is necessary to take into account the mortality improvement that takes place over the period. Medical and technological advances make it possible to improve survival conditions. New health related knowledge can spread and more and more individuals can have access to it as well. These and other elements make mortality decline over time and a cohort, that lives through a long period, is affected by this continuous change.

Figure 4.7 reports the mortality surface for the years of the follow up and the age of the population. A mortality surface is a very common and useful tool in demography. Here it shows the death rates, on logarithmic scale, by 1 year age group and 1 calendar year interval. On the y-axis there are

the different ages and on the x-axis the different years. A vertical segment represents the population age-specific death rates for a certain period. A horizontal segment represents the year-specific death rate for a certain age. The different birth cohorts are represented along the diagonals.

Period mortality improvements are clearly visible. Year by year, the same level of mortality shifts up to higher and higher ages. This means that a younger cohort experiences a level of mortality at age x that is lower than the level of mortality that on older cohort experienced at the same age x. Alternatively, the level of mortality experienced by an older cohort at age x is experienced by a younger cohort at an older age.

It is also possible to see an acceleration of the improvements from the late 80's and beginning of the 90's. A model that describes the age trajectory of mortality of different cohorts has to consider that the cohorts live through different periods.

Demographic Information
sex
date of birth
date of exit from Turin
cause of exit (death or emigration)
Macro-regions of birth
North-East
North-West
Central Italy
South and Islands
Abroad
Education categories
Not available
No degree (cannot read nor write)
No degree (but can read and/or write)
Primary education (italian *Scuola elementare*)
Lower Secondary education (italian *Scuola media inferiore*)
Upper Secondary education (italian *Scuola superiore*)
University Degree
begin of follow up: 24-10-1971
end of follow up: 31-07-2007

Table 4.2: Available variables from the Turin Longitudinal Study used for the analysis. When the variable is categorical its classification is given.

	Tot	**Men**	**Women**
Population of the follow-up from age 50	847,386	391,170	456,216
Cause of exit			
death	321,729	160,244	161,485
right censoring	525,657	230,926	294731
due to emigration	149,870	66,079	83,791
due to end of follow-up	375,787	164,847	210,940
Unknown/Not available			
macro-region of birth	0	0	0
education level	3,538	1,703	1,835

Table 4.3: Number of individuals, number of deaths, number of right censored cases and missing data of the follow up from age 50.

	Men	(%)	Women	(%)
Education				
Low	218,098	(55.75)	299,400	(65.62)
Medium	95972	(24.53)	100,926	(22.12)
High	77100	(19.71)	55890	(12.25)
Tot	391,170	(100)	456,216	(100)
Macro-Region of birth				
North-West	192,660	(49.25)	244,613	(53.61)
North-East	39,189	(10.02)	49,680	(10.89)
Center	13,198	(3.37)	19,900	(3.05)
South & Islands	128,461	(32.84)	126,041	(27.63)
Abroad	17,662	(4.51)	21,982	(4.82)

Table 4.4: Distribution of the Turin population included in the follow-up by education and macro-region of birth.

Chapter 5

Methods

5.1 Frailty Models in Survival Analysis

The aim of the study is to investigate the presence of selection processes in the educational mortality patterns of Turin. Are the estimates of the differentials affected when taking into account unobserved heterogeneity? Can the observed mortality pattern be explained by the framework of the frailty models?

The availability of individual longitudinal histories data in this study makes it possible to use survival analysis techniques.

Survival analysis is a corpus of techniques for the analysis of duration dependence processes (time until the event of interest occurs), like the age at death. The application of frailty models to this methodology is aimed to take into account the unobserved frailty component in the estimation procedures via the introduction of a random effect. A frailty model is, namely, a multiplicative hazard model that includes a frailty term (for the unobserved component), a baseline hazard function common to all the individuals (this can be parametric or non parametric) and fixed effects terms for the influence of observed covariates on the baseline hazard. For a complete and up to date treatise on the topic of frailty models in survival analysis, please see Wienke (2010).

When the investigator has no clues about the phenomenon under study (neither empirical evidence nor theoretical or biological reasons to hypothesize a specific pattern for the risk function), it is appropriate to use non parametric models that don't put any functional constraints on the hazard, like the Cox model. However, the wide demographic research conducted so far has shown that human adult mortality can be accurately described by a Gompertz function (Gompertz 1825) or by some Gompertz-like variants, like the Makeham function. This helps in the choice of the model to use for the analysis because, in the case of human mortality, the age pattern of the phenomenon is known. Therefore in this analysis, parametric models with Gompertz and Gompertz-Makeham baselines will be used.

First, the estimation will be performed without taking into account the unobserved frailty. Then, this component will be introduced in the model and the estimates obtained with the two approaches will be compared. The computations are realized with the free statistical software R (R Development Core Team 2011).

5.2 Survival analysis

Before moving onto the detailed description of the estimation strategy a brief summary of the basic concepts and functions of survival analysis will be given.

Survival analysis deals with time to event data. It models the time until the event of interest occurs and its application spans from social, economic and medical science to physics, biology, engineering and natural sciences. Many types of events can be studied. For example, age at death, age at which young adults leave the parental home, age at childbirth, time at re-entry in the labor market of unemployed individuals, failure time of a technical unit of a device in reliability engineering and any other process that involves a duration the investigator is interested in.

5.2.1 Basic concepts

Several features distinguish survival analysis from other fields of statistics and make a separate theory for it necessary. First of all, in survival analysis the response variable is a combination of continuous data (the observed duration) and discrete data (the event indicator that tells whether and when the event has occurred or the individual is censored at the end of the observation).

Moreover, inference is conditional on the information available at the time units of the analysis. Such information changes over time because the population at risk changes at every time unit and the estimates are based on the risk set that has survived (that has not experienced the event yet) up to that moment.

Survival analysis data are also peculiar because they are affected by censoring and truncation. Censored observations are incomplete observations containing only partial information about the event time and the occurrence of the event. This happens when the subject exits the study when he has not experienced the event yet and this can be due to several factors, like loss of follow up, termination of the study and observation period, drop-out and so on. Truncation, instead, refers to the situation where the subject enters the study at some delayed time in respect to the beginning of the process. In the case of a mortality follow up, for example, unless we are able to set up a study that follows a cohort of individuals from the same age, very often we are in a situation when the data are left-truncated. At the start of the observation, in fact, the population comprises many individuals who belong to different cohorts and are followed from different ages.

Several studies have shown how ignoring censoring and truncation leads to biased estimation procedure (Howards et al. 2007; Cain et al. 2011). Therefore it is very important to incorporate them in the model. According to what kind of censoring and truncation affect the data, particular adjustments of the likelihood function in the estimation procedure are needed (Klein and Moeschberger 2003: chap 3).

5.2.2 Basic functions

Given the time to event, that is a nonnegative random variable $T \geq 0$, often assumed to be continuous, the basic survival analysis functions are:

cumulative distribution function: $F(t) = P(T \leq t) = \int_0^t f(x)dx$

probability density function: $f(t) = F'(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \geq t+\Delta t)}{\Delta t}$

survival function: $S(t) = 1 - F(t) = P(t \geq t) = \int_t^\infty f(x)dx$

hazard function: $\mu(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t | T \geq t)}{\Delta t}$

cumulative hazard function: $M(t) = \int_0^t \mu(x)dx$

Thanks to relationships between the basic functions, when one of them is known it is possible to obtain all the others:

$f(t) = -S'(t)$

$\mu(t) = \frac{f(t)}{S(t)}$

$S(t) = e^{-\int_0^t \mu(x)dx} = e^{-M(t)}$

$\mu(t) = -\frac{d}{dt} \ln S(t)$

$M(t) = -\ln S(t)$

The main goal of survival analysis is to estimate an appropriate model that describes the individual experience of the event of interest by one of the basic functions and to measure the effect of observed covariates on this function. Usually the hazard, the cumulative hazard or the survival function are the functions of interest.

The model can be parametric, semi-parametric or non-parametric. In the first case, the shape of the baseline hazard is assumed to follow a certain distribution, for example weibull, gamma or exponential. In the second case some mild assumptions about the baseline hazard are done, as, for example, in the piecewise constant hazard model, where the time axis is split into time

periods and the hazard is assumed to be constant in each of these periods. Finally, in the case of a fully non-parametric model, no assumption is made about the shape of the baseline hazard and only the effect of the covariates is measured, as, for example, in the Cox model.

The estimation is performed with the maximum likelihood method, where, based on the observed data, the vector of unknown parameters regarding the baseline hazard and the covariates coefficients is estimated.

5.3 Estimations strategy

The analysis will use parametric survival analysis models.

The mortality differentials will be estimated without controlling for the unobserved heterogeneity of frailty (standard survival analysis approach) and then controlling for it, introducing frailty into the model. The estimates obtained with the two approaches will be compared. The comparison of the different non nested models (with different baseline hazards and with and without the frailty component) will be based on the Akaike Information Criterion (AIC) (Akaike 1974).

The data are both right censored and left truncated. Right censoring is due to survival up to the end of the follow up (July 31 2007) or emigration from the city of Turin at some earlier time. Left truncation, instead, derives from the fact that individuals who were older than 50 at the date of the 1971 census enter the study at different ages. Therefore, their contribution to the process must be considered conditional to their survivorship up to the age at entry (for a graphical representation of the data, please see figure 4.1).

The follow up period is long (36 years) and the dataset includes many different cohorts, each passing through the 36 years of observation at different ages. During the follow up period, from 1971 to 2007, a significant mortality improvement occurred (please see figure 4.7) and younger cohorts experience lower age specific mortality than older cohorts (at the same age). To take into account this important factor I used two strategies.

One strategy consists in considering the cohort to which the individual belongs as covariate. In the model with the cohort variable, controlling for unobserved heterogeneity is implemented within the univariate frailty model framework.

The second strategy is considering the improvement from a period perspective and splitting the time into several calendar year periods. To model this, the life spell of an individual must be split into several spells, one for each period the individual passes through. This implies an organization of the data into clusters, where each cluster represents one individual and contains its period spells. In the model with period covariates, controlling for unobserved heterogeneity cannot be achieved within the framework of the univariate frailty models. The multivariate frailty framework is needed to implement a shared frailty model, where the spells in each cluster pertain to the same individual and share the same hidden frailty.

5.3.1 Standard approach without frailty

The standard survival analysis approach does not include unobserved heterogeneity of frailty in the model. The only variability controlled for is that explained by the observed covariates u included in the model. Their effect on the baseline hazard μ_0 is estimated, as showed in equation 5.1.

$$\mu(x|u) = \mu_0(x)e^{u\beta} \tag{5.1}$$

The estimation is based on the maximum likelihood method. Equation 5.2 shows the the likelihood function for the model without frailty, in the case of right censored and left truncated data

$$L(\beta, \theta) = \prod_{i=1}^{n} \frac{(\mu(x_i, \theta)e^{u_i\beta})^{\delta_i} S(x_i, \theta)^{e^{u_i\beta}}}{S(y_i, \theta)^{e^{u_i\beta}}} \tag{5.2}$$

where for each individual i, y_i is the entry time, x_i is the exit time, δ_i is the status (1=dead, 0=right censored), u_i is the covariate profile whose effect β is estimated and $\mu(\cdot)$ denotes the hazard, $S(\cdot)$ denotes the survival function and θ is the vector of parameters of the baseline hazard.

5.3.2 Introducing frailty

While the standard approach "is based on the assumption that the study population is homogeneous up to some observed covariates [...] it is a basic observation that individuals differ greatly, for example, with respect to the effects of a drug, a treatment, or the influence of various explanatory variables." (Wienke 2010: 55) No matter how many covariates can be observed, there will always be a component of individual unobserved heterogeneity that influences the individual risk of experiencing the event under study.

This component is unobserved either because it is not available in the data or because it is difficult to measure. For example, in mortality studies, genetic or lifestyle factors can influence the risk of an individual dying; in other research fields, for example in the analysis of marriage dynamics, systems of values and opinions (when they are not observed and reported in the data) can influence the risk of marrying rather than cohabiting and so on.

Neglecting the presence of this component leads to biased estimates of both hazard rate and regression coefficients, because the time dependent hazard rate results from changes in the composition of the population under study over time with respect to the covariates in the analysis. For more details please see paragraph 3.3.

Univariate frailty

In the univariate case of independent and identically distributed survival data of unrelated individuals, an individual random effect z for the frailty is introduced in the model as a multiplicative term on the baseline hazard

$$\mu(x|u,z) = z\mu_0(x)e^{u\beta} \tag{5.3}$$

As shown in the equations 3.8 and 3.9, under the assumption of a gamma distributed frailty, the random term can be integrated out, making the model in equation 5.3 identifiable.

Equation 5.4 shows the likelihood function for the univariate gamma

frailty model with covariates, in the case of right censored and left truncated data

$$L(\beta, \theta, \sigma^2) = \prod_{i=1}^{n} \frac{\left(\frac{\mu(x_i,\theta)e^{u_i\beta}}{1+\sigma^2 M(x_i,\theta)e^{u_i\beta}}\right)^{\delta_i} (1+\sigma^2 M(x_i,\theta)e^{u_i\beta})^{-\frac{1}{\sigma^2}}}{(1+\sigma^2 M(y_i,\theta)e^{u_i\beta})^{-\frac{1}{\sigma^2}}} \tag{5.4}$$

where for each individual i, y_i is the entry time, x_i is the exit time, δ_i is the status (1=dead, 0=right censored), u_i is the covariate profile whose effect β is estimated and $\mu(\cdot)$ denotes the hazard, $M(\cdot)$ denotes the cumulative hazard, θ is the vector of parameters of the baseline hazard and σ^2 is the variance of the frailty distribution at the initial age of the study (in this case age 50).

Shared frailty

The shared frailty model is a model of the multivariate frailty models class. This class of models is used with multivariate survival data, where independence between event times within clusters cannot be assumed. In the case of repeated spells for the same individual, the shared frailty model assumes that those spells share the same hidden individual frailty, as showed in equation 5.5. Conditional on the frailty z_i, survival times in cluster i are independent and their hazard is

$$\mu(x|u_{ij}, z_i) = z_i \mu_0(x) e^{u_{ij}\beta} \tag{5.5}$$

where u_{ij} is the covariate profile of the j-th observation in the i-th cluster.

In the estimation of a parametric shared frailty model with left truncated data the likelihood function has to "properly take into account the interplay between truncation and dynamic selection." (Van den Berg and Drepper 2011: 1) The authors give the likelihood function for shared gamma frailty for the cluster i, showed in equation 5.6, that properly accounts for dynamic selection before the truncation points [1]

[1]Please note that Van den Berg and Drepper (2011) propose a corrected version of the likelihood function for shared frailty models with left truncation implemented in the

$$L_i = \left[\prod_{j=1}^{n_i} \left(\mu(x_{ij},\theta)e^{u_{ij}\beta}\right)^{\delta_{ij}}\right] \frac{\Gamma(\frac{1}{\sigma^2}+D_i)}{\Gamma(\frac{1}{\sigma^2})}(\sigma^2)^{D_i}\left(1-\sigma^2\sum_{j=1}^{n_i}\ln(S_{ij}(y_{ij},\theta)^{e^{u_{ij}\beta}})\right)^{\frac{1}{\sigma^2}} \left(1-\sigma^2\sum_{j=1}^{n_i}\ln(S_{ij}(x_{ij},\theta)^{e^{u_{ij}\beta}})\right)^{-\frac{1}{\sigma^2}-D_i} \quad (5.6)$$

where for each j-th individual in the i-th cluster, $y_{i,j}$ is the entry time for, $x_{i,j}$ is the exit time, $\delta_{i,j}$ is the status (1=dead, 0=right censored), u_{ij} is the covariate profile whose effect β is estimated, $\mu(\cdot)$ denotes the hazard, $S(\cdot)$ denotes the survival function, θ is the vector of parameters of the baseline hazard, σ^2 is the variance of the frailty distribution and $D_i = \sum_{j=1}^{n_i}\delta_{ij}$.

The overall likelihood function is simply

$$L(\beta,\theta,\sigma^2) = \prod_{i=1}^{n} L_i \quad (5.7)$$

5.3.3 Baseline hazards

The research conducted so far has shown that human adult mortality is accurately described by a Gompertz function (Gompertz 1825) or by some Gompertz-like variants, like the Makeham function (Makeham 1860). Therefore, in this study the Gompertz model and Gompertz-Makeham model will be used as baseline hazard.

Gompertz According to the Gompertz model Gompertz (1825), at any age x, the force of mortality $\mu(x)$ is expressed by an initial mortality level, a, and by an exponential mortality increase by age, the parameter b, which is also defined as the rate of aging (Shock 1967; Finch 1994) and represents the slope of the mortality curve on a logarithmic scale

$$\mu(x) = ae^{bx} \quad (5.8)$$

statistical software Stata (Stata Release 12 2012: 389), used by many of the studies on frailty models with left truncated data. The authors show that, in case of left truncation, this likelihood does not properly take into account the dynamic selection before truncation.

Gompertz-Makeham The Gompertz-Makeham model differs from the previous model by introducing in the equation an additive constant c

$$\mu(x) = ae^{bx} + c \tag{5.9}$$

This model was suggested by Makeham (1860) to achieve a better fit to real survival data at adult ages but the Makehan term is not only a useful mathematical tool. Golubev (2004) points out how this term "represents the rate of deaths resulting from causes that no organism can resist irrespective of its age," thus having a fundamental biological meaning. In other words, the Makeham term helps to capture the external mortality that is not related to the age-senescent process.

Exponential A model with an exponential baseline is used as a first step of the analysis to explore the data. The observed convergence of hazards by education at old ages has been modeled by an exponential model with the interaction between age and education, considering the age as covariate and not as part of the baseline hazard. The exponential hazard, shown in equation 5.10, allows this because it is a constant hazard model, independent of the age.

$$\mu(x) = \lambda \tag{5.10}$$

5.4 Models

5.4.1 Exponential model with age as covariate

The fact that the age is not contained in the exponential baseline risk function allows one to consider the age as covariate. The aim of this step is to investigate whether there is convergence of hazards by education groups at old ages by testing whether there is a significant interaction between the variables education and age.

The single-parameter baseline hazard is modulated over two broad age groups that, in addition to the education level, act as covariates. Equations 5.11 and 5.12 describe the hazard and survival functions of the exponential model with covariates. The identity between an exponential hazard modulated by an age covariate (whose effect is exponentiated) and the Gompertz model, where the age is an exponential term embedded in the baseline hazard, makes this model appropriate for human mortality data that, it is well-known, follow a Gompertz-like pattern.

$$\mu(x) = \lambda e^{\beta_{cov}} \tag{5.11}$$

$$S(x) = (e^{-\lambda x})^{e^{\beta_{cov}}} \tag{5.12}$$

The age span is divided into two age variables: "ages before 80" and "ages after 80". In figure 4.5 we have seen how mortality deceleration, considered to be due of selection, starts to be visible approximately from age 80 on. Moreover, selection needs time to take place, and manifests itself visibly at later ages.

Education is divided into three groups: low, medium and high education.

In addition to age and education, the model also controls for period effects, introducing a variable for "years before the 90's" and "years after the 90's". Looking at figure 4.7 this seems reasonable because it is possible to identify a general cut point in the trend of mortality improvement, approximately around the beginning of the 90's. Different ages might, indeed, present different improvement trends over the years. However, as the focus is on two broad age groups and not on a finer age classification, it is reasonable to think that the same general trends apply similarly to the two broad age ranges.

5.4.2 Age-dependent hazards models

Human adult mortality risk is known to increase with age. Therefore, when analyzing the mortality patterns, the baseline hazard must be an increasing

function of age.

Gompertz (1825) indicated that the increase with age is approximately exponential. Makeham (1860) complemented the exponential hazard by adding an age-invariant term that accounts for constant external mortality at all ages (Golubev 2004).

Equations 5.13, 5.14, 5.15, 5.16 show the hazard and survival functions of the Gompertz model and the Gompertz-Makeham model with covariates. These include the education level and the macro-region of birth, plus the the cohort indicator or the period indicator, depending on the strategy used for modeling the mortality improvement.

Gompertz model with covariates

$$\mu(x) = ae^{bx}e^{\beta_{cov}} \tag{5.13}$$

$$S(x) = \left(e^{\frac{a}{b}(e^{bx}-1)}\right)^{e^{\beta_{cov}}} \tag{5.14}$$

Gompertz-Makeham model with covariates

$$\mu(x) = (ae^{bx} + c)e^{\beta_{cov}} \tag{5.15}$$

$$S(x) = \left(e^{\frac{a}{b}(e^{bx}-1)+cx}\right)^{e^{\beta_{cov}}} \tag{5.16}$$

5.4.3 Modeling survival improvement

Survival conditions improve over time. In a long follow up, like the one available in this study, this factor must be taken into account. Different cohorts pass through the 36 years of follow up at different ages. Younger cohorts experience, at a certain age, a level of mortality that is lower than the mortality experienced by the many years older cohorts at the same age. The mortality risk of a 50 year old in 1971 is very different from the risk of a 50 year old in 2000.

As already explained in section 5.3, to address this question I used two different strategies. The first strategy consists in introducing into the model a cohort variable for each individual. The second strategy is creating calendar period variables and splitting the individual survival times into spells, one for each calendar period the individual is passing through. The estimation of the models with frailty in the two cases are different. In the case of the cohort variable a univariate frailty model is used. In the case of period variables a multivariate frailty model (in particular, shared frailty) is necessary.

Double checking of cohort and period effects is crucial also in the evaluation of the educational differences over age. Lauderdale (2001); Lynch (2003) have recently demonstrated how the dynamic relationship between education and health over age can be misspecified when either effect is ignored.

Model with cohort variable

A cohort variable has been introduced in the model. This variable is based on the individual's age at entry in the follow up, thus identifying one year birth cohort. This variable starts with 0, for the youngest birth cohort in the follow up (individuals aged 14 in 1971 who turned 50 at the very end of the follow up), and increases one by one for older and older cohorts. Its coefficient multiplied by the n index represents how much higher is the mortality of the n-th cohort in respect to the youngest cohort in the data.

The introduction of a single continuous cohort variable is justified by figure 5.1. This figure shows that the log-mortality rates for the same age in different years (that is, mortality at the same age for different cohorts) have decreased in a roughly linear fashion. As the likelihood functions presented in section 5.3 show, the effect of the covariate is introduced in the model with an exponential term that, on logarithmic scale, corresponds to a linear effect.

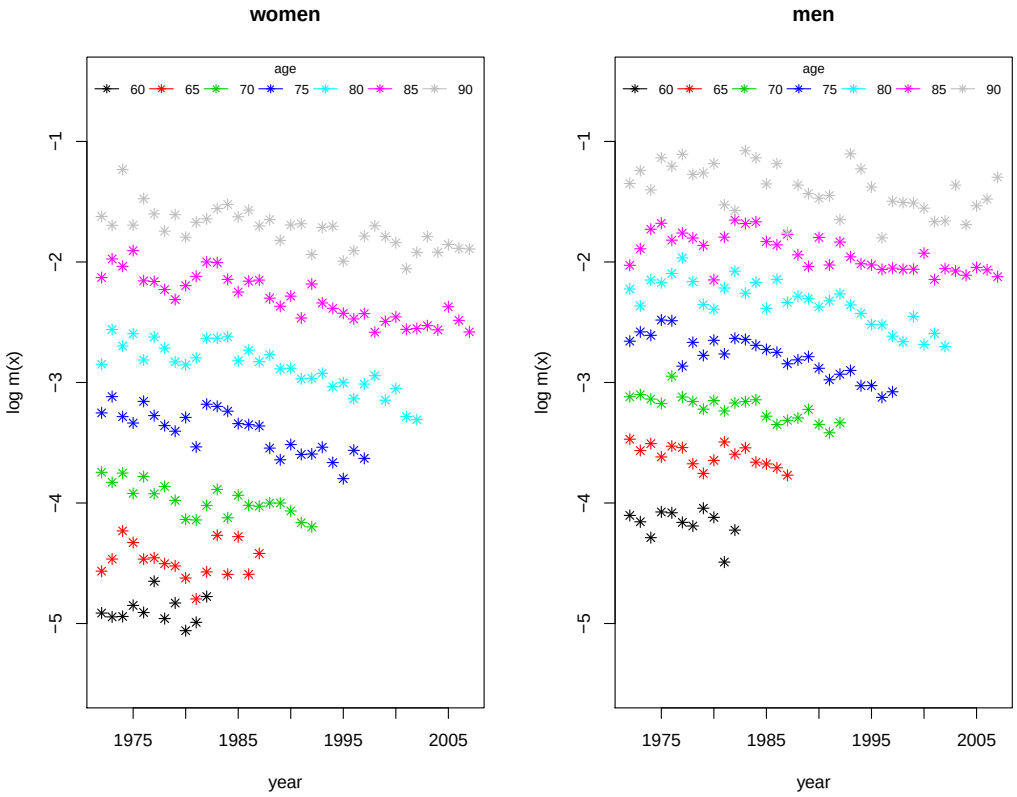


Figure 5.1: Female and male mortality rates (on logarithmic scale) for selected ages from 1971 to 2007.

Model with period variables

The second strategy is introducing period covariates for the calendar year. Mortality improvement is known to be mainly a period phenomenon and the mortality experience of an individual is determined by both his age and the calendar period that he is passing through at that age.

This is modeled by introducing time varying covariates for the periods and estimating as many coefficients as the number of periods considered, each of them affecting the portion of mortality curve that is passing through it.

The advantage of this strategy, compared to the previous one, is that

Figure 5.2: Cohort mortality age trajectory. While aging, the cohort passes through different periods (defined by the vertical lines). In each of these periods some mortality improvement occurs. This drags down the mortality curve (red line), making it differ from the curve that the cohort would have had if there had been no improvement (black line). The gray dashed line represents the standard individual mortality that, contrary to the cohort mortality, does not bend over because it is not affected by selection effects.

the "pulling down" effect of the mortality improvement can be disentangled from the cohort mortality age-trajectory, as illustrated in figure 5.2. This may help to reduce the downward bias in the estimate of the cohort rate of aging (the b coefficient of the baseline). Moreover, capturing the component of the mortality deceleration due to improvement in mortality helps to obtain a more accurate estimate of the unobserved frailty component.

The period variables are time varying covariates. Time varying covariates require episode splitting and the creation of multiple survival spells for each

individual. Episode splitting is necessary in order to allow the period variable to change according to which period the individual is passing through. The time clock is measured by the individual age at the start of each spell, that represents the left truncation time, and age at the end of each spell. At the end of the spell the individual can be right censored (status equal to 0) or can die (status equal to 1). A simplified illustrative version of the dataset is reported in table 5.4.3.

To introduce frailty in the model one has to use the shared frailty model with left truncated data. Each individual represents a cluster and the spells within the cluster share the same hidden frailty. Shared frailty models, particularly with left truncated data, are computationally heavy.

Initially the time from 1971 to 2007 had been divided into 12 periods of 3 years each. This means that the individual had up to 12 spells and the final size of the split dataset was several million lines. The computational heaviness of the model made it impossible to estimate and the number of spells had to be reduced to two, one for the period 1971-1990 and one for the period 1991-2007. The choice of the cut point is justified because it was possible to identify a change in the trend of mortality improvement, around the beginning of the 90's (see figure 4.7).

The optimization routine applied to the whole data set did not converge in more than four weeks so random subsampling has been applied, estimating the model parameters repeatedly over smaller subsets of the data randomly drawn without replacement (Hartigan 1969, 1975; Politis and Romano 1994). As pointed out by Efron (1979: 24), random subsampling "is very similar to bootstrap." The basic idea is to approximate the parameter estimates based on the distribution of the repeated estimates. 1% samples have been drawn using stratified sampling in order to have a representative sample of the total population, each stratum representing a specific combination of education and macro-region of birth, out of which individuals were randomly selected.

A major issue in bootstrapping is the number of number of repetitions able to ensure reliable estimates. It is generally advised to choose a suffi-

id	spell	status	agestart	ageend	ageclass	spellstart	spellend	1971-1980	1981-1990	1991-2000	2001-2007
1	1	0	50.00	52.36	1	1978-08-22	1981-01-01	1	0	0	0
1	2	0	52.36	52.94	1	1981-01-01	1981-08-01	0	1	0	0
2	1	0	50.00	52.75	1	1978-04-02	1981-01-01	1	0	0	0
2	2	0	52.75	62.75	1	1981-01-01	1991-01-01	0	1	0	0
2	3	0	62.75	65.00	1	1991-01-01	1993-04-02	0	0	1	0
2	4	0	65.00	72.75	2	1993-04-02	2001-01-01	0	0	1	0
2	5	0	72.75	79.33	2	2001-01-01	2007-07-31	0	0	0	1
3	1	0	55.73	64.92	1	1971-10-24	1981-01-01	1	0	0	0
3	2	0	64.92	65.00	1	1981-01-01	1981-01-31	0	1	0	0
3	3	0	65.00	74.92	2	1981-01-31	1991-01-01	0	1	0	0
3	4	0	74.92	80.00	2	1991-01-01	1996-02-01	0	0	1	0
3	6	1	84.92	85.81	3	2001-01-01	2001-11-23	0	0	0	1

Table 5.1: Example of data set with time varying covariates and episode splitting. At each point that the time varying covariates change value, a spell is created for the individual. In this case, the two time varying covariates are *calendar period* and *ageclass*. The follow up period is split into four spells for which dummy variables are created, 1971-1980, 1981-1990, 1991-2000 and 2001-2007. The individual's profile for these dummy variables is 1 for the period the individual is passing through and 0 for all the others. The age of the individuals is divided into 3 age groups, 50-65, 65-80, 80+ and every time the individual moves from one groups to the other the variable *ageclass* changes its value taking the values 1, 2 or 3. The status variable indicates whether an individual is right censored (the status of the last spell is 0) or exits the study because of death (the status of the last spell is 1).

ciently large number of repetitions (Hesterberg 2011). This does not represent a problem if the procedure is applied to simple statistical measures but it can be very hard to apply to computationally and theoretically more complex models. Efron and Tibshirani (1993) suggest that 500 repetitions are generally sufficient in most cases, while the analysis in Manly (1997) concludes that even 200 can provide relatively small error margins. Pattengale et al. (2010) have found that, typically, between 100 and 500 replications are necessary in phylogenetic bootstrapping, a technique used to make inference about phylogenetic trees, that requires the application of very complex estimators.

Given the complexity of the model and the large size of the data set, 250 estimate repetitions have been performed. By ways of illustration, one repetition for the Gamma-Gompertz model with education and macro-region of birth covariates applied to the 1% sample drawn from the men dataset requires between 12 and 13 hours on a 4CPUs - 4 GB computer.

Chapter 6

Results and discussion

6.1 Exponential model

The exponential model described by equations 5.11 and 5.12 has been estimated with covariates for education, age group and their interaction. Table 6.1 shows the mortality rate ratios by education level at the age groups 50-80 years and 80+ years. The model also controls for a period factor: period 1971-1991 and period 1991-2007. For the sake of simplicity table 6.1 does not report the estimated coefficient for the period and the λ parameter of the exponential hazard. They are shown in table A.1 in the Appendix.

The results confirm the well known mortality gradient in the literature: the risk of dying is inversely proportional to the education level. It is lower for highly educated individuals and increases from high to low educated groups. However, the relative difference respect to the high education reduces at older ages and this is true for both of the differences between medium and low education. Among men, between the ages 50 to 80, the relative risk of dying for the groups with low and medium education are 57% and 23% higher than the risk for the high education group. At ages older than 80 the difference is reduced to "only" an 18% higher risk for those with low education and 8% for those with medium education. The reduction is more pronounced among women: between the ages 50 to 80 the rate ratios to the

highly educated women are 1.59 for women with low education and 1.25 for women with medium education, while, after age 80 they reduce to 1.17 and 1.04. Moreover, the difference between the group with medium education and the group with high education becomes scarcely significant.

This model has then been compared with a model that does not include the age-education interaction term (for the results of the model without the interaction term please see table A.1 in Appendix) with a likelihood ratio test. The comparison between the two nested models shows that the interaction term is significant, indicating that the educational gradient changes significantly over age for both of the sexes.

6.2 Frailty modeling

The estimates in table 6.1 show that there is a significant reduction of the relative difference in mortality by education at old ages. However, the results could be biased if we do not consider that the observed mortality patterns for the educational subpopulations may differ from the mortality dynamics experienced at the individual level, that a selective process causes a compositional change in the population as it grows older and that this process, if not controlled for, might lead to biased estimate of the coefficients. In other words, unobserved individual characteristics might be playing a role in determining a differential selection that leads to converging population hazards by education.

Educational differences in mortality have been analyzed first estimating survival models that don't control for the component of hidden individual frailty and then including this component in the model.

Besides education level, also a variable for the macro-region of birth has been included in the analysis. Given the strong economic, social and demographic geographical differences in Italy, it is important to consider this factor. Moreover, the city of Turin is characterized by a peculiar history. After WWII, massive migration flows, especially from the southern regions,

Women				
	50-80 years		80+ years	
	estimate	95% CI	estimate	95% CI
high education	1.000	-	1.000	-
medium education	1.250	(1.213-1.289)	1.040	(1.008-1.073)
low education	1.594	(1.552-1.637)	1.170	(1.139-1.202)
Likelihood ratio test with reduced model				
D statistics: 319.833		df: 2		p-value: 0.00000
Men				
	50-80 years		80+ years	
	estimate	95% CI	estimate	95% CI
high education	1.000	-	1.000	-
medium education	1.234	(1.209-1.259)	1.082	(1.048-1.116)
low education	1.571	(1.544-1.598)	1.172	(1.143-1.202)
Likelihood ratio test with reduced model				
D statistics: 395.193		df: 2		p-value: 0.00000

Table 6.1: Mortality rate ratios between education groups at two different age groups estimated from an exponential survival hazard model with covariates education, age and their interaction. The table reports also the result of the likelihood ratio test between this model and a model without the interaction term.

were directed to Turin, center of the Italian car industry. As we have seen in chapter 4, the proportion of native individuals (natives are considered those born in the North-west region) has constantly decreased cohort by cohort. Among the resident Turin population born after 1937, about half of the individuals were born in one of the southern regions.

Education level is divided into 3 broad categories: low, medium and high education. The first category includes individuals with elementary or lower education; the second category represents individuals with lower secondary degree (the Italian *scuola media*, corresponding to 6th, 7th and 8th grades);

the third category includes individuals with high school or higher education.

The macro-region of birth is divided in 5 categories: North-West (the region that contains the city of Turin), North-East, Center, South & Islands and Abroad.

6.2.1 Frailty modeling with cohort covariate

Survival conditions improve over time and in a long follow up this factor must be taken into account. Different cohorts pass through the 36 years of follow up at different ages. Younger cohorts experience lower age specific mortality levels than those experienced by the older cohorts. The mortality risk of a 50 year old in 1971 is very different from the risk of a 50 year old in 2000.

This section presents the results of the analysis conducted introducing a cohort variable in the model based on the individual's age at entry in the follow up, thus identifying one year birth cohorts. This variable has been coded 0 for the youngest birth cohort and increases linearly one by one for each of the older cohorts. Its coefficient, multiplied by the n, represents how much higher the mortality of the n-th cohort is in respect to the youngest cohort in the data, represented by those born between 1956-1957.

Model selection

Before examining the estimation of the educational differentials it is necessary to investigate which is the most appropriate baseline hazard. The Gompertz function is known to adequately fit the human mortality trajectory at adult and old ages. However, more and more recent studies also highlight the importance of the Gompertz-Makeham model (for the sake of brevity, from now on it will simply be called Makeham). Beyond the parameters a and b of the Gompertz equation, which represent respectively the initial mortality level and the rate of mortality increase with age, the Makeham model adds a third additive component to describe adult mortality. The parameter c of the Makeham equation is used to achieve a better fit of survival data from adult ages on and represents the component of force of mortality that is not related

	Gompertz ae^{bx}	Gamma-Gompertz $\frac{ae^{bx}}{1+\sigma^2\frac{a}{b}(e^{bx}-1)}$	Makeham $ae^{bx}+c$	Gamma-Makeham $\frac{ae^{bx}+c}{1+\sigma^2\frac{a}{b}(e^{bx}-1)+cx}$
AIC Wom	1327474	1327476	1326878	1326695
AIC Men	1303693	1303655	1303695	1303693

Table 6.2: Model selection of 4 different hazard models based on the Akaike Information Criterion.

to the senescent process, also called external mortality. The equations of the models are shown in table 6.2. For more details about the models please see chapter 5, section 5.3.3.

Table 6.2 shows the comparison, via Akaike Information Criterion (AIC), of the models with Gompertz baseline and Makeham baseline hazard, without and with controlling for unobserved heterogeneity of frailty. These last models are called Gamma-Gompertz and Gamma-Makeham. Their name includes the distribution function assumed for the unobserved heterogeneity component. The comparison is based on the AIC because they are not nested models.

Based on the AIC (the lower the AIC the better the fit), the female data are better fitted by the Makeham baseline while the male data are by the Gompertz baseline. Moreover, the comparison also indicates that in both cases, the model that includes unobserved heterogeneity of frailty performs better than the model that does not control for it. In the case of women the improvement in the fit is more significant than in the case of men. This indicate that the frailty component is somehow less important for men than for women. It could also indicate that men are less heterogeneous than women in the population of Turin.

Figures 6.1 and 6.2 plot the observed mortality log-hazard for different birth-cohorts and the log-hazards predicted by the four models in table 6.2.

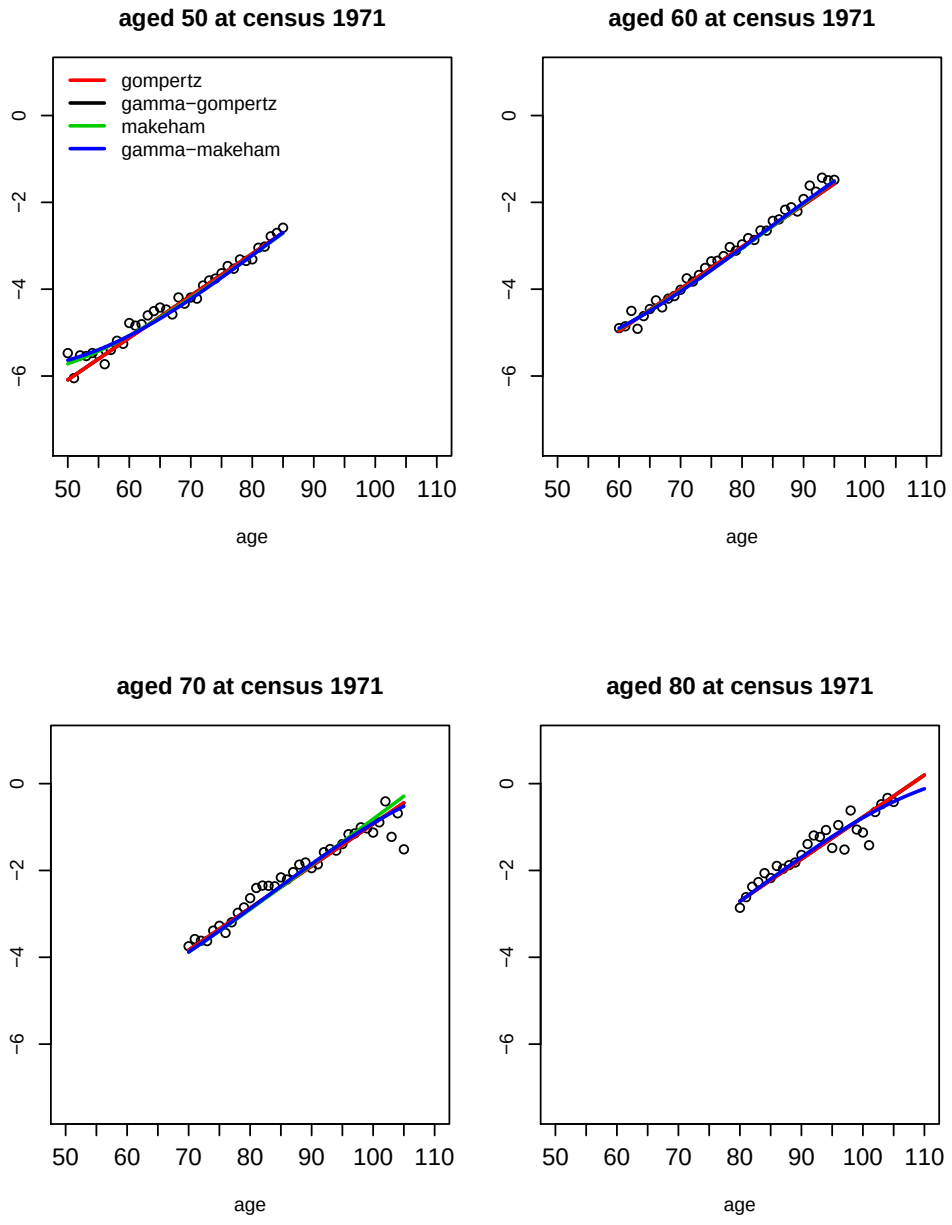


Figure 6.1: Women - Observed and predicted log mortality for different cohorts at the 1971 census and different models.

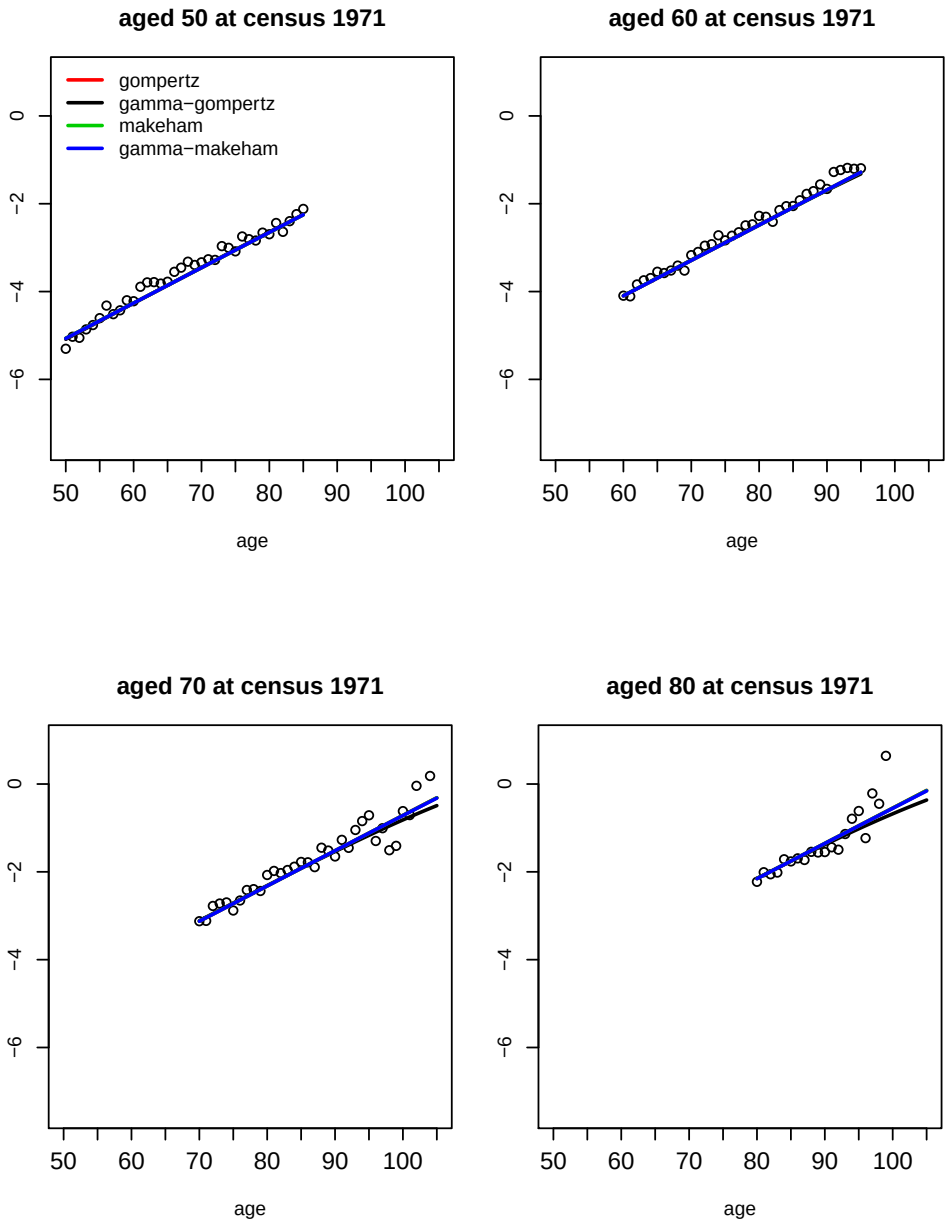


Figure 6.2: Men - Observed and predicted log mortality for different cohorts at the 1971 census and different models.

Estimation of the differentials

As indicated by the model comparison, the estimation of the differential mortality by education is performed using the Makeham baseline hazard for the women and the Gompertz baseline hazard for the men.

The first two models were estimated without unobserved frailty. Model 1 controlled only for the education level. Model 2 added also the macro region of birth. Finally, model 3 introduced the unobserved heterogeneity of frailty. All the models contained the cohort variable to take into account the cohort improvement in survival.

Tables 6.3 and 6.4 report the results of the three models: the parameters of the baseline hazards, the variance of frailty in the population (for model 3 only), the cohort variable and the rate ratios of the mortality differentials (by education and by macro-region of birth).

The estimate for the cohort variable is stable across the three models. It undergoes a minor change only for women when the unobserved heterogeneity is included. This variable has been set equal to 0 for the youngest one year birth cohort in the data set (individuals born in 1956-1957) and increases one by one for each of the older cohorts (1955-1956, 1954-1955 and so on). Its coefficient corresponds to how much higher mortality is for a certain one year birth cohort in respect to the previous younger cohort. An alternative interpretation is that the cohort coefficient shows how much mortality has been improving from one cohort to another, thus it represents the rate of mortality improvement by cohort. If the coefficient is 0.016, this means that cohort 1 has 0.016 higher mortality than cohort 0, but it also means that from cohort 1 (older) to cohort 0 (younger) mortality has decreased at the rate of -0.016.

Controlling for selective mortality by introducing the frailty component in the model allows the model to describe the observed leveling off of the hazard at old ages, attributing it to the role of selection. Compared to a model that assumes there is no hidden heterogeneity between individuals, the estimate of the b parameter in the baseline hazard, also called rate of aging (Shock 1967;

Finch 1994), should be less affected by the biasing "dragging down" effect. This is because part of the the leveling off is controlled for by the frailty component. Consistently with this hypothesis, the estimate for b increases in model 3 for the women, from 0.106 to 0.117. For men, instead, the estimate of b is 0.081 in model 1 and increases only slightly to 0.083 in models 2 and 3.

The other parameters of the baseline hazard, a and c for women, only a for men, are stable across the three models.

The estimated variance of frailty for the male population is lower than the estimated variance of frailty for the female population: σ^2 is 0.1 for women and 0.04 for men. This is consistent with the idea that the strong migration flows from the southern regions might have especially homogenized the male population in terms of frailty and susceptibility to mortality. In chapter 4 we have seen that the majority of the native population has a higher educational level. The majority of the immigrant population, instead, has low education and comes from the southern regions, whose male population is knonw to benefit from a survival advantage (Biggeri et al. 2011). This aspect, plus a possible better health status of subjects that decide to undertake a migration trip, might have contributed to make the male population more homogeneous and to reduce susceptibility differences between lowly educated men (mostly immigrant) and highly educated men (mostly native).

Research on the role of omitted covariates in duration dependence models has highlighted how the estimates for the observed covariates might be biased if the presence of such omitted covariates is ignored (Gail et al. 1984; Lin et al. 1998; Betensky et al. 2002; Aalen 1994; Aalen et al. 2008a). This is because, acting constantly on the risk set, unobserved heterogeneity, through the selection process, biases the shape of the hazard even if the omitted source of variation is uncorrelated with the predictors included in the model (Trussell and Rodriguez 1990; Rodriguez 1994).

The comparison of the coefficients in model 2 and model 3 shows that the estimates of the educational gradient in mortality among women don't

change significantly after the introduction of the frailty term in the model. Although from model 2 to model 3 the mortality rate ratios of women with low and medium education compared to the highly educated women tend to fall, the confidence intervals of the coefficient estimated by the two models consistently overlap showing that the reduction is not significant[1].

Among men, instead, the introduction of the frailty term increases the estimate of the educational gradient significantly, thus going in the direction indicated by the literature. The rate ratios compared to the highly educated men changes from 1.16 to 1.22 and from 1.24 to 1.30 for medium and low educated groups respectively.

Controlling for unobserved frailty additionally affects the estimates of the geographical differences in the macro-region of birth. The patterns estimated by the models Gompertz 2 for men and Makeham 2 for women, depict a completely different picture from the pattern estimated by the models Gamma-Gompertz and Gamma-Makeham respectively. The latter estimate differentials that are consistent with what is known to be the geographical health and mortality gradient in Italy, recently presented in the results of the paper from Biggeri et al. (2011) where the authors have analyzed the evolution of cohort mortality in the Italian regions until the cohort born in 1949-1958 (the last ones for which data were available). Although inter-regional differences have been reducing from older to younger cohort in both men and women[2], differences still persisted. Among men, those born in the southern regions show a clear survival advantage, followed by the central regions and the North-western regions, up to the least advantaged North-eastern ones.

[1] A possible explanation calls into consideration the fact that, once frailty is introduced in the model, the functional form of the mortality curve is different. More specifically, the log-mortality function is not a straight line anymore but is allowed to bend over. Given that model 3 has estimated the same a, c and cohort parameters, the slope of the curve b has increased and, consequently, mortality at the reference category is estimated to be higher, more at younger ages and a little bit less at older ages. Therefore, the educational coefficients tend to be reduced to fit the same set of data points.

[2] This is attributed to the creation of the National Health System in the 70's, that has reduced the regional differences in the health services on offer.

This is not captured by the model Gompertz 2, which finds no significant difference in the mortality risk for the macro-region of birth South and Islands. On the contrary, when controlling for unobserved heterogeneity, the advantage is detected and appears strongly significant (rate ratio 0.95 (0.94 - 0.96)). The model with frailty also finds a different pattern among women. Although the differences are known to be less pronounced, it is also known that, as concerns women, the southern regions are the most disadvantaged. The best performers in terms of women cohort mortality are the central regions, followed by the north-eastern regions and, lastly, by the north-western ones. According to the model Makeham 2, however, women born in the region South and Islands regions have a strongly significant survival advantage that becomes not significant in the Gamma-Makeham model.

6.2.2 Frailty modeling with period covariates

This section reports the results of the models that have controlled for the period mortality improvement splitting the 36 years of follow up into different sub-periods.

As already explained in chapter 5, paragraph 5.4.3, the models without frailty were estimated controlling for twelve periods of three calendar years each. The introduction of the frailty component required a more complex framework, the shared gamma-frailty model. In order to perform the estimation, the number of periods had to be reduced to two (1971-1990 and 1991-2007) and a random subsampling strategy was used. The optimization routine was repeatedly applied on a 1% sample of the data randomly drawn without replacement.

Tables 6.5 and 6.6 display the results of the three models, the first two without frailty and the third one with frailty.

For the sake of simplicity, the estimated period coefficients are omitted but the interested reader can find them in appendix A, tables A.2 and A.3. A visual representation of the mortality improvement over time is given in figure 6.3 that, in respect to the reference category 1971-1973, clearly shows

	Makeham 1		Makeham 2		Gamma-Makeham	
	estimate	95% CI	estimate	95% CI	estimate	95% CI
a	0.000	(0.000-0.000)	0.000	(0.000-0.000)	0.000	(0.000-0.000)
b	0.106	(0.105-0.107)	0.106	(0.105-0.107)	0.117	(0.115-0.119)
c	0.001	(0.001-0.001)	0.001	(0.001-0.001)	0.001	(0.001-0.001)
σ^2	-	-	-	-	0.096	(0.082-0.111)
cohort	0.016	(0.015-0.016)	0.016	(0.015-0.016)	0.017	(0.016-0.017)
Education Level						
High	1	(-)	1	(-)	1	(-)
Medium	1.146	(1.121-1.171)	1.141	(1.116-1.166)	1.111	(1.086-1.137)
Low	1.235	(1.212-1.259)	1.246	(1.222-1.270)	1.213	(1.188-1.238)
Region of Birth						
N-West	-	-	1	(-)	1	(-)
N-East	-	-	0.989	(0.973-1.004)	0.974	(0.958-0.991)
Center	-	-	0.939	(0.913-0.966)	0.968	(0.939-0.998)
South	-	-	0.932	(0.919-0.945)	0.987	(0.973-1.002)
Abroad	-	-	1.071	(1.047-1.096)	0.993	(0.968-1.018)
logLK	-663214		-663238		-663098	
AIC	1326440		1326496		1326218	

Table 6.3: Women - Results of the regression models with cohort covariate (coefficients in grey are not significant).

	Gompertz 1		Gompertz 2		Gamma-Gompertz	
	estimate	95% CI	estimate	95% CI	estimate	95% CI
a	0.000	(0.000-0.000)	0.000	(0.000-0.000)	0.000	(0.000-0.000)
b	0.081	(0.080-0.081)	0.083	(0.080-0.082)	0.083	(0.082-0.084)
σ^2	-	-	-	-	0.035	(0.027-0.045)
cohort	0.016	(0.015-0.016)	0.016	(0.015-0.016)	0.016	(0.015-0.016)
Education Level						
High	1	(-)	1	(-)	1	(-)
Medium	1.157	(1.138-1.177)	1.166	(1.147-1.186)	1.221	(1.200-1.243)
Low	1.254	(1.237-1.272)	1.239	(1.221-1.257)	1.302	(1.283-1.322)
Region of Birth						
N-West	-	-	1	(-)	1	(-)
N-East	-	-	1.053	(1.036-1.070)	1.060	(1.042-1.077)
Center	-	-	1.011	(0.984-1.038)	0.996	(0.969-1.024)
South	-	-	1.000	(0.988-1.012)	0.950	(0.938-0.962)
Abroad	-	-	1.031	(1.006-1.057)	0.998	(0.974-1.024)
logLK	-651258		-651219		-651082	
AIC	1302526		1302456		1302184	

Table 6.4: Men - Results of the regression models with cohort covariate (coefficients in grey are not significant).

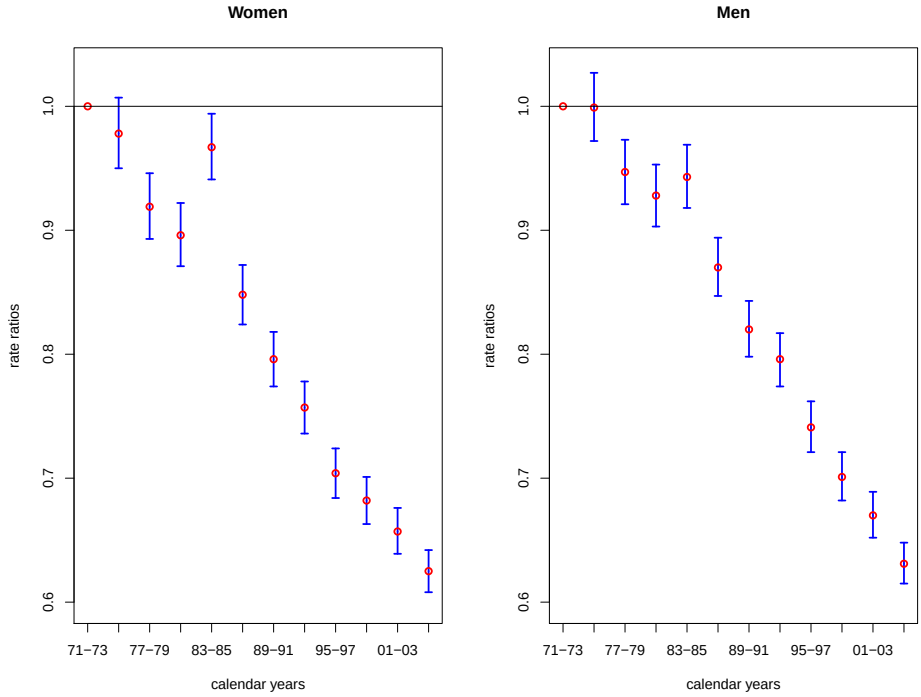


Figure 6.3: Mortality rate ratios of different calendar times compared to the first period 1971-1973, for men and women.

a declining trend for both men and women [3].

Compared to the models with cohort variables, the models with the period variables estimate a much higher unobserved heterogeneity component, for both men and women. The σ^2 term is now 0.27 for men and 0.29 for women while it was 0.04 and 0.1 for men and women respectively. This is reasonable because periods are much more heterogeneous than cohorts, being a mixture of several different cohorts.

Comparing the period models with and without frailty component, no-

[3]The trend is temporarily interrupted in 1983-1985. One could speculate that some event related to the local health system might have stopped the improvement. For example, it could be thought of a budget deficit that caused a reduction of health services. This aspect is beyond the scope of this study and will not be explored further. Nevertheless, it is an interesting finding that deserves future investigation

	Makeham 1		Makeham 2		Gamma-Makeham	
	estimate	95% CI	estimate	95% CI	Mean	0.025-0.975
a	0.000	(0.000-0.000)	0.000	(0.000-0.000)	0.008	(0.000-0.016)
b	0.121	(0.120-0.121)	0.121	(0.120-0.122)	0.084	(0.073-0.106)
c	0.001	(0.001-0.002)	0.001	(0.001-0.002)	0.000	(0.000-0.000)
σ^2	-	-	-	-	0.292	(0.174-0.367)
Education Level						
High	1	(-)	1	(-)		
Medium	1.120	(1.096-1.144)	1.107	(1.083-1.131)	1.256	(1.053-1.347)
Low	1.208	(1.186-1.232)	1.209	(1.186-1.232)	1.475	(1.103-1.641)
Region of Birth						
N-West	-	-	1	(-)		
N-East	-	-	0.963	(0.948-0.978)	1.122	(0.888-1.217)
Center	-	-	0.964	(0.938-0.992)	1.102	(0.864-1.218)
South	-	-	0.962	(0.949-0.975)	1.130	(0.904-1.220)
Abroad	-	-	0.985	(0.962-1.009)	1.082	(0.847-1.215)
logLK	-663105		-663081		Na	
AIC	1326242		1326204		Na	

Table 6.5: Women - Results of the regression models with period covariates (coefficients in grey are not significant). The estimated period coefficients are reported in appendix, table A.2. The model Gamma-Gompertz does not report conventional point estimates and confidence intervals but the mean value and the 0.025-0.975 quantiles of the empirical distribution of parameters obtained from the repeated estimates via random subsampling.

	Gompertz 1		Gompertz 2		Gamma-Gompertz	
	estimate	95% CI	estimate	95% CI	mean estim.	0.025-0.975 quant.
a	0.000	(0.000-0.000)	0.000	(0.000-0.000)	0.004	(0.000-0.010)
b	0.096	(0.096-0.097)	0.096	(0.095-0.096)	0.069	(0.061-0.163)
σ^2	-	-	-	-	0.269	(0.026-0.367)
Education Level						
High	1	(-)	1	(-)		
Medium	1.184	(1.164-1.204)	1.204	(1.184-1.225)	1.277	(1.054-1.349)
Low	1.267	(1.249-1.285)	1.301	(1.282-1.320)	1.268	(1.074-1.591)
Region of Birth						
N-West	-	-	1	(-)	1	(-)
N-East	-	-	1.040	(1.024-1.057)	1.075	(0.855-1.220)
Center	-	-	0.943	(0.917-0.969)	1.081	(0.854-1.217)
South	-	-	0.900	(0.889-0.911)	1.037	(0.854-1.216)
Abroad	-	-	0.965	(0.941-0.989)	1.082	(0.864-1.218)
logLK	-651208		-650997		Na	
AIC	1302446		1302034		Na	

Table 6.6: Men - Results of the regression models with period covariates (coefficients in grey are not significant). The estimated period coefficients are reported in appendix, table A.3. The model Gamma-Gompertz does not report conventional point estimates and confidence intervals but the mean value and the 0.025-0.975 quantiles of the empirical distribution of parameters obtained from the repeated estimates via random subsampling.

table differences in the estimates of the baseline parameters emerge.

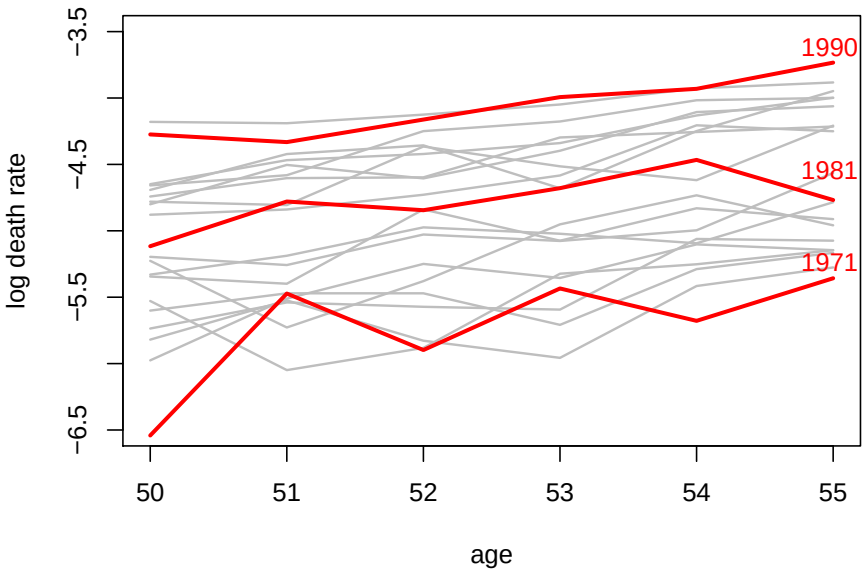


Figure 6.4: Period log death rates of women from age 50 to age 55 from 1971 to 1990, the reference category for period variable in the model with unobserved frailty.

The c term of the Makeham baseline hazard, used for the women, practically disappears. However, this change is not likely to be due to the introduction of frailty but to the different period variables used in the models. In the model without frailty it was possible to introduce twelve period variables while in the model with frailty the number had to be reduced to two broad periods. The coefficients of the baseline hazard in the first case refer to the reference category 1971-1973 while in the second case to a much broader one, 1971-1990. Figure 6.4 shows how the 'bump' in period death rates between ages 50 and 55, which is captured by the component c, is mainly visible only in the beginning of the period and then fades away. Consequently, in the narrower reference category of 1971-1973 this component was estimated to be significant, while in the much broader category of 1971-1990, the 'signal' for it was much weaker and the model estimated a c term close to zero.

When the shape of the mortality curve does not require a c term, the initial mortality level is not decomposed anymore into two components and is described only by the a term, whose estimated value increases significantly when the c term is found to be zero.

This has consequences also for the b parameter, whose estimate is now lower, since the two values are known to be negatively correlated (Strehler and Mildvan 1960; Riggs 1990). The negative correlation emerges in the model with frailty, where a was found to be higher than in the model without frailty and b lower. This model has been estimated with random subsampling, repeating several times the estimation on a small sample size randomly drawn from the dataset. Figures 6.5 and 6.6 present scatter plots of the estimated parameters, for women and men respectively. Scatter plots display the correlation between pairs of variables. Given a set of variables, they can be organized into a matrix, making it easy to look at all pairwise correlations in one place. Along the diagonal, summaries of the distribution for each variable can be placed. In this case the diagonal contains the density based on the empirical distribution of the repeated estimates. The negative correlation between a and b is clearly visible (scatter plots in 1st row-2nd column and in 2nd row-1st column).

The introduction of frailty in the model affects the mortality gradient by education of both men and women. Although the uncertainty around the estimates does not allow assessing a precise effect, it is possible to see that both medium and low educated groups have a significantly higher risk of death than high educated groups and that the differential could be underestimated by the model without frailty. The rate ratio for women with medium education, in fact, lies between 1.05 and 1.34, compared to 1.08 and 1.13 of the model without frailty; for the low education group it lies between 1.1 and 1.6, compared to 1.18 and 1.23 of the model without frailty. The same pattern can be observed among men: the rate ratios of medium and low educated groups respect to the highly educated one lies in a higher confidence region in the model with frailty compared to the models without. This is consistent

with the literature about the bias caused by not taking into account unobserved heterogeneity that might lead to underestimation of the effect of the covariates included in the model.

Interestingly, the gradient between men with medium and low education is much less defined than among women, and this seems to suggest that the male population in Turin might have undergone a process oh homogenization. This seems confirmed also by the estimated lower variance of frailty than thate for women.

Regarding the geographical differentials, in contrast, it is not possible to identify a pattern. This could be due to the fact that the signal for such differentials in the data is weaker than the one for groups according to education level. Although the number of repetitions used in the analysis, 250, is indicated by the literature as sufficient, especially for very complex models, it is possible that it was inadequate for the model to identify a significant difference. However, the densities of the geographical parameters in figure 6.6 might help reveal interesting dynamics. For the macro-region of birth South, the likely rate ratios span over a wide range of values, from 0.8 to 1.2, while for the other variables the density is unimodal. Taking into account that 1) there was no clearly distinguishable gradient for low compared to medium education, 2) the majority of men with low education in Turin come from a southern region (figure 4.4), 3) subjects that decide to migrate might be in better health and 4) Biggeri et al. (2011) have found a survivorship advantage for male cohorts from the southern regions, it is possible that the variables low education and south macro-region of birth interact. In particular, one could speculate that the migration flows might have contributed to reducing the mortality differentials by education.

Unluckily, in the case of the models with period covariates it was not possible to implement the comparison via AIC because, for the model including frailty, the procedure used did not allow the obtaining of a likelihood value comparable with the values of the models without frailty.

Figure 6.5: Women - scatterplot and density.

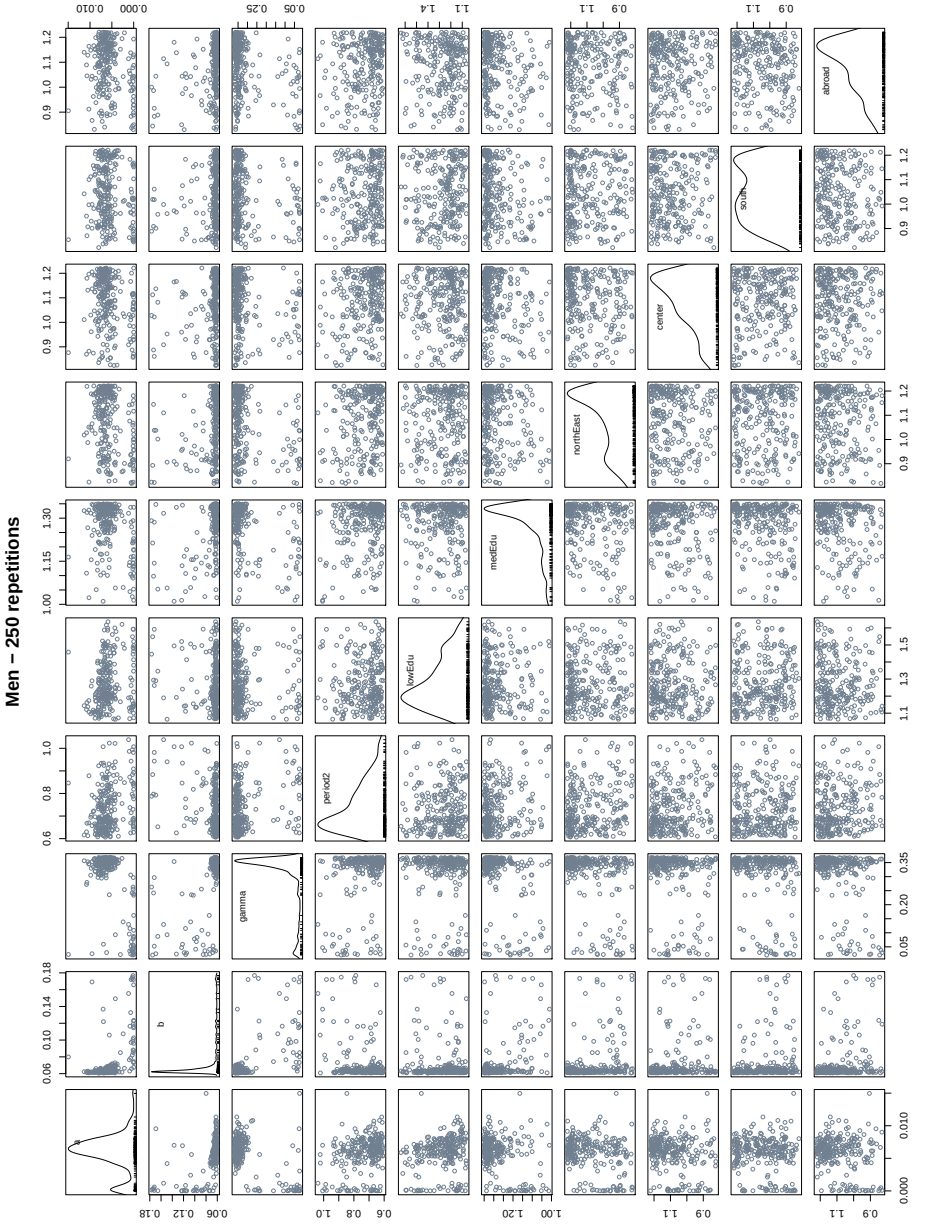


Figure 6.6: Men - scatterplot and density.

6.3 Summary of the results

Education level has been used as proxy for socioeconomic status. The choice of this indicator is consistent with the literature on the topic (for more details please see paragraph 2.3 in chapter 2). Given that the data involve several birth cohorts for a long follow up period, to facilitate the comparison, three broad educational groups were created: high (upper secondary or higher), medium (lower secondary) and low (primary or lower).

Plotting the cohort death rates for the three education levels, on a log scale (figure 4.6), showed that the three hazards gradually converged. So the data present the well known reduction of relative mortality differences at old ages.

A preliminary analysis performed using an exponential model, estimated with age group-education group interaction, has shown that the reduction of the gradient over age is statistically significant.

The next step was to examine the age-pattern of the hazard function, using parametric survival models with age dependent hazards.

After having investigated which baseline hazard was the most appropriate (the results indicate a Gompertz function for the men and a Gompertz-Makeham for the women), models with and without controlling for unobserved heterogeneity have been applied.

First, the models have been applied to the analysis of total population mortality and then the education level was taken into account. The comparison of the models via AIC showed that when unobserved heterogeneity of frailty was included, the model gave a statistically significant better fit than when this component was not included. This strengthens the validity of the selection hypothesis as an alternative explanation for the reduction of the gradient in socioeconomic mortality at old ages.

The models that controlled for period mortality improvement, as expected, estimated a higher heterogeneity (the term σ^2) than the models that took into account a cohort mortality improvement. Periods are an aggregate of different generations and, therefore, are expected to be more heterogeneous

than the cohorts themselves.

In all the implemented models, the female variance of frailty was estimated to be higher than the variance of frailty for the males. This indicates that the male population is somewhat more homogeneous than the female one, suggesting that it might have undergone a process of homogenization. This seems to confirm what was hypothesized about the effect of the extensive migration flows that characterize the history of Turin. The data description in chapter 4 showed that the majority of the lowly educated individuals were migrants born in the South, while the majority of the highly educated individuals were the native population. Migrants are supposedly healthier than the average population, because unhealthy people are less likely to migrate. Migration could have contributed to reduce the relative differences in susceptibility among men and not, or to a smaller extent, among women. Women are likely to have been passive actors in the migratory decision. Many of them, in fact, often followed their husbands only after some time.

The literature about unobserved heterogeneity invites the careful consideration of the bias caused by neglecting the presence of this component and of its selective action on the population at risk. In duration dependence models, both the hazard function and the effect of observed covariates are likely to be biased towards zero (for more details, please, see paragraph 3.3.1 in chapter 3). This means that not controlling for frailty might lead to underestimating the effect of the variables in the analysis.

The comparison of the two approaches, indeed, goes in the direction indicated by the literature. In the majority of the cases, the educational gradient estimated by the models with frailty was higher than the one estimated by the models without frailty. Even in the cases when big uncertainty around the point estimates did not allow assessing a precise value, the confidence regions in the models with frailty were spanning over higher values than those in the models without frailty.

6.4 Conclusions

Hidden differences in survival chances between individuals in a population, called unobserved frailty, influence the observed population mortality dynamic. Selection processes related to these characteristics constantly change the population, as frailer individuals die earlier than more robust ones. It is important to consider these hidden differences because, if neglected, they can lead to biased estimates of the hazard function, as well as of the effect of observed covariates on the survival chances. To control for unobserved heterogeneity of frailty and to evaluate their impact on the observed mortality dynamics, frailty models have been developed.

In differential mortality analysis it is important to control for hidden frailty.

First, because not controlling for it, in proportional hazard models, could lead to biased estimates of the effect of the socioeconomic group on the mortality risk. The statistical literature shows that the bias is towards zero, therefore, ignoring the frailty component could cause the underestimation of the differentials.

Second, because selection due to unobserved frailty could provide an alternative explanation for the phenomenon of decreasing relative differences in death rates by socioeconomic group at old ages, which is often observed in research about differential mortality. The age-as-leveler hypothesis attributes this phenomenon to factors like governmental support to the elderly and biological frailty that level off socioeconomic differences and make individuals more equally weak as they age. However, this phenomenon could be an artifact of selection at the population (group) level, due to unobserved characteristics of the individuals. Selective effects of high early mortality experienced by the disadvantaged group would leave more robust individuals at old ages, causing the convergence with the hazard of the lower mortality group that is subjected to weaker selection.

This study analyzed the socioeconomic gradient mortality pattern over age from a longitudinal perspective. Its aims were 1) to investigate whether

heterogeneity matters by analyzing if and how the estimates change when introducing into the model a component of unobserved heterogeneity of frailty and 2) to investigate whether the theoretical framework of the frailty models can better explain the pattern, thus giving an alternative valid explanation for the observed pattern of reduction of relative differences in mortality by socioeconomic status at old ages.

This study found that in the majority of the cases, when the component of unobserved frailty was included in the model, the estimated educational differences between high education and medium and low education groups were larger than those estimated in the models that did not control for unobserved frailty. Even in the cases when big uncertainty around the point estimates did not allow assessing a precise value, the confidence regions in the models with frailty were spanning over higher values than those in the models without frailty. These results are consistent with the findings of the statistical literature about the bias caused by ignoring the effect of omitted covariates in proportional hazards models. They suggest that, when estimating survival models about mortality differentials, we might be underestimating these differentials if we do not control for unobserved heterogeneity.

Moreover, the comparison of the models showed that when unobserved heterogeneity of frailty was included, the model gave a statistically significant better fit than when this component was not included. This strengthens the validity of the selection hypothesis as alternative explanation for the reduction of the gradient in socioeconomic mortality at old ages and invites to carefully consider that the dynamics at the individual level and at the population (or group) level are not necessarily the same.

Apart from the statistical aspect of making the right inference, this has also important policies facets. Specifically, when studying differential survival chances in socioeconomic groups, the tendency to dismiss the importance of such differences in old ages because they are observed to diminish should be avoided. Individuals might experience throughout their life a position of disadvantage that does not fade away when they age, and the lessening of

differences at old ages could be the result of a stronger selection due to early higher mortality that disadvantaged groups are still subject to.

Appendix A

Appendix

A.1 R source code

The following codes have been programmed by the author of the thesis in R, the software used for the analysis.

The symbol # is used to comment out the part of the code preceded by it, so that comments and notes can be added to the program.

Note that, to simplify the computation, the parameters of the models were always considered in the log scale and exponentiated back inside the functions, so that they were constrained to be positive.

A.1.1 Models withouth frailty

Exponential

```
#exponential hazard function with covariates
hazcov <- function(pars,betas,covs) {
  lambda <- exp(pars)
  out <- lambda*exp(covs%*%betas)
  return(out)
}

#exponential survival function with covariates
survcov <- function(pars,betas,covs,x) {
```

```
  lambda <- exp(pars)
  subsurv <- exp(-lambda*x)
  ccoef <- as.vector(exp(covs%*%betas))
  surv <- subsurv^(ccoef)
  out <- surv
  return(out)
}

#individual likelihood contribution
likc <- function(pars,betas,covs,status,entry,exit) {
  haz <- hazcov(pars=pars,betas=betas,covs=covs)^(status)
  surv.exit <- survcov(pars=pars,betas=betas,covs=covs,x=exit)
  surv.entry <-  survcov(pars=pars,betas=betas,covs=covs,x=entry)
  out <- haz*surv.exit/surv.entry
  return(out)
}

#summing individual contributions up
LogLikc <- function(par,covs,status,entry,exit) {
        print(par)
  params <- par[1]
  betas <- par[2:length(par)]
  liks <- likc(pars=params,betas=betas,covs=covs,
  status=status,entry=entry,exit=exit)
  out <- sum(log(liks))
  print(out)
  return(out)
}

attach(dataset)

#individual's covariate profile from dataset's variables
covprof <- matrix(c(varX,varY,varZ...),
                  byrow=F,ncol=3...)

#optimizing the log-likelihood function
start <- c(...) #vector of starting values
opt <- optim(par=start,
```

```
              covs=covprof, status=status, fn=LogLikc,
              entry=agestart, exit=ageend,
              control=list(fnscale=-1),
              hessian=T)

#agestart=age at entry (left truncation time)
#ageend=age at exit
#status=0 if right censored, 1 if dead
```

Gompertz

```
#gompertz hazard with covariates
ghc <- function(pars, betas, covs, x) {
  a <- exp(pars[1])
  b <- exp(pars[2])
  out <- a*exp(b*x)*exp(covs%*%betas)
  return(out)
}

#gompertz surv function with covariates
gsc <- function(pars, betas, covs, x) {
  a <- exp(pars[1])
  b <- exp(pars[2])
  subsurv <- exp(a/b*(1-exp(b*x)))
  ccoef <- as.vector(exp(covs%*%betas))
  surv <- subsurv^(ccoef)
  out <- surv
  return(out)
}

#individual likelihood contribution
likc <- function(pars, betas, covs, status, entry, exit) {
  haz <- ghc(pars=pars, betas=betas, covs=covs, x=exit)^(status)
  surv.exit <- gsc(pars=pars, betas=betas, covs=covs, x=exit)
  surv.entry <-  gsc(pars=pars, betas=betas, covs=covs, x=entry)
  out <- haz*surv.exit/surv.entry
  return(out)
```

```
}

#summing individual contributions up
LogLikc <- function(par,covs,status,entry,exit) {
        print(par)
  params <- par[1:2]
  betas <- par[3:length(par)]
  liks <- likc(pars=params,betas=betas,covs=covs,
  status=status,entry=entry,exit=exit)
  out <- sum(log(liks))
  print(out)
  return(out)
}

attach(dataset)

... [continue as the previous code]
```

Makeham

```
# makeham hazard with covariates
mkhc <- function(pars,betas,covs,x) {
  a <- exp(pars[1])
  b <- exp(pars[2])
  c <- exp(pars[3])
  out <- (a*exp(b*x)+c)*exp(covs%*%betas)
  return(out)
}

# makeham surv function with covariates
mksc <- function(pars,betas,covs,x) {
  a <- exp(pars[1])
  b <- exp(pars[2])
  c <- exp(pars[3])
  subsurv <- exp(a/b*(1-exp(b*x))-c*x)
  ccoef <- as.vector(exp(covs%*%betas))
  surv <- subsurv^(ccoef)
```

```
  out <- surv
  return(out)
}

#individual likelihood contribution
likc <- function(pars,betas,covs,status,entry,exit) {
  haz <- mkhc(pars=pars,betas=betas,covs=covs,x=exit)^(status)
  surv.exit <- mksc(pars=pars,betas=betas,covs=covs,x=exit)
  surv.entry <-  mksc(pars=pars,betas=betas,covs=covs,x=entry)
  out <- haz*surv.exit/surv.entry
  return(out)
}

#summing individual contributions up
LogLikc <- function(par,covs,status,entry,exit) {
        print(par)
  params <- par[1:3]
  betas <- par[4:length(par)]
  liks <- likc(pars=params,betas=betas,covs=covs,
  status=status,entry=entry,exit=exit)
  out <- sum(log(liks))
  print(out)
  return(out)
}

attach(dataset)

... [continue as the previous code]
```

A.1.2 Models with univariate frailty

Gamma-Gompertz

```
#gompertz baseline hazard
ghc <- function(pars,betas,covs,x) {
  a <- exp(pars[1])
  b <- exp(pars[2])
```

```
  out <- a*exp(b*x)*exp(covs%*%betas)
  return(out)
}

#gompertz cumulative hazard
gcumhaz <- function(pars,betas,covs,x) {
  a <- exp(pars[1])
  b <- exp(pars[2])
  sb <- a/b*(exp(b*x)-1)
  ccoef <- exp(covs%*%betas)
  out <- sb*(ccoef)
  return(out)
}

#gompertz marginal hazard
gmh <- function(pars,betas,covs,x) {
  gamma <- exp(pars[3])
  num <- ghc(pars[1:2],betas=betas,covs=covs,x=x)
  denom <- 1+gamma*gcumhaz(pars[1:2],betas=betas,covs=covs,x=x)
  out <- num/denom
  return(out)
}

#gompertz marginal survival
gms <- function(pars,betas,covs,x) {
  gamma <- exp(pars[3])
  part1 <- (1+gamma*gcumhaz(pars[1:2],betas=betas,covs=covs,x=x))
  part2 <- (-1/gamma)
  out <- part1^part2
  return(out)
}

#individual likelihood contribution
likc <- function(pars,betas,covs,status,entry,exit) {
  haz <- gmh(pars=pars,betas=betas,covs=covs,x=exit)^(status)
  surv.exit <- gms(pars=pars,betas=betas,covs=covs,x=exit)
  surv.entry <-  gms(pars=pars,betas=betas,covs=covs,x=entry)
  out <- haz*surv.exit/surv.entry
```

```
  return(out)
}

#summing individual contributions up
LogLikc <- function(par,covs,status,entry,exit) {
        print(par)
  params <- par[1:3]
  betas <- par[4:length(par)]
  liks <- likc(pars=params,betas=betas,covs=covs,
  status=status,entry=entry,exit=exit)
  out <- sum(log(liks))
  print(out)
  return(out)
}

attach(dataset)

... [continue as the previous code]
```

Gamma-Makeham

```
#makeham hazard with covariates
mkhc <- function(pars,betas,covs,x) {
  a <- exp(pars[1])
  b <- exp(pars[2])
  c <- exp(pars[3])
  out <- (a*exp(b*x)+c)*exp(covs%*%betas)
  return(out)
}

#makeham cum. hazard function with covariates
mkcumhaz <- function(pars,betas,covs,x) {
  a <- exp(pars[1])
  b <- exp(pars[2])
  c <- exp(pars[3])
  sb <- a/b*(exp(b*x)-1)+c*x
  ccoef <- exp(covs%*%betas)
```

```
  out <- sb*(ccoef)
  return(out)
}

#makeham marginal hazard
mkmh <- function(pars,betas,covs,x) {
  gamma <- exp(pars[4])
  num <- mkhc(pars[1:3],betas=betas,covs=covs,x=x)
  denom <- 1+gamma*mkcumhaz(pars[1:3],betas=betas,covs=covs,x=x)
  out <- num/denom
  out <- num/denom
  return(out)
}

#makeham marginal survival
mkms <- function(pars,betas,covs,x) {
  gamma <- exp(pars[4])
  part1 <- (1+gamma*mkcumhaz(pars[1:3],betas=betas,covs=covs,x=x))
  part2 <- (-1/gamma)
  out <- part1^part2
  return(out)
}

#individual likelihood contribution
likc <- function(pars,betas,covs,status,entry,exit) {
  haz <- mkmh(pars=pars,betas=betas,covs=covs,x=exit)^(status)
  surv.exit <- mkms(pars=pars,betas=betas,covs=covs,x=exit)
  surv.entry <-  mkms(pars=pars,betas=betas,covs=covs,x=entry)
  out <- haz*surv.exit/surv.entry
  return(out)
}

#summing individual contributions up
LogLikc <- function(par,edumat,covs,status,entry,exit) {
        print(par)
  params <- par[1:4]
  betas <- par[5:length(par)]
  liks <- mklikc(pars=params,betas=betas,covs=covs,
```

```
  status=status,entry=entry,exit=exit)
  out <- sum(log(liks))
  print(out)
  return(out)
}

attach(dataset)

... [continue as the previous code]
```

A.1.3 Models with shared frailty

This section shows the code used for the estimation of the shared frailty models by subsampling. These codes differ from the others not only in the definition of the hazard, survival and likelihood functions, but also in the optimization routine. Several routines were tried because they failed to converge.

Gamma-Gompertz (with shared frailty)

```
library(plyr)
library(DEoptim)
library(compiler)
library(snowfall)

#set number of CPUs to use for the estimation
sfInit(parallel=T,cpus=4)

load("dataset.rdata")
attach(dataset)

#grid of combinations of education and area of birth categories
        #edu=1 low, 2 medium, 3 high
        #birth=1 NWest,2 NEast, 3 Center, 4 South, 5 Abroad
combinations <- expand.grid(edu=c(1,2,3),birth=c(1,2,3,4,5))
combinations <- cbind(combinations,dim=0)
```

```
for (i in 1:15) {
a <- subset(dataset,educlass==combinations[i,1]
& areabirth==combinations[i,2])
combinations[i,3] <- length(unique(a$id))
}

#for each combination of education and birth categories,
#create list of IDs that have that combination
IDlist <- list()
for (i in 1:15) {
a <- subset(dati,educlass==combinations[i,1] &
areabirth==combinations[i,2])
IDlist[[i]] <- unique(a$id)
}

#function that takes "sampsize"% sample IDs representative
#of education and birth area and extracts from the dataset
#the randomly sampled ids

takesample <- function(IDlist,dati,sampsize){
IDsample <- list()
for (i in 1:15) {
IDsample[[i]] <- sample(IDlist[[i]],
size=round(length(IDlist[[i]])*sampsize),replace=F)
}
out <- dati[which(dati$id%in%unlist(IDsample)==T),]
return(out)
}

#cluster likelihood
clusLik <- function(pars,ids,columns,data,i) {
  a <- exp(pars[1])
  b <- exp(pars[2])
  varfr <- exp(pars[3])
  betas <- exp(pars[4:length(pars)])
  clus <- subset(dat,id==ids[i])
  covs <- as.matrix(clus[,columns])
  status <- clus$status
```

```
  entry <- clus$agestart
  exit <- clus$ageend
  #part1 <- prod(a*exp(b*exit)*exp(covs%*%betas))
  part1 <-  sum(log(a)+(b*exit)+(covs%*%betas))
  part2 <-  (gamma(1/varfr+sum(status))/gamma(1/varfr))
  *varfr^(sum(status))
  surventry <- exp(a/b*(1-exp(b*entry)))^(exp(covs%*%betas))
  survexit <-  exp(a/b*(1-exp(b*exit)))^(exp(covs%*%betas))
  if(any(as.vector(survexit)==0))
    {survexit[which(as.vector(survexit)==0)] <- 1e-128}
  part3 <-  (1-varfr*(sum(log(surventry))))^(1/varfr)
  part4 <-  (1-varfr*(sum(log(survexit))))^(-1/varfr-sum(status))
  clusLiks <- part1+log(part2)+log(part3)+log(part4)
  out <- clusLiks
  return(out)
}

LogLikc <- function(pars,ids,columns,data) {
  print(pars)
  pre.out <- sfSapply(x=1:length(ids),fun=clusLik,pars=pars,
  ids=ids,columns=columns,data=dataset)
  out <- sum(pre.out)
  print(out)
  return(-out)
}

compLogLikc <- cmpfun(LogLikc)

lower <- c(...) #vector of lower boundaries for the parameters
upper <- c(...) #vector of upper boundaries for the parameters


#specify the number of repetitions to be done
resamptimes <- n
#create a list to store results
results <- list()
for(i in 1:resamptimes) {
print(paste("this is resampling",i))
```

```
#extract 1% sample size from dataset
dat <- takesample(IDlist=IDlist,dati=dataset,sampsize=0.01)
ids <- unique(id)
#indicate the column positions of the covariates in the dataset
columns <- c(...)
results[[i]] <-  DEoptim(fn=complogLik2,lower=lower,upper=upper,
                 DEoptim.control(itermax=100,reltol=1e-03),
                 ids=ids,columns=columns,dat=dat,resamp=i)
}
```

Gamma-Makeham (with shared frailty)

```
library(plyr)
library(DEoptim)
library(compiler)
library(snowfall)

... [as previous code]

clusLik <- function(pars,ids,columns,dat,i) {
  a <- exp(pars[1])
  b <- exp(pars[2])
  m <- exp(pars[3])
  varfr <- exp(pars[4])
  betas <- pars[5:length(pars)]
  clus <- subset(dat,id==ids[i])
  covs <- as.matrix(clus[,columns])
  status <- clus$status
  entry <- clus$agestart
  exit <- clus$ageend
  #part1 <- prod((a*exp(b*exit)+m)*exp(covs%*%betas))
  part1 <-  sum(log(a*exp(b*exit)+m)+(covs%*%betas))
  part2 <-  (gamma(1/varfr+sum(status))/gamma(1/varfr))
  *varfr^(sum(status))
surventry <- exp(-a/b*(exp(b*entry)-1)+m*entry)^(exp(covs%*%betas))
survexit <-  exp(-a/b*(exp(b*exit)-1)+m*exit)^(exp(covs%*%betas))
  if(any(as.vector(survexit)==0))
```

```
    {survexit[which(as.vector(survexit)==0)] <- 1e-128}
  part3 <-  (1-varfr*(sum(log(surventry))))^(1/varfr)
  part4 <-  (1-varfr*(sum(log(survexit))))^(-1/varfr-sum(status))
  clusLiks <- part1+log(part2)+log(part3)+log(part4)
  out <- clusLiks
  return(out)
}

... [continue as the previous code]
```

A.2 Tables and figures

	Women		Men	
	estimate	95% CI	estimate	95% CI
Model without interaction between age and education				
λ	0.009	(0.009-0.009)	0.017	(0.017-0.017)
age 50-80	0	-	0	-
age 80+	2.135	(2.125-2.145)	1.855	(1.844-1.866)
high education	1	-	1	-
medium education	0.176	(0.154-0.198)	0.167	(0.150-0.184)
low education	0.345	(0.326-0.365)	0.364	(0.350-0.378)
1971-1991	1	-	1	-
1991-2007	-0.149	(-0.159 - -0.139)	-0.152	(-0.162 - -0.142)
Likelihood	-700668.8		-688641.5	
Model with interaction between age and education				
λ	0.008	(0.008-0.009)	0.016	(0.016-0.016)
age 50-80	0	-	0	-
age 80+	2.400	(2.364-2.436)	2.079	(2.052-2.107)
high education	0	-	0	-
medium education	0.223	(0.193-0.254)	0.210	(0.190-0.230)
low education	0.466	(0.440-0.493)	0.452	(0.435-0.469)
low edu * age 80+	-0.309	(-0.347 - -0.271)	-0.293	(-0.323 - -0.262)
medium edu * age 80+	-0.184	(-0.227 - -0.140)	-0.132	(-0.169 - -0.094)
1971-1991	0	-	0	-
1991-2007	-0.146	(-0.156 - -0.136)	-0.152	(-0.162 - -0.142)
Likelihood	-700508.8		-688443.9	

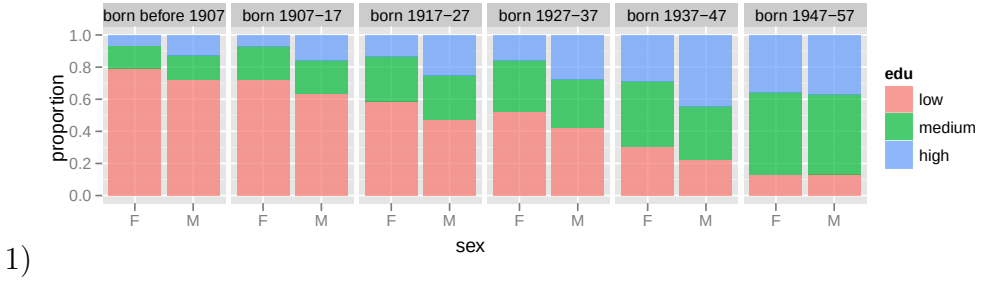
Table A.1: Results of the exponential hazard models with covariates.

	Makeham 1		Makeham 2		Gamma-Makeham		
	estimate	95% CI	estimate	95% CI	mean estim.		0.025-0.975 quant.
Calendar year							
1971-1973	1	1	1	1			
1974-1976	0.965	(0.938-0.994)	0.978	(0.950-1.007)			
1977-1979	0.902	(0.876-0.928)	0.919	(0.893-0.946)			
1980-1982	0.879	(0.855-0.905)	0.896	(0.871-0.922)			
1983-1985	0.947	(0.922-0.974)	0.967	(0.941-0.994)			
1986-1988	0.832	(0.809-0.856)	0.848	(0.824-0.872)	1971-1990:	1	(-)
1989-1991	0.779	(0.758-0.801)	0.796	(0.774-0.818)	1991-2007:	0.888	(0.671-1.035)
1992-1994	0.743	(0.723-0.764)	0.757	(0.736-0.778)			
1995-1997	0.690	(0.671-0.710)	0.704	(0.684-0.724)			
1998-2000	0.669	(0.651-0.688)	0.682	(0.663-0.701)			
2001-2003	0.642	(0.625-0.660)	0.657	(0.639-0.676)			
2004-2007	0.609	(0.593-0.626)	0.625	(0.608-0.642)			

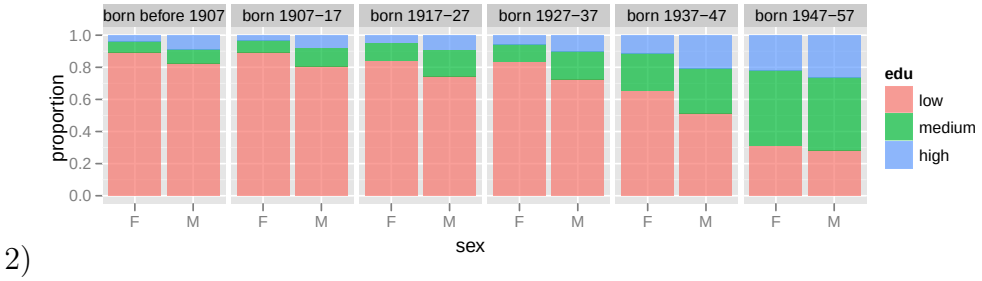
Table A.2: Women - Results of the regression models with period covariates (coefficients in grey are not significant).

	Gompertz 1		Gompertz 2		Gamma-Gompertz		
	estimate	95% CI	estimate	95% CI	mean estim.		0.025-0.975 quant.
Calendar year							
1971-1973	1	(-)	1	(-)			
1974-1976	1.006	(0.978-1.034)	0.999	(0.972-1.027)			
1977-1979	0.995	(0.968-1.023)	0.947	(0.921-0.973)			
1980-1982	0.923	(0.898-0.948)	0.928	(0.903-0.953)			
1983-1985	0.960	(0.935-0.987)	0.943	(0.918-0.969)			
1986-1988	0.873	(0.849-0.897)	0.870	(0.847-0.894)	1971-1990:	1	(-)
1989-1991	0.822	(0.800-0.844)	0.820	(0.798-0.843)	1991-2007:	0.728	(0.613-0.985)
1992-1994	0.797	(0.775-0.819)	0.796	(0.774-0.817)			
1995-1997	0.738	(0.718-0.758)	0.741	(0.721-0.762)			
1998-2000	0.690	(0.671-0.709)	0.701	(0.682-0.721)			
2001-2003	0.673	(0.655-0.692)	0.670	(0.652-0.689)			
2004-2007	0.616	(0.600-0.632)	0.631	(0.615-0.648)			

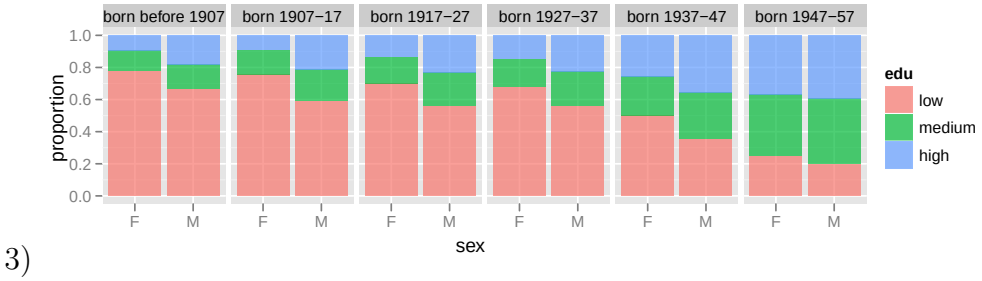
Table A.3: Men - Results of the regression models with period covariates (coefficients in grey are not significant).



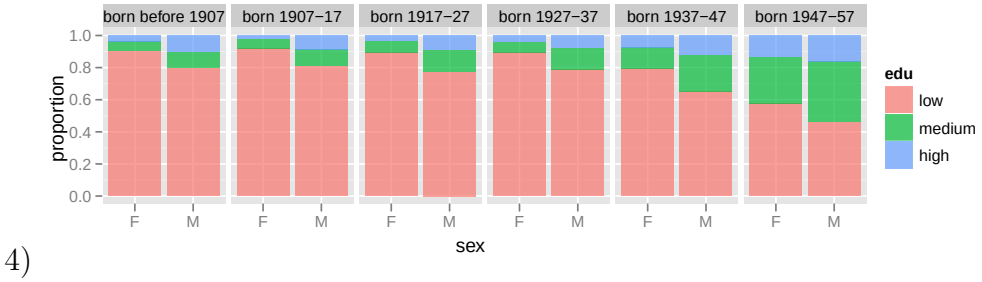
1)



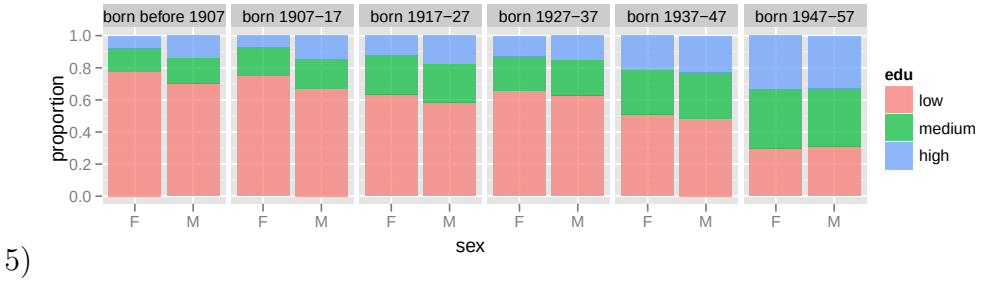
2)



3)



4)



5)

Figure A.1: Distribution of the population by education, birth cohorts and sex (M men and F women) for five macro regions of birth: (1) North-West, 2) North-East, 3) Center, 4) South and Islands, 5) Abroad).

Bibliography

Aalen, O. 1988. "Heterogeneity in survival analysis." *Statistics in medicine* 7(11):1121–1137.

———. 1994. "Effects of frailty in survival analysis." *Statistical Methods in Medical Research* 3(3):227.

Aalen, O., Ø. Borgan and H. Gjessing. 2008a. *Survival and Event History Analysis*, chap. 6: Unobserved heterogeneity: The odd effects of frailty, pp. 231–270. Springer.

———. 2008b. *Survival and event history analysis: a process point of view*. Springer Verlag.

Akaike, H. 1974. "A new look at the statistical model identification." *Automatic Control, IEEE Transactions on* 19(6):716–723.

Aldabe, B., R. Anderson, M. Lyly-Yrjänäinen, A. Parent-Thirion, G. Vermeylen, C. Kelleher and I. Niedhammer. 2011. "Contribution of material, occupational, and psychosocial factors in the explanation of social inequalities in health in 28 countries in Europe." *Journal of Epidemiology and Community Health* 65(12):1123–1131.

Antonovsky, A. 1967. "Social class, life expectancy and overall mortality." *The Milbank Memorial Fund Quarterly* 45(2):31–73.

Baeriswyl, S. 2008. *Etude de l'altération du vieillissement de C. elegans par sa flore intestinale*. Ph.D. thesis.

Barbi, E. 2003. "Assessing the rate of ageing of the human population." *MPIDR Working Papers* .

Beard, R. 1959. "Appendix: Note on some mathematical mortality models." In *Ciba Foundation Symposium- The Lifespan of Animals (Colloquia on Ageing)*, pp. 302–311.

Beckett, M. 2000. "Converging health inequalities in later life-an artifact of mortality selection?" *Journal of Health and Social Behavior* pp. 106–119.

Beckett, M., N. Goldman, M. Weinstein, I. Lin, Y. Chuang et al. 2002. "Social environment, life challenge, and health among the elderly in Taiwan." *Social Science & Medicine* 55(2):191–209.

Ben-Shlomo, Y. and D. Kuh. 2002. "A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives." *International Journal of Epidemiology* 31(2):285–293.

Betensky, R., D. Louis and J. Gregory Cairncross. 2002. "Influence of unrecognized molecular heterogeneity on randomized clinical trials." *Journal of clinical oncology* 20(10):2495.

Biggeri, A., G. Accetta and V. Egidi. 2011. "Evoluzione del profilo di mortalita 30-74 anni per le coorti di nascita dal 1889 al 1968 nelle regioni italiane [Mortality Time Trends 30-74 years by Birth Cohorts 1889-1968 in the Italian Regions]." *Epidemiologia & Prevenzione* 35(5-6):50–67.

Black, D., J. Morris, C. Smith and P. Townsend. 1980. "Inequalities in health: report of a Research Working Group." *London: Department of Health and Social Security* .

Blakely, T. and N. Wilson. 2005. "The contribution of smoking to inequalities in mortality by education varies over time and by sex: two national cohort studies, 1981–84 and 1996–99." *International journal of epidemiology* 34(5):1054–1062.

Bretagnolle, J. and C. Huber-Carol. 1988. "Effects of omitting covariates in Cox's model for survival data." *Scandinavian Journal of Statistics* pp. 125–138.

Brooks, A., G. Lithgow and T. Johnson. 1994. "Mortality rates in a genetically heterogeneous population of Caenorhabditis elegans." *Science* 263(5147):668.

Browning, C. and K. Cagney. 2003. "Moving beyond poverty: neighborhood structure, social processes, and health." *Journal of Health and Social Behavior* pp. 552–571.

Cain, K., S. Harlow, R. Little, B. Nan, M. Yosef, J. Taffe and M. Elliott. 2011. "Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes." *American journal of epidemiology* 173(9):1078–1084.

Cardano, M., G. Costa and M. Demaria. 2004. "Social mobility and health in the Turin longitudinal study." *Social science & medicine* 58(8):1563–1574.

Carey, J., P. Liedo, D. Orozco, J. Vaupel et al. 1992. "Slowing of mortality rates at older ages in large medfly cohorts." *SCIENCE-NEW YORK THEN WASHINGTON-* 258:457–457.

Carey, J., P. Liedo and J. Vaupel. 1995. "Mortality dynamics of density in the Mediterranean fruit fly." *Experimental gerontology* 30(6):605–629.

Caselli, G. and V. Egidi. 1981. "L'analyse des données multidimensionnelles dan l'étude des relations entre mortalité et variable socio-économiques d' envirnment et de comportement individuel [Multivariate methods in the analysis of the relations between mortality and socio-economic, environmental and behavioural variables]." *Genus* 37(3/4):57–91.

Caselli, G., J. Vaupel and A. Yashin. 2000. "Longevity, heterogeneity and selection." *Atti della XL Riunione Scientifica della Società Italiana di Statistica* pp. 49–72.

Cavelaars, A., A. Kunst, J. Geurts, R. Crialesi, L. Grötvedt, U. Helmert, E. Lahelma, O. Lundberg, J. Matheson, A. Mielck et al. 2000. "Educational differences in smoking: international comparison." *Bmj* 320(7242):1102.

Chamberlain, G. 1985. "Heterogeneity, omitted variable bias, and duration dependence." *Longitudinal Analysis of Labor Market Data, ed. JJ Heckman, B. Singer* pp. 3–38.

Christensen, K. and J. Vaupel. 1996. "Determinants of longevity: genetic, environmental and medical factors." *Journal of Internal Medicine* 240(6):333–341.

Coale, A. and E. Kisker. 1986. "Mortality crossovers: Reality or bad data?" *Population Studies* 40(3):389–401.

Cockerham, W. 2007. *Social causes of health and disease*. Polity Pr.

Costa, G., E. Cadum, F. Faggiano, M. Cardano and M. Demaria. 1999. "Social inequalities in the mortality due to cardiovascular diseases in Italy." *Giornale italiano di cardiologia* 29(6):684.

Curtsinger, J., H. Fukui, D. Townsend and J. Vaupel. 1992. "Demography of genotypes: failure of the limited life-span paradigm in Drosophila melanogaster." *Science* 258(5081):461.

d' Errico, A., M. Filippi, M. Demaria, G. Picanza, R. Crialesi, G. Costa, G. Campo and M. Passerini. 2005. "Mortalità per settore produttivo in Italia nel 1992 secondo le storie lavorative INPS [Occupational Mortality in Italy during 1992, assessed through record-linkage between pension records and death certificate]." *Med Lav* 96:s52–s65.

Dalstra, J., A. Kunst, J. Mackenbach et al. 2006. "A comparative appraisal of the relationship of education, income and housing tenure with less than good health among the elderly in Europe." *Social Science & Medicine* 62(8):2046–2060.

Davey Smith, G., D. Gunnel and Y. Ben-Sholmo. 2001. "Life course approaches to socio-economic differentials in cause-specific adult mortality." In *Poverty, Inequality and Health: an International Perspective*, pp. 88–125, Oxford University Press.

Decker, S. and C. Rapaport. 2002. *Medicare and disparities in women's health*. Tech. Rep., National Bureau of Economic Research.

Demaria, M. 2010. "La gestione integrata dei dati demografici e sanitari: lo Studio Longitudinale Torinese [The integrated management of demographic and health data: the Turin Longitudinal Study]."

Demaria, M., M. Cardano and G. Costa. 2001. "La gestione integrata dei dati demografici: l'esempio dello Studio Longitudinale Torinese [The integrated management of demographic data: the example of the Turin Longitudinal Study]."

Doblhammer, G., R. Hoffmann, E. Muth, C. Westphal and A. Kruse. 2009. "A systematic literature review of studies analyzing the effect of sex, age, education, marital status, obesity, and smoking on health transitions." *Demographic Research* 20(5):37–64.

Dor, A., J. Sudano and D. Baker. 2006. "The effect of private insurance on the health of older, working age adults: evidence from the Health and Retirement Study." *Health services research* 41(3p1):759–787.

Drapeau, M., E. Gass, M. Simison, L. Mueller, M. Rose et al. 2000. "Testing the heterogeneity theory of late-life mortality plateaus by using cohorts of Drosophila melanogaster." *Experimental Gerontology* 35:71–84.

Duchateau, L. and P. Janssen. 2008. *The frailty model*. Springer.

Duncan, G., M. Daly, P. McDonough and D. Williams. 2002. "Optimal indicators of socioeconomic status for health research." *American Journal of Public Health* 92(7).

Dupre, M. 2007. "Educational differences in age-related patterns of disease: Reconsidering the cumulative disadvantage and age-as-leveler hypotheses." *Journal of Health and Social Behavior* 48(1):1–15.

Efron, B. 1979. "Bootstrap methods: another look at the jackknife." *The annals of Statistics* 7(1):1–26.

Efron, B. and R. Tibshirani. 1993. *An introduction to the bootstrap*, vol. 57. Chapman & Hall/CRC.

Elo, I., P. Martikainen and K. Smith. 2006. "Socioeconomic differentials in mortality in Finland and the United States: the role of education and income." *European Journal of Population/Revue Européenne de Démographie* 22(2):179–203.

Elo, I. and S. Preston. 1996. "Educational differentials in mortality: United States, 1979-1985." *Social Science & Medicine* 42(1):47–57.

Faggiano, F., E. Versino and P. Lemma. 2001. "Decennial trends of social differentials in smoking habits in Italy." *Cancer Causes and Control* 12(7):665–671.

Ferraro, K. and M. Farmer. 1996. "Double jeopardy, aging as leveler, or persistent health inequality? A longitudinal analysis of white and black Americans." *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 51(6):S319.

Finch, C. 1994. *Longevity, senescence, and the genome*. University of Chicago Press.

Finch, C. and T. Kirkwood. 2000. *Chance, development, and aging*. Oxford University Press, USA.

Gail, M., S. Wieand and S. Piantadosi. 1984. "Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates." *Biometrika* 71(3):431.

Galobardes, B., J. Lynch and G. . 2004. "Childhood socioeconomic circumstances and cause-specific mortality in adulthood: systematic review and interpretation." *Epidemiologic Reviews* 26(1):7–21.

Galobardes, B., M. Shaw, D. Lawlor, J. Lynch and G. Smith. 2006. "Indicators of socioeconomic position (part 1)." *Journal of Epidemiology and Community Health* 60(1):7–12.

Gampe, J. 2010. "Human mortality beyond age 110." In *Supercentenarians*, pp. 219–230, Springer.

Gavrilov, L. and N. Gavrilova. 2001. "The reliability theory of aging and longevity." *Journal of theoretical Biology* 213(4):527–545.

Geyer, S., Ö. Hemström, R. Peter and D. Vågerö. 2006. "Education, income, and occupational class cannot be used interchangeably in social epidemiology. Empirical evidence against a common practice." *Journal of Epidemiology and Community Health* 60(9):804–810.

Goldblatt, P. 1989. "Mortality by social class, 1971-85." *Population Trends* 56:6–15.

Goldman, N. 2001. "Social inequalities in health." *Annals of the New York Academy of Sciences* 954(1):118–139.

Goldman, N., S. Korenman and R. Weinstein. 1995. "Marital status and health among the elderly." *Social Science & Medicine* 40(12):1717–1730.

Golubev, A. 2004. "Does Makeham make sense?" *Biogerontology* 5(3):159–167.

Gompertz, B. 1825. "On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies." *Philosophical transactions of the Royal Society of London* 115:513–583.

Hambright, T. 1969. "Comparison of information on death certificates and matching 1960 census records: age, marital status, race, nativity and country of origin." *Demography* 6(4):413–423.

Hartigan, J. 1969. "Using subsample values as typical values." *Journal of the American Statistical Association* pp. 1303–1317.

———. 1975. "Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values." *The Annals of Statistics* 3(3):573–580.

Hayward, M. and B. Gorman. 2004. "The long arm of childhood: The influence of early-life social conditions on mens mortality." *Demography* 41(1):87–107.

Henderson, R. and P. Oman. 1999. "Effect of frailty on marginal regression estimates in survival analysis." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(2):367–379.

Herd, P. 2006. "Do functional health inequalities decrease in old age?" *Research on Aging* 28(3):375–392.

Herskind, A., M. McGue, N. Holm, T. Sörensen, B. Harvald and J. Vaupel. 1996. "The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870–1900." *Human Genetics* 97(3):319–323.

Hesterberg, T. 2011. "Bootstrap." *Wiley Interdisciplinary Reviews: Computational Statistics* 3(6):497–526.

Hoffmann, R. 2005. "Do socioeconomic mortality differences decrease with rising age?" *Demographic Research* 13(2):35–62.

———. 2008. *Socioeconomic Differences in Old Age Mortality*. Springer.

———. 2011a. "Illness, Not Age, Is the Leveler of Social Mortality Differences in Old Age." *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 66(3):374–379.

———. 2011b. "Socioeconomic inequalities in old-age mortality: A comparison of Denmark and the USA." *Social Science & Medicine* 72(12):1986 – 1992.

Horiuchi, S. 1997. "Postmenopausal acceleration of age-related mortality increase." *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 52(1):B78.

Horiuchi, S. and A. Coale. 1990. "Age patterns of mortality for older women: An analysis using the age-specific rate of mortality change with age." *Mathematical Population Studies* 2(4):245–267.

Horiuchi, S. and J. Wilmoth. 1997. "Age patterns of the life table aging rate for major causes of death in Japan, 1951–1990." *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 52(1):B67.

———. 1998. "Deceleration in the age pattern of mortality at olderages." *Demography* 35(4):391–412.

Hougaard, P. 1984. "Life table methods for heterogeneous populations: distributions describing the heterogeneity." *Biometrika* 71(1):75–83.

———. 1986. "Survival models for heterogeneous populations derived from stable distributions." *Biometrika* 73(2):387–396.

House, J., R. Kessler and A. Herzog. 1990. "Age, socioeconomic status, and health." *The Milbank Quarterly* pp. 383–411.

House, J., J. Lepkowski, A. Kinney, R. Mero, R. Kessler and A. Herzog. 1994. "The social stratification of aging and health." *Journal of Health and Social Behavior* pp. 213–234.

Howards, P., I. Hertz-Picciotto and C. Poole. 2007. "Conditions for bias from differential left truncation." *American journal of epidemiology* 165(4):444–452.

Huisman, M., A. Kunst, O. Andersen, M. Bopp, J. Borgan, C. Borrell, G. Costa, P. Deboosere, G. Desplanques, A. Donkin et al. 2004. "Socioeconomic inequalities in mortality among elderly people in 11 European populations." *Journal of Epidemiology and Community Health* 58(6):468–475.

Huisman, M., A. Kunst and J. Mackenbach. 2003. "Socioeconomic inequalities in morbidity among the elderly; a European overview." *Social Science & Medicine* 57(5):861–873.

Istat. 1990. *La mortalità differenziale secondo alcuni fattori socio-demografici, anni 1981-1982 [Differential Mortality by some Socio-demographic Characteristics, 1981-1982]*. Tech. Rep., Italian National Institute of Statistics, Rome, Italy.

———. 2001. *La mortalità differenziale secondo alcuni fattori socio-demografici, anni 1991-1992 [Differential Mortality by some Socio-demographic Characteristics, 1981-1982]*. Tech. Rep., Italian National Institute of Statistics, Rome, Italy.

Jazwinski, S., S. Kim, C. Lai and A. Benguria. 1998. "Epigenetic stratification: the role of individual change in the biological aging process." *Experimental gerontology* 33(6):571–580.

Johnson, J., W. Stewart, E. Hall, P. Fredlund and T. Theorell. 1996. "Long-term psychosocial work environment and cardiovascular mortality among Swedish men." *American Journal of Public Health* 86(3):324–331.

Karasek, R. and T. Theorell. 1994. *Healthy work: stress, productivity, and the reconstruction of working life*. New York: Basic books.

Kelley-Moore, J. 2010. "Disability and ageing: The social construction of causality." In *The SAGE handbook of social gerontology*, pp. 96–110, Sage Publications Ltd.

Kim, J. and E. Durden. 2007. "Socioeconomic status and age trajectories of health." *Social science & medicine* 65(12):2489–2502.

Kitagawa, E. and P. Hauser. 1973. *Differential mortality in the United States: A study in socioeconomic epidemiology*. Harvard University Press Cambridge, MA.

Klein, J. and M. Moeschberger. 2003. *Survival analysis: Techniques for Censored and Truncated Data*, chap. 3, pp. 63–90. Springer.

Knoops, K., L. de Groot, D. Kromhout, A. Perrin, O. Moreiras-Varela, A. Menotti and W. van Staveren. 2004. "Mediterranean diet, lifestyle factors, and 10-year mortality in elderly European men and women." *JAMA: the journal of the American Medical Association* 292(12):1433.

Kohler, H. and I. Kohler. 2000. "Frailty modelling for adult and old age mortality: The application of a modified DeMoivre hazard function to sex differentials in mortality." *Demographic Research* 3(8):32.

Koskinen, S., K. Joutsenniemi, T. Martelin and P. Martikainen. 2007. "Mortality differences according to living arrangements." *International journal of epidemiology* 36(6):1255–1264.

Koskinen, S. and T. Martelin. 1994. "Why are socioeconomic mortality differences smaller among women than among men?" *Social Science & Medicine* 38(10):1385–1396.

Krieger, J. and D. Higgins. 2002. "Housing and health: time again for public health action." *American journal of public health* 92(5):758.

Krieger, N., D. Williams and N. Moss. 1997. "Measuring social class in US public health research: concepts, methodologies, and guidelines." *Annual Review of Public Health* 18(1):341–378.

Kristenson, M., Z. Kucinskiene, B. Bergdahl and K. Orth-Gomér. 2001. "Risk factors for coronary heart disease in different socioeconomic groups of Lithuania and Swedenthe LiVicordia study." *Scandinavian journal of public health* 29(2):140–150.

Kunst, A., F. Groenhof, J. Mackenbach and D. Leon. 1998a. "Occupational class and cause specific mortality in middle aged men in 11 European countries: comparison of population based studiesCommentary: Unequal inequalities across Europe." *Bmj* 316(7145):1636–1642.

Kunst, A., F. Groenhof, J. Mackenbach et al. 1998b. "Mortality by occupational class among men 30-64 years in 11 European countries." *Social science & medicine* 46(11):1459–1476.

Laaksonen, M., R. Prättalä and A. Karisto. 2001. "Patterns of unhealthy behaviour in Finland." *The European Journal of Public Health* 11(3):294–300.

Lancaster, T. 1979. "Econometric methods for the duration of unemployment." *Econometrica: Journal of the Econometric Society* pp. 939–956.

Lauderdale, D. 2001. "Education and survival: Birth cohort, period, and age effects." *Demography* 38(4):551–561.

Leombruni, R., M. Richiardi and G. Costa. 2008. "Aspettative di vita, lavori usuranti e equità del sistema previdenziale. Prime evidenze dal Work Histories Italian Panel [Life expectancies, physically demanding occupations and pension system equity. First evidences from the Work Histories Italian Panel]." *Laboratorio R. Revelli, Collegio Carlo Alberto, Working Paper* (75).

Liang, J., J. Bennett, N. Krause, E. Kobayashi, H. Kim, J. Brown, H. Akiyama, H. Sugisawa and A. Jain. 2002. "Old age mortality in Japan." *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 57(5):S294.

Lin, D., B. Psaty and R. Kronmal. 1998. "Assessing the sensitivity of regression results to unmeasured confounders in observational studies." *Biometrics* pp. 948–963.

Lopez, A., N. Collishaw and T. Piha. 1994. "A descriptive model of the cigarette epidemic in developed countries." *Tobacco control* 3(3):242.

Lumey, L., A. Stein, H. Kahn, K. van der Pal-de Bruin, G. Blauw, P. Zybert, E. Susser et al. 2007. "Cohort profile: the Dutch Hunger Winter families study." *International journal of epidemiology* 36(6):1196.

Luy, M., P. Di Giulio and G. Caselli. 2011. "Differences in life expectancy by education and occupation in Italy, 1980–94: Indirect estimates from maternal and paternal orphanhood." *Population Studies* 65(2):137–155.

Lynch, S. 2003. "Cohort and life-course patterns in the relationship between education and health: A hierarchical approach." *Demography* 40(2):309–331.

Mackenbach, J., V. Bos, O. Andersen, M. Cardano, G. Costa, S. Harding, A. Reid, Ö. Hemström, T. Valkonen and A. Kunst. 2003. "Widening socioeconomic inequalities in mortality in six Western European countries." *International journal of epidemiology* 32(5):830.

Mackenbach, J., E. Commission and C. C. A. G. Britain. 2006. *Health inequalities: Europe in profile*. Produced by COI for the Dept. of Health.

Mackenbach, J., M. Huisman, O. Andersen, M. Bopp, J. Borgan, C. Borrell, G. Costa, P. Deboosere, A. Donkin, S. Gadeyne et al. 2004. "Inequalities in lung cancer mortality by the educational level in 10 European populations." *European Journal of Cancer* 40(1):126–135.

Mackenbach, J., A. Kunst, A. Cavelaars, F. Groenhof, J. Geurts et al. 1997. "Socioeconomic inequalities in morbidity and mortality in western Europe." *The Lancet* 349(9066):1655–1659.

Mackenbach, J., A. Kunst, F. Groenhof, J. Borgan, G. Costa, F. Faggiano, P. Jozan, M. Leinsalu, P. Martikainen, J. Rychtarikova et al. 1999. "Socioeco-

nomic inequalities in mortality among women and among men: an international study." *American Journal of Public Health* 89(12):1800.

Mackenbach, J., I. Stirbu, A. Roskam, M. Schaap, G. Menvielle, M. Leinsalu and A. Kunst. 2008. "Socioeconomic inequalities in health in 22 European countries." *New England Journal of Medicine* 358(23):2468–2481.

Mackenbach, J., K. Stronks and A. Kunst. 1989. "The contribution of medical care to inequalities in health: differences between socio-economic groups in decline of mortality from conditions amenable to medical intervention." *Social Science & Medicine* 29(3):369–376.

Makeham, W. 1860. "On the law of mortality and the construction of annuity tables." *The Assurance Magazine, and Journal of the Institute of Actuaries* 8(6):301–310.

Mamo, C., A. dErrico, O. Pasqualini, S. Bruzzone and M. Dalmasso. 2005a. "La mortalità per professioni nella popolazione occupata al Censimento 1991 [Occupational Mortality in the population emplyed at the 1991 Census]." *Med Lav* 96:s39–s51.

Mamo, C., C. Marinacci, M. Demaria, D. Mirabelli and G. Costa. 2005b. "Factors other than risks in the workplace as determinants of socioeconomic differences in health in Italy." *International journal of occupational and environmental health* 11(1):70–76.

Manly, B. 1997. *Randomization, bootstrap and Monte Carlo methods in biology*. Chapman & Hall/CRC.

Manton, K. 1992. "Mortality and life expectancy changes among the oldest old." In *The Oldest Old*, pp. 157–182, Oxford University Press.

Manton, K., E. Stallard and J. Vaupel. 1981a. "Methods for comparing the mortality experience of heterogeneous populations." *Demography* 18(3):389–410.

———. 1986. "Alternative models for the heterogeneity of mortality risks among the aged." *Journal of the American Statistical Association* pp. 635–644.

Manton, K., E. Stallard et al. 1981b. "Methods for evaluating the heterogeneity of aging processes in human populations using vital statistics data: explaining the black/white mortality crossover by a model of mortality selection." *Human biology; an international record of research* 53(1):47.

Manton, K., M. Woodbury and E. Stallard. 1995. "Sex differences in human mortality and aging at late ages: the effect of mortality selection and state dynamics." *The Gerontologist* 35(5):597.

Marinacci, C., T. Spadea, A. Biggeri, M. Demaria, A. Caiazzo and G. Costa. 2004. "The role of individual and contextual socioeconomic circumstances on mortality: analysis of time variations in a city of north west Italy." *Journal of epidemiology and community health* 58(3):199–207.

Marmot, M. and M. McDowall. 1986. "Mortality decline and widening social inequalities." *The Lancet* 328(8501):274–276.

Marmot, M. and M. Shipley. 1996. "Do socioeconomic differences in mortality persist after retirement? 25 year follow up of civil servants from the first Whitehall study." *Bmj* 313(7066):1177–1180.

Marmot, M., M. Shipley and G. Rose. 1984. "Inequalities in death–specific explanations of a general pattern?" *The Lancet* 323(8384):1003–1006.

Martelin, T. 1994. "Mortality by indicators of socioeconomic status among the Finnish elderly." *Social Science & Medicine* 38(9):1257–1278.

Martikainen, P., T. Valkonen and T. Martelin. 2001. "Change in male and female life expectancy by social class: decomposition by age and cause of death in Finland 1971–95." *Journal of Epidemiology and Community Health* 55(7):494.

Martikainen, P., T. Valkonen and H. Moustgaard. 2009. "The effects of individual taxable income, household taxable income, and household disposable income on mortality in Finland, 1998–2004." *Population Studies* 63(2):147–162.

Materia, E., L. Cacciani, G. Bugarini, G. Cesaroni, M. Davoli, M. Mirale, L. Vergine, G. Baglio, G. Simeone and C. Perucci. 2005. "Income inequality and mortality in Italy." *The European Journal of Public Health* 15(4):411–417.

McDonough, P., G. Duncan, D. Williams and J. House. 1997. "Income dynamics and adult mortality in the United States, 1972 through 1989." *American journal of public health* 87(9):1476–1483.

McGilchrist, C. and C. Aisbett. 1991. "Regression with frailty in survival analysis." *Biometrics* pp. 461–466.

McGue, M., J. Vaupel, N. Holm and B. Harvald. 1993. "Longevity is moderately heritable in a sample of Danish twins born 1870–1880." *Journal of Gerontology* 48(6):B237.

McMunn, A., J. Nazroo and E. Breeze. 2009. "Inequalities in health at older ages: a longitudinal investigation of the onset of illness and survival effects in England." *Age and ageing* 38(2):181.

Merler, E., A. Benvenuti, P. Baldi, M. Nardulli, L. Olmastroni, A. Biggeri, M. Marchi and E. Dreassi. 1999. "Socioeconomic inequalities in health in the Tuscany Longitudinal Study (SLTO): persistence and changes over time in overall mortality and selected causes (lung cancer, liver cirrhosis, AIDS and overdose)]." *Epidemiologia e prevenzione* 23(3):207.

Michelozzi, P., C. Perucci, F. Forastiere, D. Fusco, C. Ancona and V. Dell'Orco. 1999. "Inequality in health: socioeconomic differentials in mortality in Rome, 1990-95." *Journal of epidemiology and community health* 53(11):687–693.

Mirowsky, J. and C. Ross. 2003. *Education, social status, and health*. Aldine de Gruyter.

Missov, T. and M. Finkelstein. 2011. "Admissible mixing distributions for a general class of mixture survival models with known asymptotics." *MPIDR Working Papers, WP 2011-004* .

Nakamura, E. and K. Miyao. 2008. "Sex differences in human biological aging." *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 63(9):936.

Oakes, J. and P. Rossi. 2003. "The measurement of SES in health research: current practice and steps toward a new approach." *Social Science & Medicine* 56(4):769–784.

Olausson, P. 1991. "Mortality among the elderly in Sweden by social class." *Social science & medicine* 32(4):437–440.

Pahl, J. 1990. "Household spending, personal spending and the control of money in marriage." *Sociology* 24(1):119–138.

Painter, R., T. Roseboom and O. Bleker. 2005. "Prenatal exposure to the Dutch famine and disease in later life: an overview." *Reproductive Toxicology* 20(3):345–352.

Pamuk, E. 1985. "Social class inequality in mortality from 1921 to 1972 in England and Wales." *Population studies* 39(1):17–31.

Pattengale, N., M. Alipour, O. Bininda-Emonds, B. Moret and A. Stamatakis. 2010. "How many bootstrap replicates are necessary?" *Journal of Computational Biology* 17(3):337–354.

Pearlin, L. 1989. "The sociological study of stress." *Journal of health and social behavior* pp. 241–256.

Pearlin, L., S. Schieman, E. Fazio and S. Meersman. 2005. "Stress, health, and the life course: Some conceptual perspectives." *Journal of Health and Social Behavior* 46(2):205–219.

Pensola, T. and T. Valkonen. 2002. "Effect of parental social class, own education and social class on mortality among young men." *The European Journal of Public Health* 12(1):29–36.

Petrelli, A., R. Gnavi, C. Marinacci and G. Costa. 2006. "Socioeconomic inequalities in coronary heart disease in Italy: a multilevel population-based study." *Social Science & Medicine* 63(2):446–456.

Phillimore, P., A. Beattie and P. Townsend. 1994. "Widening inequality of health in northern England, 1981-91." *Bmj* 308(6937):1125–1128.

Politis, D. and J. Romano. 1994. "Large sample confidence regions based on subsamples under minimal assumptions." *The Annals of Statistics* 22(4):2031–2050.

Preston, S. and I. Elo. 1995. "Are educational differentials in adult mortality increasing in the United States?" *Journal of Aging and Health* 7(4):476–496.

Preston, S., I. Elo, I. Rosenwaike and M. Hill. 1996. "African-American mortality at older ages: Results of a matching study." *Demography* 33(2):193–209.

Psn. 2011. "Programma Statistico Nazionale 2011-2013 (Aggiornamento 2012-2013)[National Statistical Program 2011-2013 (Update 2012-2013)]." URL `http://www.sistan.it/psn/aggiornamento2012-2013.htm`.

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

Rahkonen, O., M. Laaksonen, P. Martikainen, E. Roos and E. Lahelma. 2006. "Job control, job demands, or social class? The impact of working conditions on the relation between social class and health." *Journal of epidemiology and community health* 60(1):50–54.

Reynolds, R., S. Temiyasathit, M. Reedy, E. Ruedi, J. Drnevich, J. Leips and K. Hughes. 2007. "Age specificity of inbreeding load in Drosophila melanogaster and implications for the evolution of late-life mortality plateaus." *Genetics* 177(1):587.

Riggs, J. 1990. "Longitudinal Gompertzian analysis of adult mortality in the US, 1900-1986." *Mechanisms of Ageing and Development* 54(3):235–247.

Rodriguez, G. 1994. "Statistical issues in the analysis of reproductive histories using hazard models." *Annals of the New York academy of sciences* 709(1):266–279.

Rosamond, W., H. Tyroler, L. Chambless, A. Folsom, L. Cooper and D. Conwill. 1997. "Educational achievement recorded on certificates of death compared with self-report." *Epidemiology* pp. 202–204.

Rose, G. and M. Marmot. 1981. "Social class and coronary heart disease." *British heart journal* 45(1):13–19.

Rose, M., M. Drapeau, P. Yazdi, K. Shah, D. Moise, R. Thakar, C. Rauser and L. Mueller. 2002. "Evolution of late-life mortality in Drosophila melanogaster." *Evolution* 56(10):1982–1991.

Ross, C. and C. Wu. 1996. "Education, age, and the cumulative advantage in health." *Journal of Health and Social Behavior* pp. 104–120.

Rosso, S., F. Faggiano, R. Zanetti and G. Costa. 1997. "Social class and cancer survival in Turin, Italy." *Journal of epidemiology and community health* 51(1):30–34.

Schmoor, C. and M. Schumacher. 1997. "Effects of covariate omission and categorization when analysing randomized trials with the Cox model." *Statistics in medicine* 16(3):225–237.

Schrijvers, C., H. van de Mheen, K. Stronks and J. Mackenbach. 1998. "Socioeconomic inequalities in health in the working population: the contribution of working conditions." *International Journal of Epidemiology* 27(6):1011.

Schumacher, M., M. Olschewski and C. Schmoor. 1987. "The impact of heterogeneity on the comparison of survival times." *Statistics in medicine* 6(7):773–784.

Selye, H. 1985. "The nature of stress." URL `http://www.icnr.com/articles/the-nature-of-stress.html`.

Shai, D. and I. Rosenwaike. 1989. "Errors in reporting education on the death certificate: some findings for older male decedents from New York State and Utah." *American journal of epidemiology* 130(1):188–192.

Shock, N. 1967. "Physical activity and the" rate of ageing"." *Canadian Medical Association Journal* 96(12):836.

Smith, G., E. Brunner et al. 1997. "Socio-economic differentials in health: the role of nutrition." *Proceedings of the Nutrition Society* 56(1):75–90.

Smith, J. 1999. "Healthy bodies and thick wallets: the dual relation between health and economic status." *The Journal of Economic Perspectives* 13(2):145–166.

Sorlie, P. and N. Johnson. 1996. "Validity of education information on the death certificate." *Epidemiology* pp. 437–439.

Spadea, T., A. d'Errico, M. Demaria, F. Faggiano, S. Pasian, R. Zanetti, S. Rosso, P. Vicari and G. Costa. 2009. "Educational inequalities in cancer incidence in Turin, Italy." *European Journal of Cancer Prevention* 18(3):169.

Spadea, T., N. Zengarini, A. Kunst, R. Zanetti, S. Rosso and G. Costa. 2010. "Cancer risk in relationship to different indicators of adult socioeconomic position in Turin, Italy." *Cancer Causes and Control* 21(7):1117–1130.

Sparén, P., D. Vågerö, D. Shestov, S. Plavinskaja, N. Parfenova, V. Hoptiar, D. Paturot and M. Galanti. 2004. "Long term mortality after severe starvation during the siege of Leningrad: prospective cohort study." *bmj* 328(7430):11.

Stanner, S., K. Bulmer, C. Andres, O. Lantseva, V. Borodina, V. Poteen and J. Yudkin. 1997. "Does malnutrition in utero determine diabetes and coronary heart disease in adulthood? Results from the Leningrad siege study, a cross sectional study." *Bmj* 315(7119):1342–1348.

Stata Release 12. 2012. *Subject Specific: Survival Analysis and Epidemiological Tables Reference Manual*.

Steinsaltz, D. and S. Evans. 2004. "Markov mortality models: implications of quasistationarity and varying initial distributions." *Theoretical Population Biology* 65(4):319–337.

Steinsaltz, D. and K. Wachter. 2006. "Understanding mortality rate deceleration and heterogeneity." *Mathematical Population Studies* 13(1):19–37.

Strehler, B. and A. Mildvan. 1960. "General theory of mortality and aging. A stochastic model relates observations on aging, physiologic decline, mortality, and radiation." *Science See Saiensu* 132.

Stringhini, S., S. Sabia, M. Shipley, E. Brunner, H. Nabi, M. Kivimaki and A. Singh-Manoux. 2010. "Association of socioeconomic position with health behaviors and mortality." *JAMA: the journal of the American Medical Association* 303(12):1159.

Thoits, P. 1995. "Stress, coping, and social support processes: Where are we? What next?" *Journal of health and social behavior* pp. 53–79.

Trussell, J. and G. Rodriguez. 1990. "Heterogeneity in demographic research." In *Convergent issues in genetics and demography*, Oxford University Press, USA.

Tsimbos, C. 2010. "An assessment of socio-economic inequalities in health among elderly in Greece, Italy and Spain." *International journal of public health* 55(1):5–15.

Vallin, J. 1980. "Socio-economic determinants of mortality in industrialized countries." *Population bulletin of the United Nations* (13):26.

Van den Berg, G. J. and B. Drepper. 2011. "Inference for Shared-Frailty Survival Models with Left-Truncated Data." *IZA Discussion Paper No. 6031* .

Van Doorslaer, E., C. Masseria, X. Koolman et al. 2006. "Inequalities in access to medical care by income in developed countries." *Canadian medical association journal* 174(2):177.

van Raalte, A., A. Kunst, P. Deboosere, M. Leinsalu, O. Lundberg, P. Martikainen, B. Strand, B. Artnik, B. Wojtyniak and J. Mackenbach. 2011. "More variation in lifespan in lower educated groups: evidence from 10 European countries." *International journal of epidemiology* 40(6):1703–1714.

Vaupel, J. 2010. "Biodemography of human ageing." *Nature* 464(7288):536–542.

Vaupel, J., J. Carey, K. Christensen, T. Johnson, A. Yashin, N. Holm, I. Iachine, V. Kannisto, A. Khazaeli, P. Liedo et al. 1998. "Biodemographic trajectories of longevity." *Science* 280(5365):855.

Vaupel, J., T. Johnson, G. Lithgow, J. Curtsinger, H. Fukui, L. Xiu, A. Khazaeli, S. Pletcher, J. Wang, H. Muller et al. 1994. "Rates of mortality in populations of Caenorhabditis elegans." *Science* 266:826–828.

Vaupel, J., K. Manton and E. Stallard. 1979. "The impact of heterogeneity in individual frailty on the dynamics of mortality." *Demography* 16(3):439–454.

Vaupel, J. and A. Yashin. 1983. *The deviant dynamics of death in heterogeneous populations*. International Institute for Applied Systems Analysis.

———. 1985. "Heterogeneity's ruses: some surprising effects of selection on population dynamics." *American statistician* pp. 176–185.

Vogler, C. and J. Pahl. 1994. "Money, power and inequality within marriage." *The Sociological Review* 42(2):263–288.

Wachter, K. 2003. "Hazard curves and lifespan prospects." *Population and development review* 29(03):270.

Waldron, I. 1985. "What do we know about causes of sex differences in mortality? A review of the literature." *Population Bulletin of the United Nations* (18):59.

Weitz, J. and H. Fraser. 2001. "Explaining mortality rate plateaus." *Proceedings of the National Academy of Sciences* 98(26):15383.

Wienke, A. 2010. *Frailty models in survival analysis*, vol. 37. Chapman & Hall/CRC.

Wilkinson, R. 1992. "Income distribution and life expectancy." *British medical journal* 304(6820):165–168.

———. 1996. *Unhealthy societies: the afflictions of inequality*. Psychology Press.

———. 1997. "Socioeconomic determinants of health: Health inequalities: relative or absolute material standards?" *Bmj* 314(7080):591.

Wing, S., K. Manton, E. Stallard, C. Hames and H. Tryoler. 1985. "The black-/white mortality crossover: investigation in a community-based study." *Journal of Gerontology* 40(1):78.

Xue, X. and R. Brookmeyer. 1996. "Bivariate frailty model for the analysis of multivariate survival time." *Lifetime Data Analysis* 2(3):277–289.

Yang, Y. and M. Kozloski. 2011. "Sex Differences in Age Trajectories of Physiological Dysregulation: Inflammation, Metabolic Syndrome, and Allostatic Load." *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 66(5):493.

Yashin, A., I. Iachine, A. Begun and J. Vaupel. 2001. "Hidden frailty: myths and reality." Research Report 34, Department of Statistics and Demography, Odense University.

Zajacova, A., N. Goldman and G. Rodriguez. 2009. "Unobserved heterogeneity can confound the effect of education on mortality." *Mathematical Population Studies* 16(2):153–173.

Zarulli, V., D. Jasilionis and D. Jdanov. 2012. "Changes in educational differentials in old-age mortality in Finland and Sweden between 1971-1975 and 1996-2000." *Demographic Research* 26(19).