1-26-2015

Longitudinal Analysis of Mortality Risk Factors for Actuarial Valuation

Ushani D. Kariyawasam
University of Connecticut - Storrs, ushanidias@yahoo.com

Longitudinal Analysis of Mortality Risk Factors for Actuarial Valuation

Ushani Dias Kariyawasam Majuwana Gamage, Ph.D.

University of Connecticut, 2015

ABSTRACT

Today, people are living longer and the world population is getting older. Recent statistics indicate that a 65-year-old female in the United States is estimated to live to 88.8 years old, while a 65-year-old male to 86.6 years old. This translates to about a two-year increase in life expectancy from birth compared to that more than a decade ago. Understanding these trends and their potential impact are ever more relevant to the insurance industry. In this thesis, we emphasize the longitudinal modeling framework of pension and long term care insurance using some advanced techniques to analyze patterns and trends in longevity and to investigate potential covariates to further characterize the nature of the risk. Using data from the Health and Retirement Study (HRS), our work finds that factors that incorporate demographic, health, lifestyle, and financial information help improve model projections of mortality. We used multiple state framework to develop models for understanding the utilization of long term care. Some key findings indicate that female tends to be more vulnerable to exposure for long term care needs, and so with low educated people. Finally, motivated by the data obtained from an insurer, this thesis also examined the effect of policy termination on the survival of policyholders with life insurance contracts.

We modeled the time until a policy lapses and its subsequent mortality pattern and found some evidence of mortality selection. We subsequently examined the financial cost of policy termination. The lack of available data precluded us from extending this analysis to pension plans and long term care insurance products; such can be done as further studies.

Longitudinal Analysis of Mortality Risk Factors for Actuarial Valuation

Ushani Dias Kariyawasam Majuwana Gamage

M.S. University of Connecticut, 2008

B.S. University of Sri Jayewardenepure, 2005

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2015

APPROVAL PAGE

Doctor of Philosophy Dissertation

Longitudinal Analysis of Mortality Risk Factors for Actuarial Valuation

Presented by

Ushani Dias Kariyawasam Majuwana Gamage, B.S., M.S.

Major Advisor _____
Professor Emiliano A. Valdez

Associate Advisor _____
Professor James G. Bridgeman

Associate Advisor _____
Professor Brian M. Hartman

University of Connecticut

2015

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to those exceptional individuals who had contributed in many ways, and extended unstinted support, from the inception of my efforts on this thesis.

Top and foremost, I would like to express sincere and heartfelt gratitude to my major advisor, Professor Emiliano A. Valdez, for his continued and untiring support towards my Ph.D research. He imparted immense knowledge, gave me excellent guidance, with unwavering patience and enthusiasm. His faith in my ability throughout this process, had been extremely helpful to accomplish this formidable task. Professor Valdez is a true inspiration for me.

Besides my major advisor, I owe my deepest gratitude to Professors James H. Bridgeman and Brian M. Hartman for being extraordinary committee members who were more than generous with their expertise and precious time. In addition, I could not be more than thankful enough to Professor Jeyaraj Vadiveloo who inspired me and suggested to do the work related to pension plans and long term care. His guidance and generosity is deeply appreciated.

Next on my list to thank is the University of Connecticut Department of Mathematics together with its Quantitative Learning Center for generously providing me financial assistance throughout my graduate studies. Special thanks go to Professors Thomas Roby, Sarah Frey, David Gross, Reed Solomon, Charles Vinsonhaler and James Bridgeman for arranging the financial support especially at times when I

needed it most.

I am also deeply humbled by the help during the completion of my Ph.D. extended to me by several other people connected with the Department of Mathematics. Monique Roy, Tammy Prentice, and Cara Light; were at the top of the list of those who supported and assisted me.

I am indebted to the many friendly and bright colleagues in the department and the University who also extended support in many indirect ways. Here I want to acknowledge the following people: Shirani, Asiri, Milanthi, Bernardo, Gao, Rozita, David, Ji, Sudath, Chandrika, Priyanga, Amali, and Januka.

I would like to thank members of my family for their endless love and support throughout my life, and especially their understanding during the time of completing my dissertation. There was certainly a limitation on time available, to love and share. Special thanks go to my mother Lily Dias and my father Nimal Dias, sisters Dulani, Priyani, and brother Kasun. Not forgotten are my brother-in law Dimuthu, my beloved grandmother Aalen, uncle Justin, and of course, my in-laws.

Forever, I will be thankful to Mr. Sardha J. Rasaputra for his generous support to complete my B.S. degree, and all other unconditional support given throughout; in memory of his beloved father, late Mr. Buddhadasa Rasaputram.

A special thank goes to William J. Thompson, Stephen J. Kaczmarek, and Andrea Sheldon for giving me the opportunity to work at Milliman Hartford Health and to help secure my actuarial career.

Last but not the least, I would like to thank my husband, Priyantha, for his encouragement, patience, love and support, while going through this difficult process during a period when he too, was racing against time, to complete his own thesis. Special thanks to him for preparing delicious meals of my choice, while I spent time

on research, to write my thesis.

Contents

Chapter 1

Introduction

The society is living longer and world population is getting older than ever. Within the last century, significant medical progress such as the pioneering discoveries of anesthesia, anesthetics and insulin, socio-demographic changes such as educational and technological developments, improvements in lifestyles such as diet and exercises, the absence of major pandemic crises like cholera and the Black Death and probably a combination of these factors, contributed to a significant reduction in mortality across several countries in the world. As shown in the Figure 1.0.1, due to extraordinary improvements in life expectancy, world elderly population live longer at their retirement phase.

In the United States, the country expects a rapid growth in the retired population from 2010 to 2050 as there will now the aging of the "baby boomers generation"; this generation consists of those born post-World War II, i.e. years between 1946 and 1964, who will now enter the older age population. According to the recently published Wall Street Journal article, Fitzpatrick (2014), which is based on 2014 Society of Actuaries

findings about lifespan of 65+ people in the United States, a 65-year-old female is estimated to live to 88.8 years old, and similarly, a 65-year-old male is estimated to live to 86.6 years old. This translates about a two-year increase in life expectancy from birth when compared to year 2000 estimates.

As we go through these phases, understanding trends in mortality rates are ever more relevant in the insurance industry. As Americans are expected to spend more time in retirement years, the allocation of economic resources associated with aging becomes even more important. Pension and long term care insurance are two crucial resources for the elderly population. In this dissertation for which we titled "Longitudinal Analysis of Mortality Risk Factors for Actuarial Valuation", we emphasize the longitudinal modeling framework of pension and long term care insurance using some more advanced techniques and we investigate many potential covariates which can be used to further characterize longevity risk. Lapse or termination for pension and long term care is also an important aspect that our empirical data used in our investigation did not have. However, to supplement our work and motivated by the available life insured data we obtained from an insurance company, this dissertation examined the effect of policy termination on the survival of policyholders with life insurance contracts. If data is available, similar analysis can be done for pension plans and long term care insurance products.

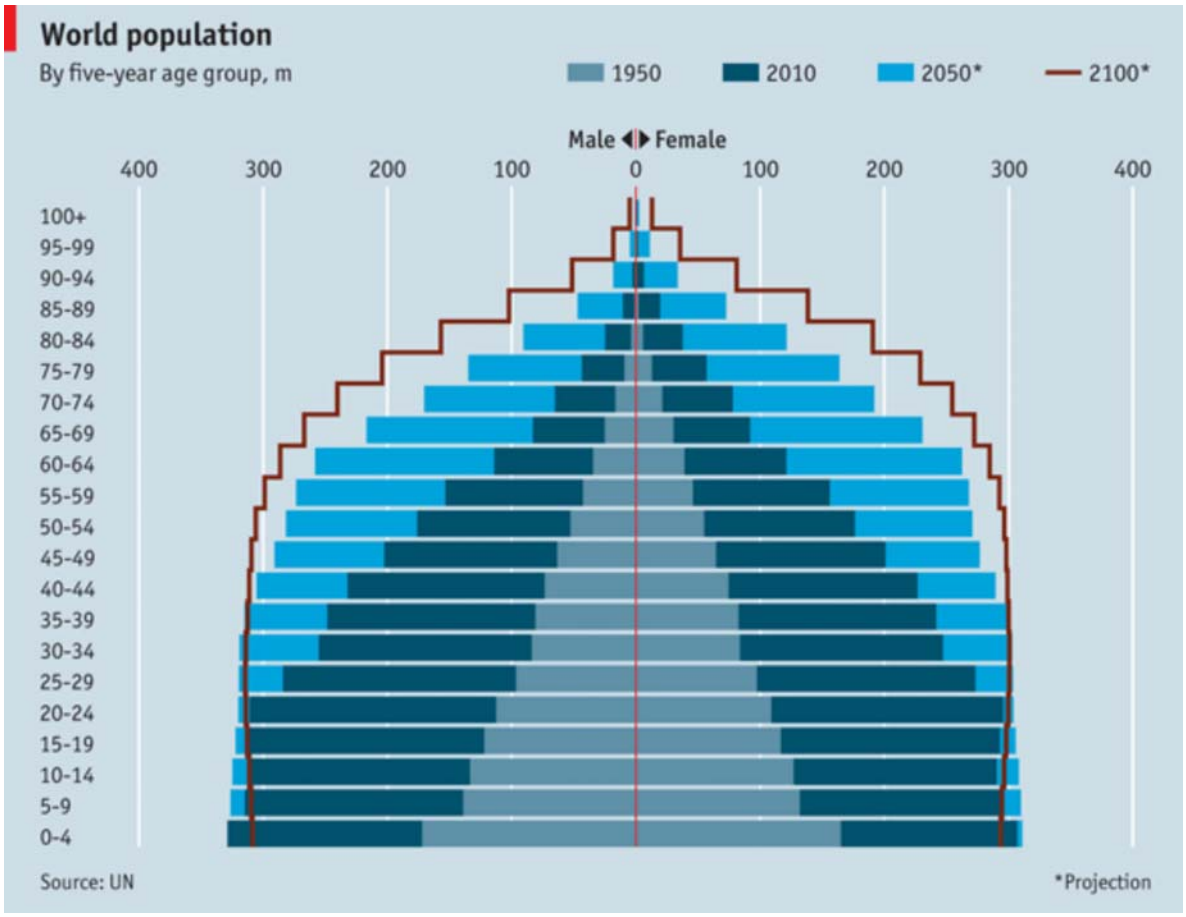


FIGURE 1.0.1: World Population 1950 - 2100

Several research works are motivated to understand the remarkable gain in life expectancy within the context of demographic, economic, lifestyle, and health effects. These analyses led several mortality studies to focus on extending explanatory variables for survival other than simply age and gender. It is widely accepted that both age and gender play a key and foremost role in understanding mortality or survival effects. These developments on mortality analysis are advanced by researches in multidisciplinary areas encompassing epidemiology, psychology, socio-economics, and biology for which may of their findings provide some crucial predictive variables

for mortality. To illustrate, educational level, marital status, income, smoking status, alcohol intake, occupation, body mass index, self-rated health, exercise, health indicators, and even genetic influence are now some of the powerful variables that help us better understand mortality. These effects have beyond the traditional age and gender effects that have appeared in the literature; please refer Brown and McDaid (2003), Crimmins (2011) and U.S.Census Bureau (2014).

Education and incomes are two most significant predictors for mortality, health, and even longevity. Several studies confirmed that low income and low educated people are more likely to die than high income and better educated people; see Pappas et al. (1993) and Rogers, Hummer, and Nam (1999). Due to the extensive association with mortality, Sorlie, Backlund, and Keller (1995) suggested to include education, income, occupation, race, and marital status for any morbidity and mortality studies.

In addition to, the impact of alcohol consumption on mortality have been examined by different researchers. Findings in this area are conflicting. For example, based on the middle-aged and elderly population in the United States, Thun et al. (1997) found that a moderate drinker who had up to one or two alcohol drinks daily had a lower mortality rates than nondrinkers. People with low cardiovascular risk, alcohol consumption and mortality rates show a J-shaped relationship which indicates that light to moderate drinkers have lower mortality than none or heavy drinkers. The relationship between body mass index and mortality is another debatable area. Some studies show hazard rates increase with body weights as in Allison et al. (1999), and in contrast, some studies show the reverse relationship as in Kalantar-Zadeh et al. (2003). It appears that this controversy requires further investigation.

Throughout this dissertation we are going to use these extended variables within the context of both pension plans and long term care. This dissertation is organized

as follows. This chapter describes briefly the purposes of the thesis.

Chapter 2 provides an overview of Health and Retirement Study (HRS) data, how several research analysts interested in understanding survivorship continue to use this rich data, and how we choose to use the HRS data for for our pension plan and long term care studies. HRS is a most advanced national resource providing longitudinal information about aging in the United States. This chapter explores the importance of HRS data, how it has been utilized in several disciplines, and its suitability for our purposes. It is important to note that this data is from an excellent national representation of the general population, and not necessary a reflection of insured data. Selection is a common characteristics of insured data so that some deviations about the insured population from the general population may become possible. Some other information about insured population can be used for pension and long term care modeling. Finally, as described in Chapter 2, the HRS data has been widely used in many disciplines including, but not limit to, economics, finance, health economics, medical statistics, and actuarial science. It is not surprising to see that some of the findings resulting from using the HRS data can have reverberating policy implications.

A detailed description of survival models for pension liabilities are dealt with in Chapter 3. In this chapter, we examined a set of key areas such as health, demographic, lifestyle, and financial factors that affect the health of individuals and thereby presumably their survival patterns. For modeling purposes, we find the use of Cox regression model, which in some sense extends the usual age-gender factor model, to project future mortality. We calibrated the Cox regression model and used these estimates to project just some simple pension liabilities to illustrate its usefulness.

Chapter 4 focuses on the long term care insurance, including the use of multiple

state modeling within a longitudinal framework. In this chapter, to understand the use of long term care services among the elderly population, we find the usefulness of the rich and comprehensive data obtained from HRS. For illustrative proposes, we calibrated multiple state models, one where only age and gender were used and another one where extended time varying covariates were used, in order to understand the implications of various factors in the use of long term care services. A state structure which is fairly similar to the standard state structure of illness-death or disability models has used. The use of long term care were examined by looking at the number of difficulties to perform Activities of Daily Living (ADL) with death considered as the absorbing state.

Chapter 5 provides a discussion of the life insurance policy termination and survivorship. This chapter focuses on policy termination together with understanding the survivorship pattern resulting from terminated policies. Parametric models based on Generalized Gamma Distributions, GB2 Distributions, and Log-Normal Distributions are considered for time-until-withdrawal and while we have examined several survival models for the age at death, the two models finally used which were both commonly known to actuaries are the Gompertz and the Weibull distributions. This chapter is published in *Insurance: Mathematics and Economics*, vol. 58, 2014, pp. 138-149.

Finally, some concluding remarks are added in Chapter 6.

Chapter 2

The Health and Retirement Study

2.1 Introduction

The Health and Retirement Study (HRS) is a prospective national longitudinal study about the health, retirement, economic, and other aspects of Americans who are over 50 years of age. It is one of the most advanced biennial national resources providing information about aging in the United States since 1992. Although the observations are sampled from the population, it provides an excellent national representation of the general population. As detailed in Juster and Suzman (1995), well-funded and designed HRS's origin has been created to recognize the antecedents and consequences of growing retirement population in the United States. This is a collaborative work between the National Institute of Aging (NIA) and the Institute for Social Research at the University of Michigan.

The Social Security Administration (SSA) plays a key supporting role in this study by providing income information of HRS participants. The mortality information of

the participants can be obtained from the National Death Index (NDI) and via the spouse's report or an HRS exit interview.

This study contains a rich, comprehensive database that facilitates research to explore both cross-sectional and longitudinal aspects about America's older adults' social, economic, health, psychological, and other critical age-related changes. As a clear consequence of the importance of this study and HRS's open access policy to the data, investigators and research scholars in the United States and the rest of the world have shown a growing awareness and interest about HRS data. This had led to numerous studies using the HRS data appearing in all forms of publication media.

As shown in the Figure 2.2.1, since the inception of the HRS data, about 159 books and book chapters, 338 dissertations, 624 reports and $1,510$ journals have incorporated the use of the HRS data. These publications have ranged in various disciplines such as health economics, medical statistics and even insurance, and it is not surprising to see that some of these results have leading decision making decision policy implications.

We see the value and importance of this rich and comprehensive HRS data, and therefore we find its usefulness in this thesis. The underlying purpose of this chapter is to explore the importance of this data and how it has been utilized in several disciplines.

2.2 Contributions in different disciplines

Various research published in $1,510$ journal articles, 159 books and book chapters, and 338 dissertations, according to *Scientific Productivity of HRS* (2013), have been

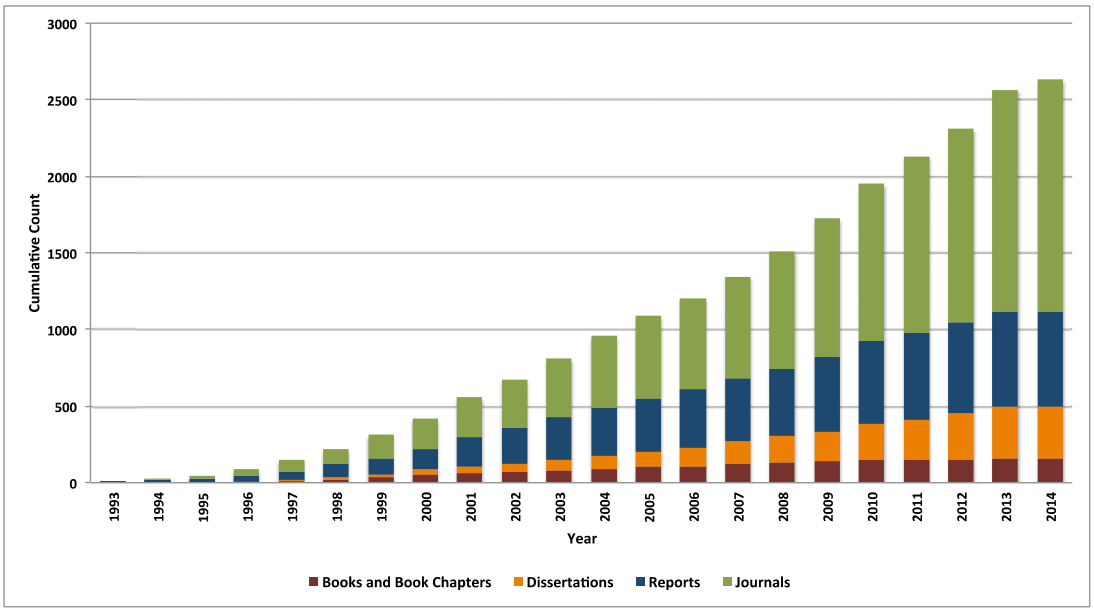


FIGURE 2.2.1: Cumulative Count of HRS Publications by Type
Source: University of Michigan

using the HRS data since it became available and accessible in 1992. Economics, finance, health economics, medical statistics, and actuarial science are some of the major research areas. The following is a sampling of some of these works.

Numerous economists have relied on HRS data since it was launched. For a start, the work done by Barsky et al. (1997) estimates the individual preference parameters linked to risk tolerance, time preference, and intertemporal substitution by analyzing the responses to some questions, originating from economic experts, of HRS participants. They discovered a positive relationship between risk tolerances and risky behavior while finding heterogeneity in the preference parameters. In yet another work, using the HRS data, McGarry and Schoeni (1995) assesses the family support network and further analyzes the redistribution aspects of family resources from better-off parents to children as money or wealth and to elderly parents as money or time. Using different statistical modeling techniques such as ordinary regression, logit, and fixed effects models, this paper examined the relationship between household transfers and income of children/elderly parents after controlling for several characteristics. The key outcome suggests negative correlations between transfers from prosperous parents and income of children/elderly parents. Another interesting work on retirement decisions is made by Rogowski and Karoly (2000). Such retirement decisions are always a fascinating concern for most of the economists as the baby boomers arrive close to their retirement phase. These authors reviewed the context of health insurance and retirement choices using the HRS data from 1992 to 1996. According to their findings, the financial factors that influence the speedy retirement and the access to have retiree health insurance boost early retirement decisions.

Medical scientists also have extended their research on elderly population as this group rapidly grows. Life depression in later ages is found to be a major concern

for senior citizens because it is believed to be linked to major health issues including disability. By examining 1996 HRS participants, Dunlop et al. (2004) discover that functional limitation is a major cause associated with depression and it leads to chronic illnesses such as arthritis and heart disease. This study also promotes screening elderly for functional limitations. A large body of existing literature indicates that involuntary job loss burden the life events of elderly by causing undesirable consequences such as financial difficulties, loss of pension benefits, weak mental health, and increased boost of tobacco use. After investigating 10 years of HRS data, Gallo et al. (2006) added heart attack and stroke to this list. Banks et al. (2006) have done a cross-country morbidity comparison between the United States and the U.K., using the HRS for the United States and its U.K. counterpart called the English Longitudinal Study of Aging. The authors examined self-reported and biological health data of non-Hispanic whites elderly population in the United States and its counterpart in the U.K. population. Results from ordinary least squares regression methods provided evidence that after adjusting for major socio-economic factors like income and education, late middle-aged U.S. population has been exposed to more chronic diseases like diabetes, heart disease, hypertension, and stroke than its British counterpart across the wide range of socio-economic distributions.

Concerns about health economics have motivated a number of research as aging population increases in recent decades. The following articles have used HRS data to investigate any burden concerns this group in the population might contribute. Due to numerous reasons, there has an expansion of US grandparent-headed households in recent years. By analyzing 8 years of the Center for Epidemiologic Studies-Depression (CES-D) scale data within HRS, Blustein, Chan, and Guanais (2004) have revealed that among members of this group who act as caregiving grandparents, women are

more sensitive to symptoms of increasing signs of depression than men. In the work of Byrne et al. (2009), elder care arrangements in the form of either formal and informal care have become a major concern for the community. To get a statistical understanding of caregiving decisions of families, they used HRS data to fit a game-theoretic model and discovered that both children and children-in-law find informal care and quality caregiving results from own children to the parents difficult tasks to do. Within these findings, even educated elders prefer formal care. Cigolle et al. (2007) examined the prevalence of geriatric conditions such as cognitive impairment, low body mass index, dizziness, and vision impairment. These conditions are increasingly becoming important issues for older adults. These authors studied the HRS cross-sectional data for the year 2000 and find evidence of a strong association between geriatric conditions and disability among older adults, even after controlling for demographic characteristics such as age, sex, race, education, to name a few, and for chronic diseases such as heart disease, lung disease, and diabetes.

Mitchell et al. (2006) investigated the global financial market's reactions to universal aging trends and the importance of insurance products innovation to manage old-age risks. Here in this work, the HRS data was used to get an overview about the existing retirement resources for pensioners. Their findings indicate that Social Security benefits, pension benefits, and homes fully paid are the major assets of the retireds, whereas financial assets share only a tiny portion of their retirement wealth. This analysis emphasizes the potential importance of financial products and product innovations including life annuities, long-term care benefits, reverse mortgages, and other products providing guarantees, to provide financial security to retirees against longevity risk. Defined Benefit (DB) pension plan can help retirees against longevity risk, and using the HRS data, Friedberg and Webb (2005) investigated how

the pension structure changes the retirement framework as the market moves towards a preference to Defined Contribution (DC) plans. Due to various reasons such as low interest rates, increasing longevity risk, and high volatility in the market, the number of DB pension plans are decreasing while the number of DC pension plans are increasing. The finding indicates that persons holding DC plans tend to retire about two years later than persons with DB plans. This result validates slightly raising retirement age as a result of the shift in pension structure from DB to DC. A debating concern about retirement decisions and health insurance is investigated by Blau and Gilleskie (2006). In their study, an attempt is made to find a quantitative association between retiree health insurance and employment behavior of older married couples using the $1992 - 1998$ HRS data. A dynamic model is applied for the information about employment history, medical expenditures, health insurance, and income of married couples. The outcome of the study implies a moderate influence between health insurance and the retirement decisions of older married couples, and at the same time, health insurance does not appear to be a key, influential factor for retirement decisions of older married couples.

2.3 The HRS data for our purpose

Longitudinal data contains information about many subjects, with a series of observations over a period time. Such data allows us to examine both cross-sectional characteristics and time-varying relationships. Many attribute the use of longitudinal data to have several advantages that include more flexibility in research design, the ability to identify inter-temporal patterns, and the incorporation of time-varying

predictors. For additional advantages of the use of longitudinal data, please see Frees (2004) and Singer and Willett (2003).

For purposes of this thesis, we drew a sample of observations from the HRS longitudinal survey data that finally consisted of $7,607$ adult participants who are non-institutionalized, living in the contiguous United States, with age 50 and above, and are financially responsible. These participants joined the HRS study between 1992 and 2010, but with a final follow up extending to 2011. This HRS sample is selected using the Four-stage Area Probability Sample Design technique as employed within the HRS database. As mentioned in Heeringa and Connor (1995), to get closer eyesight of senior population at the State of Florida, Black and Hispanics, an oversample of these three cohorts are included in the HRS data. As a result, we also considered the use of sampling weights mainly to offset this geographic and race differences.

As stated in Juster and Suzman (1995), two of the main HRS design decisions were the accuracy of data and the rate of response. However, as a consequence of advance planning done by HRS, the overall response rates have been impressive. At each follow-up wave till 2008, these response rates have range from 85% to 93%. These impressive response rates are acceptable for our advanced studies. Moreover, as demonstrated below, we are able to track the observations in our sample that have missed responses during certain survey periods.

Our observations are best illustrated by Figure 2.3.1. Here we observe five different types of HRS members, labeled A, B, C, D, and E. The (blue) dotted line indicates the censored date referring to year 2011 when the latest follow up of mortality is made. We now describe each of these members. Member A participated in the surveys contacted between 1992 to 2000 and died before the 2002 survey can be conducted. Member B participated in the surveys all the way to the end of 2010 but no records of

mortality in 2011 and therefore considered censored. Member C participated surveys but missed some periods of survey and died prior to 2011. Member D participated survey only in 1992, missed the rest of the survey and presumed alive by 2011. Finally, Member E participated survey only in 1992 and died before 2011. For our purposes,

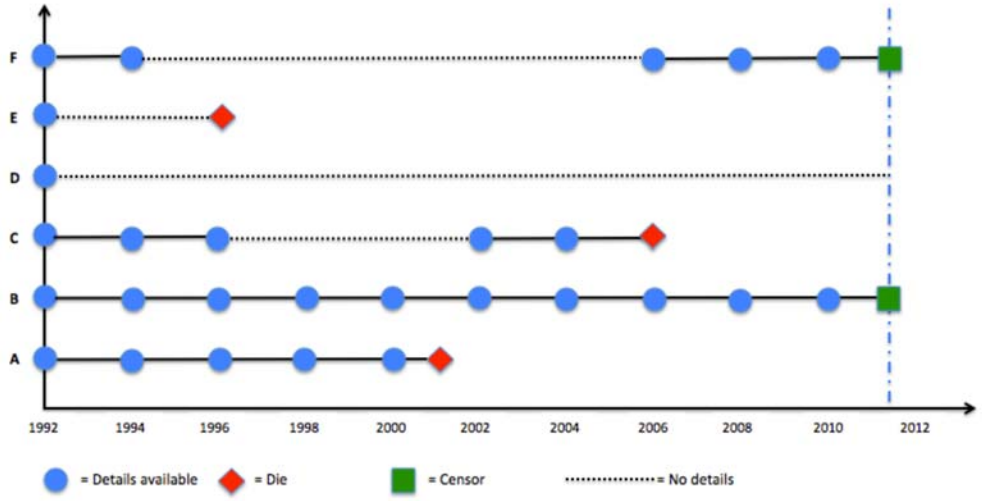


FIGURE 2.3.1: Flow of participants in the survey

these HRS data were used to analyze survival models applicable to pension liabilities as explained Chapter 3 and again used to study multiple state models for long term care (LTC) as explained in Chapter 4.

Chapter 3

Survival Models for Pension Liabilities

3.1 Introduction

Within the last century, significant medical progress, socio-demographic changes, improvements in lifestyles, the absence of major pandemic crises, and probably a combination of these factors, contributed to a significant reduction in mortality across several countries in the world. As a result, several developed, and even developing, countries have recorded a marked decline in mortality during this period. Some are still indeed continually experiencing improvements in life expectancy. Figure 3.1.1 shows some evidence of this reduction in mortality around the world.

As we go through these phases, understanding trends in mortality rates are ever more relevant in the insurance industry. As a result, despite the investment component being important in many insurance products nowadays, exploring and modeling

longevity exposure have become a key component in the growing pension market that has witnessed steadily decreasing patterns of mortality rates over time. To advance global thinking of future trends of mortality rates, the International Actuarial Association (IAA) has set up a Mortality Working Group. The main focus of the IAA Mortality Working Group is to investigate the impact of mortality on insurance products such as life insurance and pensions, by studying trends and patterns of longevity globally and within a nation. Some findings of the IAA Mortality working group show a greater impact caused by social changes rather than mere medical improvements towards improved longevity. See, for example, Ridsdale and Gallop (2010).

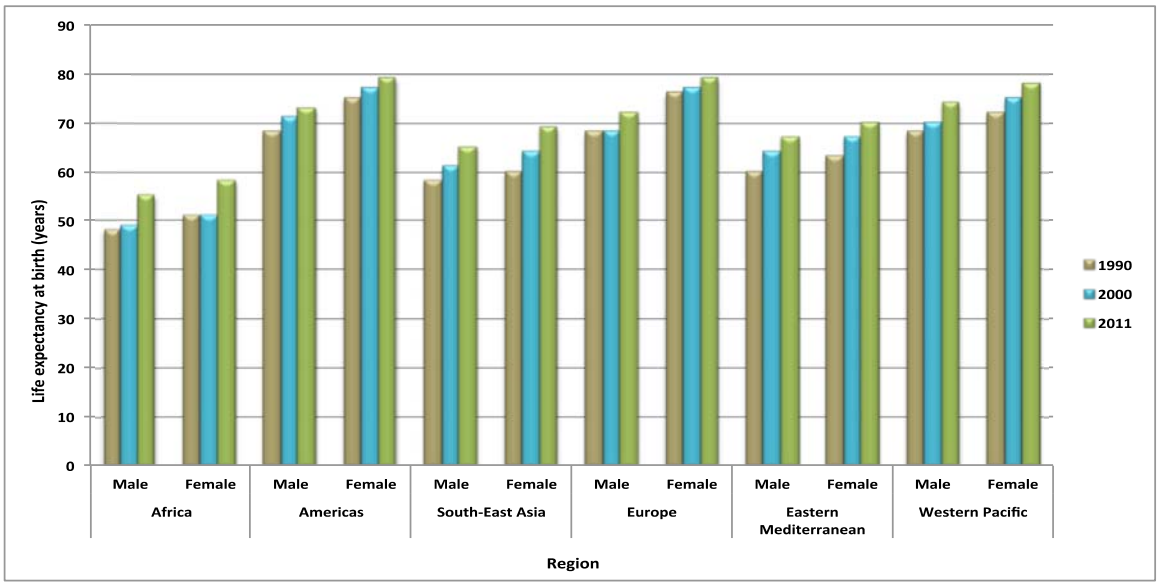


FIGURE 3.1.1: Life Expectancy at Birth by Regions

Source: World Health Organization

Several studies are now saying that the United States is expected to experience a rapid growth in the retired population from 2010 to 2050. First, the average lifespan of Americans will continue to increase and this increase extends well beyond normal

therefore sponsoring a number of research projects on mortality and longevity studies, and at the same time, encouraging its actuarial memberships to search for innovative ways to minimize the possible catastrophic effects of surviving longer. Rosner et al. (2013) and *Mortality Improvement Scale BB* (2012) are some of most recent major experience studies sponsored by the SOA. As Americans are expected to spend more time in retirement years, the allocation of economic resources associated with aging becomes even more important. This is not only applicable to the individuals but also to the society as a whole. With lack of strategic planning, these unforeseen costs may impose a serious strain on personal savings and wealth of the individuals, and possibly even drain government resources. The Social Security will continue to be the primary source of retirement income, but private pension is becoming a second essential source. See Moore (2011) and Quadagno and Pederson (2012).

There are generally two primary types of pension plans: the defined contribution (DC) and the defined benefit (DB) pension plans. A DC plan helps build a retirement account for the participants by making regular contributions, during years employed, coming from both the employee and the sponsor, which is usually the employer. A DB plan, on the other hand, will provide retirement benefits according to a predetermined formula which depends on earnings made during employment, age and years of service. The primary difference between these two plans has to do with who bears the investment risk during the accumulation phase and who bears the longevity risk during the retirement phase. It is clear that in a DB plan, the plan will continue to bear the longevity risk so that retirement income will continue as promised so long as the retired is alive.

Clearly then, outliving resources is one of the primary risks for defined benefits plans. The American Academy of Actuaries (AAA) Pension Committee recently pub-

lished a public policy practice note intended to spur the pension actuaries professional judgments about the impact of potential future mortality trends, especially for DB pension plans. As per the Pension Committee Practice, see Note (2011), collar type of employment, income, gender, occupation, and geographic location of residence are just some of the demographic factors which may affect mortality improvements.

In this chapter, we intend to examine and investigate key socio-demographic factors that affecting the health of individuals and thereby presumably their survival pattern. Using the source of information obtained from the HRS data, we calibrated a survival model, based on Cox regression, using various categories that may affect mortality: health, demographic, lifestyle, and financial factors. We find that each of these categories contribute to mortality at varying degres; we try to express their relative importance. Finally, we used this calibrated model, which in some sense extends the usual age-gender factor model, to project future mortality and thereby project pension liabilities.

3.2 Survival modeling

The history of mortality modeling extends long beyond the efforts of de Moivre (1725) and Gompertz (1825). The history of mortality projections can be traced back to 1875 when a Swedish Astronomer Gyldèn fitted a straight line to the Swedish mortality data as noted by Cramér and Wold (1935). Additional details about the seminal contributions of mortality modeling in actuarial science can be found in Haberman (1996), Pitacco (2004), and Booth and Tickle (2008).

Survival analysis is the most generally used statistical method for analyzing the

timing of events. Applications of survival analysis are widely utilized in several disciplines such as sociology, engineering, economics, and not surprisingly, actuarial science. The terminology varies from discipline to discipline. It can be time-to-event analysis in the social sciences, reliability analysis in the engineering sciences, and duration analysis in economics. For more details, please refer to Allison (2012). Survival analysis can be flexibly performed with either retrospective or prospective data. Censoring, where some observations may be incomplete due to random causes, and time-dependent covariates, that may change in value over the observation period, are two regular important features of data that survival modeling can handle with ease.

Survival models also come in various forms: parametric, nonparametric, semi parametric, and discrete. For example, Exponential and Weibull are some parametric survival models where some functional form is assumed for the distribution of survival time. Life table is an example of a nonparametric model which directly uses the observed data to define survival and hazard functions. Cox model is an example of semi-parametric survival models. Logit and Probit models are some discrete survival models.

Driven by the nature of our data, we intend to use the semi-parametric approach to calibrate the survival model, although the parametric approach, which has been used in actuarial models, may be used for comparison purposes. Here the data consists of the information about the attributes of multiple persons which are taken with respect to time. This aspect gives a longitudinal flavor for this survival analysis.

3.3 Cox proportional hazard models

Sir David Cox, a British statistician, proposed the extremely popular, semi-parametric Cox regression model as it appeared in Cox (1972). According to this model, the hazard function contains both fully parametric components and a nonparametric baseline component. By generalizing the ideas of conditional and marginal likelihood, Cox (1975), presented an innovative approach of partial likelihood method to estimate the covariate parameters associated with the survival model. The partial likelihood method enables one to handle censoring and time-varying explanatory variables. Because of these two main attractive features, Cox model is exceedingly popular in the literature on statistics and even in other disciplines.

3.3.1 The theory of Cox regression models

Let T be a nonnegative random variable representing failure time and its probability distribution can be defined by the survivor function:

$$S(t) = \Pr(T \geq t). \tag{3.3.1}$$

The probability density function is

$$f(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t)}{\Delta t}.$$

The hazard function is

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{\Pr(t \leq T < t + \Delta t \mid t \leq T)}{\Delta t} = \frac{f(t)}{S(t)} \tag{3.3.2}$$

For discrete T,

$$\lambda(t) = \sum \lambda_j \delta(t - j)$$

where $\delta(t)$ is the Dirac delta function and $\lambda_j = Pr(T = j \mid T \geq j)$. Using the product law of probability and product integral, we have

$$S(t) = \lim_{(\tau_{k+1} - \tau_k) \to 0} \prod_{k=0}^{r-1} [1 - \lambda(\tau_k)(\tau_{k+1} - \tau_k)] \tag{3.3.3}$$

with $0 = \tau_0 < \tau_1 < \tau_2 .... < \tau_r = t$. If $\lambda(t)$ is integrable, then the survivor function becomes

$$S(t) = e^{-\int_0^t \lambda(u) du} \tag{3.3.4}$$

and in the case where T is discrete, then

$$S(t) = \prod_{j<t} (1 - \lambda_j) \tag{3.3.5}$$

Following Cox (1972), Kaplan and Meier (1958) has extended the work by incorporating regression-like concepts into life table analysis. In addition, Cox further developed a method to assess the relationship between the distribution of failure time and covariates.

Let $\mathbf{z} = (z_1, z_2, z_3, ...z_r)$ be time independent or dependent covariates with $r > 1$ so that the hazard function is

$$\lambda(t; \mathbf{z}) = \lambda_0(t) e^{\mathbf{z}\beta} \tag{3.3.6}$$

where β is a $r \times 1$ vector of unknown regression coefficients and $\lambda_0(t)$ is an unknown

arbitrary nonnegative function. By developing Cox's ideas where covariate processes have a proportional effect on the intensity process of a multivariate counting process, Fleming and Harrington (2005) proposed the multiplicative hazards models for failure times using Martingale theory. This method permits to analysis of the intensity of a recurrent event with complicated censoring patterns and time dependent covariates. The multiplicative hazards models can be considered as a superset of the Cox model.

As demonstrated by Andersen and Gill (1982) and Fleming and Harrington (2005), consider the right-censored failure time data for independent observations on $(X, \delta, \mathbf{Z})$ where $X = \min(T, U)$, T is failure and U is censoring times, $\delta = I_{[T \leq U]}$ are failure indicators, $\mathbf{Z}$ is a p-dimensional column vector of covariates. The stochastic basis with the right-continuous filtration $\{F_t : t \geqslant 0\}$ is defined as

$$F_t = \sigma \left\{ \boldsymbol{Z}, N(u), Y(u+) : 0 \leq u \leq t \right\}.$$

According to the Doob-Meyer Decomposition, for the increasing process N, there is a unique predictable process A with respect to F_t such that $N - A$ is a Martingale. When A' exists, it is called the intensity process for N. Aalen (1978) has shown that

$$\lim_{h \to 0} \frac{1}{h} \wp \left[N(t+h) - N(t) = 1 | F_t \right] = \lambda(t+)$$

where

$$\lambda_i(t) = Y_i(t) \lambda_0(t) \exp^{[\beta_0 \mathbf{Z_i(t)}]}$$

with $Y_i(t)$ defined to be a predictable process taking values $\{0, 1\}$, λ_0 is a fixed underlying hazard function, β_0 is a fixed column vector of p coefficients, and $\mathbf{Z_i}$ is a column vector of p covariates.

To estimate β_0, partial (Cox's) likelihood technique is used and this results in the partial likelihood for n independent triplets $(N_i, Y_i, \mathbf{Z_i})$ with possible ties in observed failure times so that for $i = 1, 2, \ldots, n$,

$$L(\beta, t) = \prod_{i=1}^{n} \prod_{s \geq 0} \left\{ \frac{Y_i(s)\exp[\beta' \mathbf{Z_i}(\mathbf{s})]}{\sum_{j=1}^{n} Y_i(s)\exp[\beta' \mathbf{Z_i}(\mathbf{s})]} \right\}^{\Delta N_i(s)} \tag{3.3.7}$$

where $\Delta N_i(s) = 1$ if $N_i(s) - N_i(s-) = 1$, and otherwise $\Delta N_i(s) = 0$.

3.4 Model estimation

It is well known that age and gender affects mortality. However, various studies in several disciplines such as medical sciences, gerontology, sociology, and economics have discovered significant effects of additional relevant factors on adult mortality. See, for example, Crimmins (2011) and Brown and McDaid (2003).

To calibrate the Cox proportional hazard model as described in the previous section, we used the HRS data that consisted of $7,067$ survey participants with follow up studies that was fully described in Chapter 2. In our model estimation, we considered broad categories of factors possibly affecting mortality and its improvement. The following describes such four broad categories:

- Demographic risk factors such as age, gender, education level, current partnership status, and number of kids.

- Health indicators such as current and past high blood pressure, and the presence or absence of diseases like diabetes, cancer, and cardio-related problems.

- Lifestyle factors such as smoking habits, alcohol consumption, body mass index, and the level of cardio activities.

- Financial factors such as the level of income and wealth, to name a few.

In the subsequent sections, we describe the data characteristics and other observable information of each categories from our data set.

3.4.1 Variables

The information obtained from the HRS data has been summarized in four tables with each table represent one of the four broad categories described previously. While these tables are self-explanatory, we briefly described some of the key characteristics we can draw from each table.

- In Table 3.4.1, we summarize demographic variables. Here, for example, we find that we have a slightly larger proportion of male respondents (55%). About 75% individuals who have at least graduated from high schools. Only 36.14% of persons who are living alone. About 33% deaths recorded during the period from 1992-2011.

- In Table 3.4.2, we summarize health variables. Here, for example, we used 3 different categorical variables describing health. CONDE refers to the total number of conditions that one ever had. About a fifth of the respondents never experienced any of these conditions, 29% experienced exactly one of these conditions, 24% experience 2 of these conditions, and the rest have at least 3 of these conditions. SHLT is the self-reported health, whether the respondent

feels excellent, very good, good or fair/poor health. Finally CESD is a mental health index.

- In Table 3.4.3, we summarize the lifestyle variables. We chose the variables describing smoking habits, drinking habits, level of physical activity and BMI to fall in this category. The summary statistics in the table are self-explanatory.

- In Table 3.4.4, we summarize the financial variables. Here the primary variables of interest fall in either a categorical variable (e.g. collar type of employment, household income) or a continuous variable (level of wealth, total assets and capital income, to name some).

TABLE 3.4.1: Data Characteristics and other Observable Information for the Demographic Variables

Categorical Variables	Description		Proportions	
RAGENDER	Gender of the respondent:	Male=0	54.57%	
		Female=1	45.43%	
RAEDUC	Education:	College and above= 0	39.14%	
		High-school graduate = 1	36.05%	
		Lt High-school = 2	24.81%	
RMARRY	Current Partnership Status:	Single=0	36.14%	
		Married/Partnered=1	63.86%	
CENDIV	Census Division:	New England = 1	3.47%	
		Mid Atlantic = 2	13.32%	
		EN Central = 3	15.56%	
		WN Central = 4	8.32%	
		S Atlantic = 5	25.82%	
		ES Central = 6	6.78%	
		WS Central = 7	9.84%	
		Mountain = 8	4.77%	
		Pacific = 9	11.99%	
		Not US = 11	0.11%	
CENSOR	Censoring indicator for death:	Alive = 0	67.29%	
		Died= 1	32.71%	
Continuous Variables		Minimum	Mean	Maximum
HKIDS	Number of living children of household	0	3	20
AGE	Age of the respondent	50	63	107
PAVAGE	Parents Average Age	24	74	100
Date		Minimum	Mean	Maximum
DEATHY	Death Year	1992	2002	2011
RABYEAR	Birth year of the respondent	1900	1935	1942

TABLE 3.4.2: Data Characteristics and other Observable Information for Health Variables

Categorical Variables	Description		Proportions
CONDE	Sum of conditions (high blood pressure, diabetes, cancer, lung disease, heart attack, stroke, psychiatric problems, arthritis) ever had:	None = 0	21.00%
		One = 1	28.96%
		Two = 2	24.04%
		More than Three = 3	26.00%
SHLT	Self-reported health:	Excellent = 1	13.66%
		Very good = 2	28.22%
		Good = 3	30.12%
		Fair/Poor = 4	27.99%
CESD	Mental Health Index:	None = 0	53.52%
	sum of depression, effort, restless sleep, alone, sad, (1- happy), (1- enjoy life):	One or more = 1	46.48%

TABLE 3.4.3: Data Characteristics and other Observable Information for Lifestyle Variables

Categorical Variables	Description		Proportions	
SMOKEV	Smoking Status:	Non-smoker = 0	35.50%	
		Former smoker = 1	45.04%	
		Current smoker = 2	19.46%	
DRINKR	Alcohol Drinking Status:	$<$ 1 drink per day = 0	61.00%	
		1-2 drinks per day = 1	31.49%	
		$\geq$ 3 drinks per day = 2	7.51%	
VIGACT	Physical activity or Exercise 3+ times a week:	No = 0	67.09%	
		Yes = 1	32.91%	
Continuous Variable		Minimum	Mean	Maximum
BMI	Body Mass Index (kg/m^2)	10.80	27.84	102.70

TABLE 3.4.4: Data Characteristics and other Observable Information for Financial Variables

Categorical Variable	Description		Proportions	
RCOLLAR	Collar:	White = 1	26.75%	
		Mixed = 2	27.35%	
		Blue = 3	45.91%	
HTINC	Total Household Income:	Low = 0	25.11%	
		Average = 1	49.95%	
		High = 2	24.94%	
Continuous Variables		Minimum	Mean	Maximum
HTOTW	Total Wealth(excluding IRAs and less all debt)	-4,733,000	282,791	77,165,000
HTOTA	Total Assets	-4,733,000	330,380	77,225,000
HTOTN	Total Non-housing Wealth	-770,000,000	230,475	76,625,000
HICAP	Household Capital Income	0	11,911	7,331,325

3.4.2 Variable selection

Variable selection is an important procedure and plays a pivotal role in statistical modeling. Many statistical discoveries have been done with model selection techniques to help trim down spurious effects to a manageable level. For our purpose, we examine various statistical methods for detecting the best subsets for our model. Many of these methods are classified as standard automated methods. We can classify standard variable selection methods as forward selection, backward selection, and stepwise selection. All three selection procedures were comparatively used. For different combinations of risk factors, Tables 3.4.5 and 3.4.6 indicate the results from these automated methods. A check means the variable is statistically significant at a 5% level of significance; otherwise, it is listed with a cross mark.

As shown by Tables 3.4.5 and 3.4.6, only controlling for demographic variables such as age, gender, marital status, region and parents' average age, the education variable is significant predictor for mortality. However after adjusting for both demographic and financial variables including income and wealth, the significance of education level seems to decline; these are similar to findings with Kwon and Jones (2006) and Bassuk, Berkman, and Amick (2002). This result is presumably because education induces higher income and the variable selection has chosen the strongest predictor for the model. According to the U.S. Census Bureau as in Day and Newburger (2002), for full time employees aged 25 to 64, the average earnings for the period $1997 - 1999$ were about $\$18,900$ for high school dropouts, $\$25,900$ for high school graduates and $\$45,400$ for college graduates.

SAS is a predominant statistical language with some extremely powerful survival analysis tools. For this purpose, the analysis has been conducted using SAS; please

refer to Allison (2012).

TABLE 3.4.5: Variable Selection Part I

Variables	Forward	Backward	Stepwise
Demographic			
RAGENDER	√	√	√
AGE	√	√	√
RAEDUC	√	√	√
RMARRY	√	√	√
CENDIV	√	√	√
HKIDS	×	×	×
PAVAGE	√	√	√
Financial			
HTOTA	×	√	×
HTOTW	×	×	×
HTOTN	×	√	×
HICAP	×	×	×
RCOLLAR	√	√	√
HTINC	√	√	√
Health			
SHLT	√	√	√
CONDE	√	√	√
CESD	√	√	√
Lifestyle			
VIGACT	√	√	√
DRINKR	√	√	√
SMOKEV	√	√	√
BMI	√	√	√
Demographic and Financial			
RAGENDER	√	√	√
AGE	√	√	√
RAEDUC	√	√	√
RMARRY	√	√	√
CENDIV	√	√	√
HKIDS	×	×	×
PAVAGE	√	√	√
HTOTA	×	√	×
HTOTW	×	×	×
HTOTN	×	√	×
HICAP	×	×	×
HICAP	×	×	×
RCOLLAR	×	×	×
HTINC	√	√	√
Demographic and Health			
RAGENDER	√	√	√
AGE	√	√	√
RAEDUC	×	×	×
RMARRY	√	√	√
CENDIV	√	√	√
HKIDS	×	×	×
PAVAGE	√	√	√
SHLT	√	√	√
CONDE	√	√	√
CESD	√	√	√

TABLE 3.4.6: Variable Selection Part II

Variables	Forward	Backward	Stepwise
Demographic and Lifestyle			
RAGENDER	√	√	√
AGE	√	√	√
RAEDUC	×	×	×
RMARRY	√	√	√
CENDIV	√	√	√
HKIDS	×	×	×
PAVAGE	√	√	√
VIGACT	√	√	√
DRINKR	√	√	√
SMOKEV	√	√	√
BMI	√	√	√
Demographic, Financial and Health			
RAGENDER	√	√	√
AGE	√	√	√
RAEDUC	√	√	√
RMARRY	√	√	√
CENDIV	√	√	√
HKIDS	×	×	×
PAVAGE	√	√	√
HTOTA	×	√	×
HTOTW	×	×	×
HTOTN	×	√	×
HICAP	×	×	×
HICAP	×	×	×
RCOLLAR	×	×	×
HTINC	√	√	√
SHLT	√	√	√
CONDE	√	√	√
CESD	√	√	√
Demographic, Financial, Health and Lifestyle			
RAGENDER	√	√	√
AGE	√	√	√
RAEDUC	√	√	√
RMARRY	√	√	√
CENDIV	√	√	√
HKIDS	×	×	×
PAVAGE	√	√	√
HTOTA	×	√	×
HTOTW	×	×	×
HTOTN	×	√	×
HICAP	×	×	×
RCOLLAR	×	×	×
HTINC	√	√	√
SHLT	√	√	√
CONDE	√	√	√
CESD	√	√	√
VIGACT	√	√	√
DRINKR	√	√	√
SMOKEV	√	√	√
BMI	√	√	√

3.4.3 Final model

After the procedure for variable selections, Table 3.4.7 provides the maximum likelihood estimates of our final model selected. In this table, we present the variables selected together with their estimated coefficients, in parenthesis the standard errors, and their corresponding hazard ratios. In order not to overwhelm the reader, we give some interpretations to these results:

- AGE is obviously a significant factor and a hazard ratio of about 1.058 indicates that as AGE increases by one year, the hazard rate will increase by approximately 5.8%.

- BMI gives an otherwise intuitive result. We find that BMI is also a significant factor in predicting mortality, but then a higher BMI, according to our model, will slightly decrease the hazard rate.

- GENDER is a categorical variable and our table gives the coefficient for male (=0). This table gives a positive estimated coefficient which means that on the average, males tend to have higher hazard rates than females.

- RMARRY refers to whether the respondent is single (=0) or married (=1). This table gives a hazard ratio of 1.183 for single respondents, indicating that hazard rates for single is worse by about 18% than married respondents.

TABLE 3.4.7: Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Hazard Ratio
AGE		1	0.056 (0.006)	1.058
RAGENDER	0	1	0.510 (0.067)	1.666
PAVAGE		1	-0.007 (0.002)	0.993
CENDIV	1	1	-0.243 (0.179) *	0.784
CENDIV	2	1	-0.241 (0.101)	0.786
CENDIV	4	1	-0.172 (0.117) *	0.842
CENDIV	5	1	-0.115 (0.088) *	0.892
CENDIV	6	1	-0.354 (0.124)	0.702
CENDIV	7	1	-0.214 (0.115)	0.807
CENDIV	8	1	0.007 (0.133) *	1.007
CENDIV	9	1	-0.281 (0.108)	0.755
CENDIV	11	1	0.138 (0.712) *	1.148
RMARRY	0	1	0.168 (0.071)	1.183
CONDE	0	1	-1.578 (0.145)	0.207
CONDE	1	1	-0.988 (0.089)	0.372
CONDE	2	1	-0.560 (0.072)	0.571
SHLT	1	1	-1.112 (0.158)	0.329
SHLT	2	1	-1.211 (0.101)	0.298
SHLT	3	1	-0.759 (0.072)	0.468
CESD	0	1	0.307 (0.061)	1.360
HTINC	0	1	0.340 (0.104)	1.406
HTINC	1	1	0.197 (0.087)	1.218
VIGACT	0	1	0.637 (0.081)	1.890
DRINKR	0	1	0.212 (0.112)	1.236
DRINKR	1	1	-0.043 (0.123) *	0.958
SMOKEV	0	1	-0.691 (0.085)	0.501
SMOKEV	1	1	-0.345 (0.071)	0.708
BMI		1	-0.041 (0.006)	0.960

Notes:

a. Standard errors are in parenthesis.

b. An asterisk * identifies 'not significant' at the 10% level.

3.5 Model implications

To understand the implications of the final model in Table 3.4.7, we isolated the effects of two important variables: the self-reported health and the average age of the parents. For pension calculation purposes, we relied on calculating annuity immediate values assuming annual payments of $52,000 for a total of 18 years at a 5% compounded interest rate. For comparative purposes, we compared the survival patterns of selected characteristics by varying each isolated effect. In order to understand how these affect mortality and pension values, we also compared the respective values using the RP-2000 (combined healthy; base) pension mortality table with a Scale BB improvement rates (herewith we called RP-2000).

3.5.1 Self-reported health

Self-reported health status rapidly appears as a significant predictor for most of the mortality studies in global literature even controlling for key covariates; see Idler and Benyamini (1997), DeSalvo et al. (2006) and Jylhä (2009). For different self-reported health status, this section investigates the survival rates for the following characteristics in Table (3.5.1). SHLT is the self-reported health index (1=Excellent, 2=Very good, 3=Good, 4=Fair/Poor).

TABLE 3.5.1: Selected characteristics

ID	RAGE	RAGENDER	PAVAGE	CENDIV	RMARRY	CONDE	SHLT	CESD	HTINC	VIGACT	DRINKR	SMOKEV	BMI
M1	65	0	65	2	1	0	1	0	1	0	0	0	25
F1	65	1	65	2	1	0	1	0	1	0	0	0	25
M2	65	0	65	2	1	0	2	0	1	0	0	0	25
F2	65	1	65	2	1	0	2	0	1	0	0	0	25
M3	65	0	65	2	1	0	3	0	1	0	0	0	25
F3	65	1	65	2	1	0	3	0	1	0	0	0	25
M4	65	0	65	2	1	0	4	0	1	0	0	0	25
F4	65	1	65	2	1	0	4	0	1	0	0	0	25

Figure 3.5.1 shows the model predicted survival rates for the above selected characteristics and survival rates based on the RP-2000 pension mortality table with Scale BB improvement rates. First, because we know that females tend to live longer, their survival rates are worse than males. This is true for the selected characteristics as well as the rates produced by the RP-2000 table. Second, the RP-2000 Table tends to underestimate the survival rates of individuals when they know how feel. Individuals with excellent self-reported health have a much higher survival rate than everyone else.

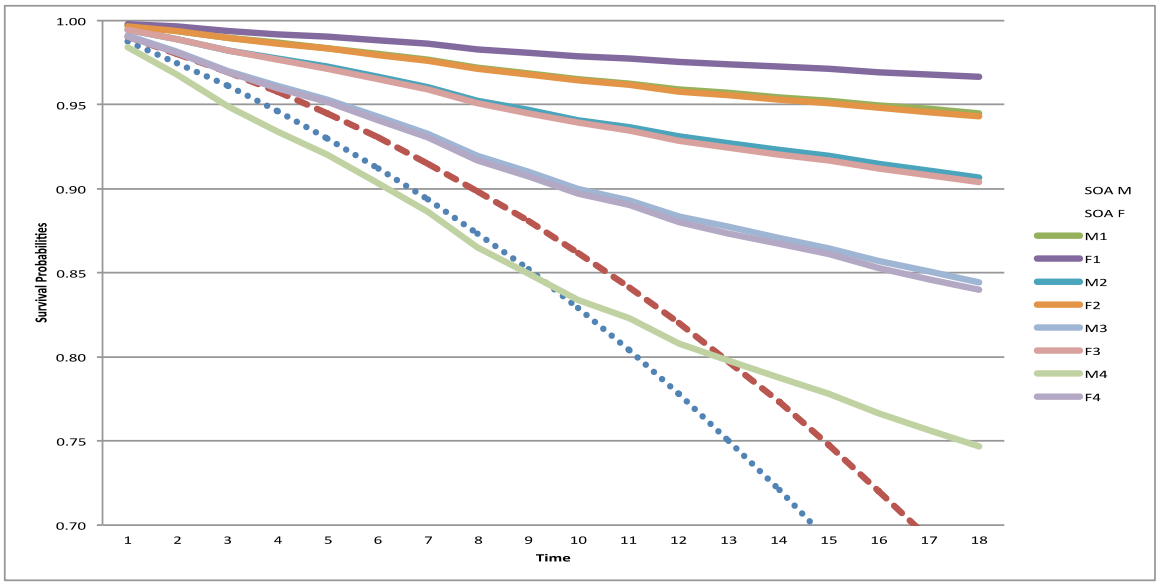


FIGURE 3.5.1: Effects of SHLT on Survival Probabilities

Figure 3.5.2 provides a comparison of the annuity immediate values, again using the RP-2000 Table in comparison to the final model as earlier described, for the selected characteristics. This same story can be drawn from the survival rates earlier discussed. The RP-2000 Table, which produces the worse survival rates, will give much lower annuity values. The degree of self-reported health status also have an effect on the annuity values. Those that give an Excellent self-reported health status are those that tend to have better survival rates will then have higher annuity values.

FIGURE 3.5.2: Annuity immediate values ($)

3.5.2 Parent average age

Some studies have identified a strong influence between increases in lifespan and possible genetic impacts; see Hjelmborg et al. (2006) and Yashin et al. (2000). In this section, by varying parent average age for the following characteristics in Table 3.5.2, we show possible genetic impacts on survival rates and pension liabilities. Here, we vary PAVAGE (parent's average age) from 50 to 75, with equal increments of 5 years.

TABLE 3.5.2: Characteristics

ID	RAGE	RAGENDER	PAVAGE	CENDIV	RMARRY	CONDE	SHLT	CESD	HTINC	VIGACT	DRINKR	SMOKEV	BMI
M1	65	0	50	2	1	0	1	0	1	0	0	0	25
F1	65	1	50	2	1	0	1	0	1	0	0	0	25
M2	65	0	55	2	1	0	1	0	1	0	0	0	25
F2	65	1	55	2	1	0	1	0	1	0	0	0	25
M3	65	0	60	2	1	0	1	0	1	0	0	0	25
F3	65	1	60	2	1	0	1	0	1	0	0	0	25
M4	65	0	65	2	1	0	1	0	1	0	0	0	25
F4	65	1	65	2	1	0	1	0	1	0	0	0	25
M5	65	0	70	2	1	0	1	0	1	0	0	0	25
F5	65	1	70	2	1	0	1	0	1	0	0	0	25
M6	65	0	75	2	1	0	1	0	1	0	0	0	25
F6	65	1	75	2	1	0	1	0	1	0	0	0	25

Similar to understanding the effect of self-reported health status, Figure 3.5.3 shows the model predicted survival rates for the above characteristics and survival rates produced by the RP-2000 Table. In this figure, we can observe three clusters. The bottom cluster refers to the survival rates produced by the RP-2000 Table and as suspected, they tend to under-estimate survival rates. The middle cluster refers to survival rates of males and those with parents whose average age are lower will have worse survival rates. Finally, the third cluster producing the best survival rates refer to the females. Here, parents whose average age are lower will also have worse survival rates.

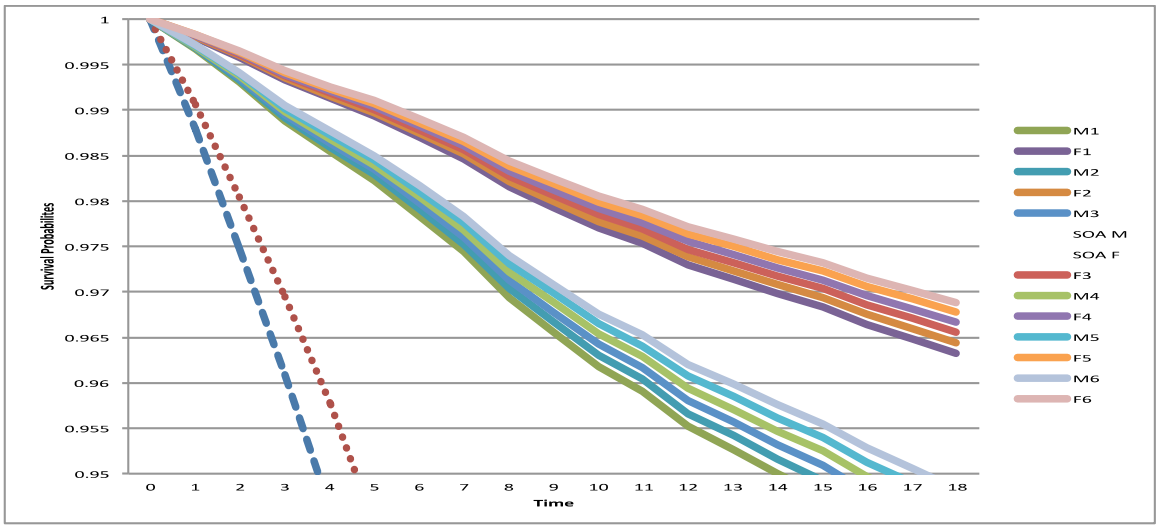


FIGURE 3.5.3: Effects of PAVEAGE on Survival Probabilities

Figure (3.5.4) is self explanatory now that we understand how survival rates affect annuity immediate values.

FIGURE 3.5.4: Annuity immediate values ($)

Chapter 4

Multiple State Models for Long Term Care

4.1 Introduction

The world's older population aged 60 and over is projected to increase dramatically during 2010 to 2100 period; the older population in 2100 is projected to be more than triple that of 2013. Also, by 2100 the oldest-old segment, persons aged more than 80, is projected to reach 7.6 percent of the world population where there were only 1.7 percent of the oldest-old in 2013. For more details about the fastest growing older population, refer to DESA (2013).

As detailed in Ortman, Velkoff, and Hogan (2014), a similar aging trend can be seen in the United States. As the baby boomers are aging, it is projected that the US older (65+) population in 2050 will be 83.7 million as compared to 43.1 million in 2012. Sex and age structure of the US population in 2012, 2030, and 2050 is illustrated

in Figure 4.1.1 and clearly, a dramatical increase of the US older population can be expected over the next few decades.

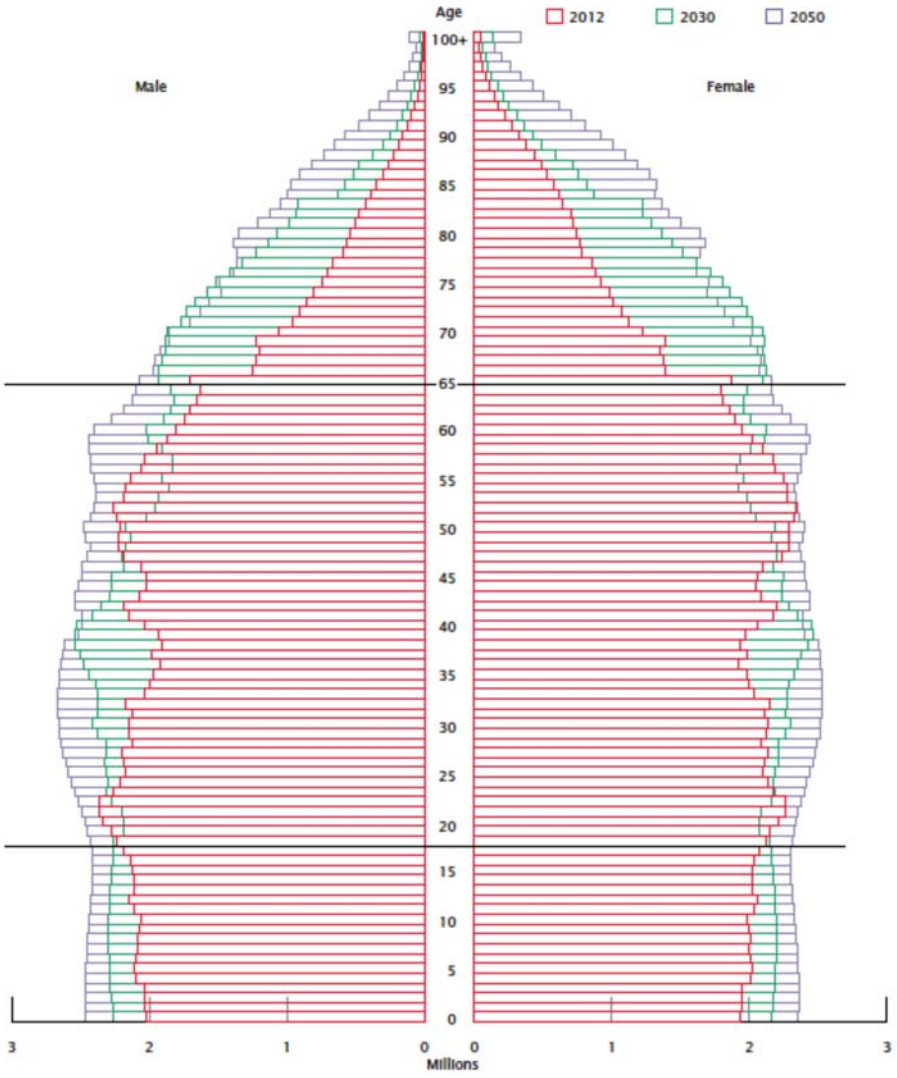


FIGURE 4.1.1: The United States Population Pyramid

Source: U.S. Census Bureau

As population continues to age, allocating resources for the costs associated with aging become much more important both for the individuals and the society as a

whole. With lack of proper planning, the resultant (un-envisaged) costs may impose a serious strain on personal savings as well as government resources. Associated with this aging population is the need for long term care for the elderly, and this includes a wide range of medical requirements and an array of support services such as nursing homes and residential care. The cost of long term care will be some of the burdens caused by aging. To illustrate, elderly individuals usually need assistance due to physical and/or cognitive impairments. At age 65 and over, the probable risk of needing long term care is about 70%; see Kemper, Komisar, and Alecxih (2005). Not surprising to note that at the annual research meeting of the Academy of Health (2005), the cost of long term care was considered one of the largest skyrocketing risk that uninsured retirees may face. In the United States, Medicare is the nation's social insurance program providing health benefits to the elderly; in year 2011 a total of $357 billion was spent for LTC, it spent approximately 21% of the total cost of long term care services. See KAISER (2013). Meanwhile, Medicaid is a federal-state welfare program that provides the nation's primary long term care support and finances approximately 40% of the total cost of long term care. However, due to recent constraints on state budgets, several states are now also limiting the expenses on Medicaid and long term care services. Many of these states are now also advocating the private sector to be prepared to face the rising cost of long term care and to become a dominant partner in the private health care system in the nation. See KAISER (2011) and Spillman (2012).

The main purpose of this chapter is to simply illustrate the use of multiple state models for long term care. The chapter is structured as follows. Section 2 considers the assumptions and notation for modeling. The mathematical ideas and techniques behind the multiple state models for panel data are described in Section 3. Section

4 details the data and the model using a state structure which is fairly similar to the standard state structure of illness-death or disability models typically found in multi-state modeling.

Multiple state models are recent development being used in actuarial practice. As detailed in Dickson, Hardy, and Waters (2009), some of the traditional actuarial applications such as term insurance with increased benefit on accidental death, the permanent disability insurance, the disability income insurance, and the joint life and last survivor insurance can be discussed within the multiple state model framework. Even though multiple state modeling is new to actuarial science, it has been widely used in the areas of medicine (e.g. modeling chronic diseases), geology, zoology, sociology, and economics. An overview of multi-state models is discussed from several points of view in Hougaard (1999), Hougaard and Hougaard (2000), and Andersen and Keiding (2002).

4.2 Assumptions and notation

Let us consider a general multiple state model framework. There is a finite set of $n+1$ states with instantaneous transitions that are possible between selected pairs of states. At some random time T, the individual moves to some other state including the absorbing states for which transition out of these states is not possible. Assume $Y(t)$ denotes the state occupied by time t; see Beyersmann, Allignol, and Schumacher (2012). Let $Y(t) \in \{0, 1, 2, \ldots, n\}$ is the cadlag (right continuous with left limits) stochastic process that denotes state occupied at time t.

The probability distribution of T is typically described in terms of the hazard

function, $\lambda(t)$, which can be defined as

$$\lambda(t) = -\frac{\mathrm{d}\log(S(t))}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{P(T \in [t, t+\Delta t) | T \geq t)}{\Delta t} \tag{4.2.1}$$

The cumulative hazard $A(t)$, distribution function $F(t)$, and survival function $S(t)$ of T can be defined respectively as

$$A(t) = \int_0^t \lambda(u) du, \tag{4.2.2}$$

$$F(t) = P(T \leq t) = 1 - S(t) = 1 - \exp(-A(t)), \tag{4.2.3}$$

and

$$S(t) = P(T > t) = 1 - F(t) = \exp(-A(t)) \tag{4.2.4}$$

As an application of product integrals to survival and hazard functions, Gill and Johansen (1990) has shown that

$$S(t) = \prod_0^t (1 - dA(u)). \tag{4.2.5}$$

As a result of the Fundamental Theorem of Calculus, we can derive $dA(u) = \lambda(u)du$ and using equation 4.2.2, we can write $1 - dA(u) = P(T \geq u + du | T \geq u)$.

Assumption 1 (Markov property): For any states i and j and any time t and $t+s$ where $s \geq 0$, the conditional probability $P[Y(t+s) = j | Y(t) = i]$ is well defined and its value does not depend on any information about the process before time t. That is

$$P[Y(t+s) = j | Y(t) = i] = P[Y(t+s) = j | Y(t) = i, \text{Past}] \tag{4.2.6}$$

Intuitively, it says that future evolutions are completely determined by the current state. How the process arrived to the current state and the time span in the current state is irrelevant.

This assumption is unnecessary to the alive-dead model and the accidental death model because the past of the process is hidden in the current process alive. However this assumption makes more sense for the permanent disability model and the disability income insurance model.

Assumption 2: Assume that for any function $g(h)$ and any positive interval of time h, we have:

$$P[\text{2 or more transitions within a time period of length } h] = \lim_{h \to 0} \frac{g(h)}{h} = o(h) \tag{4.2.7}$$

Assumption 3: For all states i and j and all ages $x \geq 0$, assume that ${}_tp_x^{ij}$, the probability that a life age x in current state i is in future state J at age $x + t$, is a differentiable function of t.

It is important to notice the limitation of any model is that it may be unable to represent the perfect reality of the world.

4.3 Multiple state models for panel data

In order to overcome the existing limitation of the analysis of panel data under a continuous time Markov model, a pioneer work by Kalbfleisch and Lawless (1985) proposed a procedure with a very efficient way of calculating the maximum likelihood estimates using quasi-Newton method that uses the first derivative of log likelihood.

Kay (1986) gave an extension for this procedure within the context of survival studies of cancer disease. Kalbfleisch and Lawless (1985) proposed the following.

There is a finite set of n states with instantaneous transitions being possible between selected pairs of states including transient sates, states out of which transitions are possible, and absorbing states, states out of which transitions are not possible. Let $Y(t) \in \{0, 1, 2, \ldots, n\}$ is a cadlag (right continuous with left limits) stochastic process that denotes state occupied at the time t. Define $\mathbf{P}(\mathbf{s}, \mathbf{t})$ be the $n \times n$ transition probability matrix with entries

$$p_{ij}(s,t) = P[Y(t+s) = j | Y(s) = i] \ , \qquad \text{where } i, j = 1, 2, \ldots, n.$$

The transition rates $q_{ik}(t)$ over very short time period Δt are defined as

$$q_{ik}(t) = \lim_{\Delta t \to 0} \frac{P(T \in [t, t+\Delta t) | T \geq t)}{\Delta t} \ , \qquad \text{where } i, j = 1, 2 \ldots, n.$$

As shown by Cox and Miller (1977) , the relationship between the transition intensities $q_{ik}(t)$ and transition probability $p_{ik}(t, t+\Delta t)$ can be written as

$$p_{ik}(t, t+\Delta t) = q_{ik}\Delta t + o(\Delta t), \quad \text{for } i \neq k$$
$$p_{ii}(t, t+\Delta t) = 1 + q_{ii}\Delta t + o(\Delta t)$$

For the absorbing states i, $q_{ii} = 0$. Using the above two equations, we can see

$$q_{ii} + \textstyle\sum_{k \neq i}^{n} q_{ik} = 0, \qquad i = 1, 2, \ldots, n.$$

which makes q_{ik} a conservative process.

Consider a matrix $\mathbf{Q}$ with dimension $n \times n$ and elements denoted by q_{ij}. For a time-homogeneous model, the independence of t implies that

$$q_{ij}(t) = q_{ij} \qquad \text{for } i, j = 1, 2, \ldots, n,$$

and

$$p_{ij}(t) = p_{ij}(s, s+t) = p_{ij}(0, t),$$
$$P(t) = P(s, s+t) = P(0, t).$$

Using Kolmogorov forward and backward equations by assuming Q is a time-independent matrix, we have

$$p'_{ik}(t) = \sum_j p_{ij}(t) q_{jk} = \sum_k q_{ik} p_{kj}(t) \tag{4.3.1}$$

If $\mathbf{p_i}(t)$ denotes the row vector of probabilities at time t with respect to the initial state i, the equation 4.3.1 leads to

$$\mathbf{p'_{ik}}(\mathbf{t}) = \mathbf{p_i}(\mathbf{t}) \cdot \mathbf{Q} = \mathbf{Q} \cdot \mathbf{p_i}(\mathbf{t}). \tag{4.3.2}$$

By generalizing equation 4.3.2, we can write the matrix $\mathbf{P(t)}$ with (i, j) entries as $p_{ij}(t)$ where

$$\mathbf{P'(t)} = \mathbf{P(t)Q} = \mathbf{QP(t)},$$

with the initial condition that $\mathbf{P(0)} = \mathbf{I}$, where $\mathbf{I}$ is the unit diagonal matrix. By extending the argument for which the transition rates are functions of time, this implies that

$$q_{ij} = q_{ij}(t),$$
$$\mathbf{Q} = \mathbf{Q(t)}.$$

Now the forward equation becomes

$$\frac{\partial}{\partial t}\mathbf{P(s, t)} = \mathbf{P(s, t)}\boldsymbol{Q(t)}, \tag{4.3.3}$$

and the backward equation becomes

$$-\frac{\partial}{\partial s}\mathbf{P}(\mathbf{s},\mathbf{t}) = \boldsymbol{Q}(\boldsymbol{s})\mathbf{P}(\mathbf{s},\mathbf{t}). \tag{4.3.4}$$

Equations 4.3.3 and 4.3.4 are called the Kolmogorov differential equations. The general solution to the Kolmogorov equation admits the form

$$\mathbf{P}(\mathbf{t}) = e^{Qt} = \sum_{m=0}^{n} \mathbf{Q}^m \frac{t^m}{m!} \tag{4.3.5}$$

Since n is finite, the above series is convergent and has unique solutions.

4.4 Data and estimation

A basic long term care multiple state structure which specifies the states and the transitions allowed from state to state is exhibited in the following Figure 4.4.1. The given state structure is fairly similar to the standard state structure of illness-death or disability models in multi-state modeling. Routine activities such as eating, bathing, dressing, toileting, and transferring, which are essential and basic tasks of everyday life, are called Activities of Daily Living (ADL). The number of ADL impairments is a key indicator used to measure "the functional status of a person" especially when it comes to people with disabilities and the elderly; see Crimmins et al. (2009) and Bandeen-Roche et al. (2006). With the available surveys' information in the HRS data, the individual's evidence about the difficulties to perform ADL is primarily used to define the states.

An individual who has difficulties to perform more than 2 ADLs may need some

form of Long Term Care (LTC) services and support such as home care or nursing home care. For each wave in the HRS data, State 1 represents the individual's situation with ADL difficulties less than or equal to 2, and for people who are in this state, it is assumed that they do not need any form of LTC support. On the other hand, HRS individuals who are in State 2 have more than 2 ADL difficulties, and for people who are in this state, it is assumed that they need some form of LTC support. Thus, individuals who are in State 2 are in some form of LTC facility.

Finally, State 3 is considered to be the absorbing state which is death. The arrows in Figure 4.4.1 indicate the possible transitions. As displayed on this figure, allowed transitions are from State 1 to State 3, State 2 to State 3, State 1 to State 2 and vice versa. As some studies show, e.g. de Leon et al. (1999) and Chemerinski, Robinson, and Kosier (2001), ADL recovery is possible with age so that transition from State 2 to State 1 is also considered.

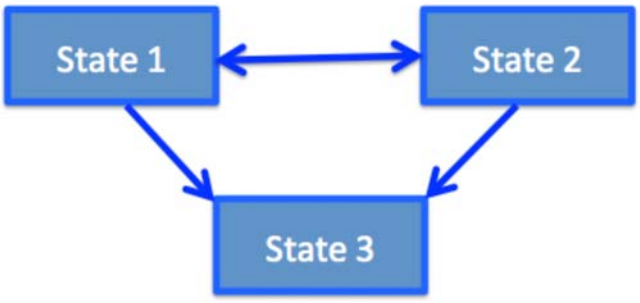


FIGURE 4.4.1: Multiple State Structure for Long Term Care

The U.S.Census Bureau (2014) report is one of the best recent sources that provides investigation about the fundamental aspects of aging Americans. The findings in the report are incorporated heavily on the data from 2010 Census, Current Population Survey, American Community Survey and National Health Interview Survey.

A few of the key highlights which motivated the variable selection in our modeling are the following:

- Changing Marital Trends: Noticeable divorce rate increases can be seen among the older population over the past few decades. As a consequence, older population may even need more social support than before because now more of them live alone.

- Increasing Education Level: A sharp increase in education attainment can be seen among both older men and women. Educated older population tend to live longer life.

- Increasing Overweight or Obesity: Approximately 30% of older men and women are considered obese and nearly 40% are considered overweight during the period of $2003 - 2006$. Being obese is associated with high risk for ADL difficulties.

- Prevalence of Chronic Diseases: It has been reported that about 40% older Americans in year 2008 have three or more chronic illnesses such as heart disease, hypertension, diabetes, cancer, and osteoporosis. Some chronic diseases may cause for shrinking the independence of older people.

Some of the research works that confirmed the above findings are in Kramarow et al. (2007), Bentler et al. (2009), and Håkansson et al. (2009). In addition to the above mentioned variables, there are a number of studies that assessed the genetic influence and the effect of mental health in human lifespan. Indeed, some of these research studies assessed the importance of genetic contributions towards human longevity especially at more advanced ages. See Hjelmborg et al. (2006), Schachter et al. (1994).

There are also ample evidences, e.g. Branch and Jette (1982) and Black, Rabins, and German (1999), supporting that mental or emotional disorder is one of the significant variables for LTC institutionalization. The Center for Epidemiological Studies Depression Scale (CESD) is widely used as a standard and most useful tool to measure the mental or emotional disorders in epidemiological studies; see Radloff (1977). The CESD scale reported in the HRS data is a sum of five negative indicators (depression, everything is an effort, sleep is restless, felt alone, and felt sad and could not get going) minus two positive indicators (felt happy and enjoyed life). As a consequence, we have incorporated the information about parent average age to measure genetic influence, and CESD scales to measure mental or emotional health, for LTC modeling.

We extracted the information of HRS individuals whose ADL information were available. Table 4.4.1 summarizes this available information. There are nearly $41,000$ non-missing states recorded from about $6,200$ HRS individuals over the period $1998-2011$ and considered for our modeling work as described. As shown in the table, the data includes:

- a slightly larger number of male responders (53%) than female (47%);
- more than 75% individuals who have at least graduated from high schools;
- only 35% of individuals who are living alone;
- a comparably large proportion (68%) of individuals' parents live or lived beyond age 70; and
- only less than 10% who are underweight.

TABLE 4.4.1: Data Characteristics and Other Observable Information

Categorical Variables	Description		Proportions
RAGENDER	Gender of the respondent:	Male=0	53.30%
		Female=1	46.70%
RAEDUC	Education:	College and above= 0	39.28%
		High-school graduate = 1	36.39%
		Lt High-school = 2	24.33%
RMARRY	Current Partnership Status:	Single=0	35.34%
		Married/Partnered=1	64.66%
STATE	State indicator:	ADL$\leq$ 2=State 1	90.81%
		ADL$>$ 2=State 2	5.06%
		Death=State 3	4.12%
CESD	Mental Health Index	None = 0	53.93%
	sum of depression, effort, restless sleep, alone, sad, (1- happy), (1- enjoy life):	At most two= 1	27.47%
		More than Three = 2	18.60%
CONDE	Sum of conditions (high blood pressure, diabetes, cancer, lung disease,	None = 0	28.50%
	heart attack, stroke, psychiatric problems, arthritis) ever had:	At most two = 1	49.88%
		More than Three = 2	21.61%
BMI	Body Mass Index:	Underweight($<$ 18.5)=0	8.43%
		Normal weight (18.5-24.9)=1	27.08%
		Overweight of Obesity (25 $\geq$)=2	64.49%
PAVAGE	Parents Average Age:	Less than 70=0	31.57%
		More than 70=1	68.43%

Continuous Variables		Minimum	Mean	Maximum
AGE	Age of the respondent	50	62	89

For illustration purposes, a series of observations grouped by a set of HRS respondents with time independent and dependent covariates are listed in Table 4.4.2. As shown in this table, HRS respondent with ID 1010 was in State 1 at the beginning of the survey ($t = 0$) and transitioned to State 3 at $t = 2$. As yet another example, HRS respondent with ID 11479010 was in State 1 during the time $t = 0 - 8$, moved to State 2 at $t = 10$ and stayed there for 4 years before finally moving to absorbing state (State 3) at $t = 14$. Finally, the last person shown in Table 4.4.2 has been living in State 1 for the time period from $t = 0 - 14$.

TABLE 4.4.2: HRS Sample for the Long Term Care Study

HHIDPN	START	AGE	GENDER	PAVEAGE	RMARRY	RAEDUC	CESD	CONDE	BMI	STATE
1010	0	54	M	0	0	0	0	3	31	1
1010	2	56	M	0	0	0	0	3	24	3
11479010	0	50	F	0	1	0	0	1	51	1
11479010	2	53	F	0	1	0	0	1	38	1
11479010	4	55	F	0	0	0	2	1	40	1
11479010	6	56	F	0	1	0	1	1	35	1
11479010	8	58	F	0	1	0	2	1	63	1
11479010	10	61	F	0	1	0	6	3	36	2
11479010	12	63	F	0	1	0	4	3	41	2
11479010	14	64	F	0	1	0	1	3	36	3
11218010	0	56	F	1	0	1	4	2	24	2
11218010	2	58	F	1	0	1	4	2	23	2
11218010	4	61	F	1	0	1	2	2	22	3
10989010	0	56	M	1	1	2	1	1	26	1
10989010	2	58	M	1	1	2	1	2	27	1
10989010	4	60	M	1	1	2	2	2	27	1
10989010	6	62	M	1	1	2	4	2	27	1
10989010	8	64	M	1	1	2	1	2	27	1
10989010	10	66	M	1	1	2	2	2	27	1
10989010	12	68	M	1	1	2	1	2	27	1
10989010	14	70	M	1	1	2	2	2	26	1

Especially in academia, R is a widely used free software programming language which has more advanced packages for statistical analysis. The illustrative models that are built in this chapter are fitted using the R package msm which specializes

in the estimation of continuous time Markov models; see Jackson (2011). The msm package in R allows time varying or constant explanatory variables as well as censoring that is included in transition intensities.

Table 4.4.3 summarizes the frequency of the multi-state data used in our study over the years $1992 - 2011$. There are 1,410 deaths from State 1 (ADL ≤ 2) and 252 deaths from State 2 (ADL > 2). Furthermore, there are 926 transitions recorded from State 1 to State 2 and another 760 transitions from State 2 to State 1.

TABLE 4.4.3: Frequency for Consecutive States

From	State 1	State 2	Death
State 1	$33,826$	926	$1,410$
State 2	760	$1,006$	252

One possible initial intensity matrix 4.4.1 with the allowed transitions which can be used as initial starting point to search for global maximum likelihood estimates is given by

$$\begin{pmatrix} 0 & 0.250 & 0.250 \\ 0.166 & 0 & 0.166 \\ 0 & 0 & 0 \end{pmatrix} \tag{4.4.1}$$

For illustrative purposes, we have fitted two models using the msm package under the assumptions that time dependent explanatory variables are constant between the waves and the exact death year is known. In our study, we considered two models: Model 1 and Model 2. In Model 1, the intensities are based only on two basic variables, one of which is age, a time varying variable, and another one is gender, a constant

variable. Model 2 includes some extended covariates such as education level, marital status, parent's age, mental health indices, health condition indices and body mass index, in addition to age and gender. In model 2, only gender and education level are the constant variables and all the other variables are considered time varying covariates.

Quasi-Newton optimization algorithm (also called a variable metric algorithm) is considered to explore the global maximum likelihood. To ensure the global maximum has been found rather than local, $10,000$ iterations are considered and log-likelihood is rescaled by $23,000$ because -2 x log-likelihood is around $23,000$. Convergence criteria has tightened to $1e-16$ to solve "false convergence".

For individual i at the time j, the transition intensity in Model 1 with age and gender (z_{ij}) as the explanatory variables is given by

$$q_{rs}(z_{ij}) = q_{rs}^{(0)} \exp(\beta_{rs}^T \mathbf{z}_{ij}) \tag{4.4.2}$$

where r, s are states, $r = 1, 2, 3$ and $s = 1, 2$, $q_{rs}^{(0)}$ and β_{rs} can be found using maximum likelihood estimates. The maximum likelihood estimates of model 4.4.2 are displayed in Table (4.4.4). According to this table, it indicates that deaths from State 2 are about 11% and deaths from State 1 are about 1.5%. This implies that increasing ADL difficulties increases the risk for death while controlling for age and gender.

TABLE 4.4.4: Maximum Likelihood Estimates of Model 1

From	State 1	State 2	Death
State 1	−0.03669	0.02185	0.01484
State 2	0.27935	−0.38664	0.10729

The corresponding hazard ratios (HR), $\exp(\beta_{rs})$, and 95% confidence intervals (CI) for model 4.4.2 for age are given in Table 4.4.5. It shows that after controlling for gender, 1 year of age increase is associated with 1.6% risk to transfer from State 1 to State 2, 7.4% risk from State 1 to death, and 4.8% risk to transfer from State 2 to death.

TABLE 4.4.5: For Age: Hazard Ratios and Confidence Intervals

Transition	HR	CI-Low	CI-Upper
State 1 - State 2	1.0158814	1.0025649	1.0293748
State 1 - Death	1.0746513	1.0610410	1.0884363
State 2 - State 1	0.9678685	0.9547287	0.9811892
State 2 - Death	1.0480366	1.0246385	1.0719690

The corresponding hazard ratios, $\exp(\beta_{rs})$, and 95% confidence intervals for model 4.4.2 for gender are given in Table 4.4.6. It shows that after controlling for age and when compared to male, female are associated with 63% high risk to transfer from State 1 to State 2, 27%(100% − 63%) lower risk to transfer from State 1 to death, and 20% lower risk to transfer from State 2 to death. This indicates

- that there is greater female exposure to more ADL difficulties than males;
- that male recovery rates are higher than female; and
- that female stay longer in State 1 or 2 before death.

TABLE 4.4.6: For Gender: Hazard Ratios and Confidence Intervals

Transition	HR	CI-Low	CI-Upper
State 1 - State 2	1.6290613	1.3997271	1.8959702
State 1 - Death	0.6309823	0.5258537	0.7571282
State 2 - State 1	0.8251634	0.7079921	0.9617264
State 2 - Death	0.8017438	0.5898918	1.0896798

In the estimation for Model 2, several models were initially tested with the extended variables such as education level, marital status, parent's average age, mental health indices, health condition indices, body mass index, in addition to age and gender. The likelihood ratio test which is a statistical test to compare two different models based on the likelihood functions of the models is used to select the best model. The hazard ratios for the best selected model are displayed in Table 4.4.7.

TABLE 4.4.7: Hazard Ratios for Model 2

Transition	AGE	GENDER	CONDE	BMI	RAEDUC	CESD	PAVEAGE
State 1 - State 2	0.9913886	1.1311527	2.4695750	1.1260210	1.4711665	1.9081600	0.8462561
State 1 - Death	1.0520725	0.4907620	2.2074568	0.5451874	1.2768869	1.2016374	0.8193125
State 2 - State 1	0.9713729	0.8176575	0.8645839	1.2256513	0.9929570	1.0505129	1.0987310
State 2 - Death	1.0395031	0.9885445	1.7946957	0.7015372	0.8529524	0.7111332	0.8777219

As shown in Table 4.4.7, after adjusting for other variables, age and gender effect show a lower intensity for the transitions while as expected, worse health indices show higher hazard rates to transition between states. For example, BMI shows high risk to transfer from State 1 to State 2 but lower risk with other transitions especially from State 1 to Death and from State 2 to Death. Furthermore, as the average parents age increases, lower hazard can be observed.

The corresponding fitted transition probability matrix at the time $t = 2$ is shown in Table 4.4.8 and values inside the parentheses are confidence intervals derived from the asymptotic normal distribution of the estimates. For example, it indicates an individual in State 1 has 95.6% chance to stay in State 1, 2.0% chance to move to State 2, and 2.4% probability of dying in two year's time. For an individual in State 2, there is a 4.4% probability for recovery from ADL, 4.5% chance to stay in the same state, and 11.2% probability of dying in two year's time.

TABLE 4.4.8: Transition Probability Matrix

From	State 1	State 2	Death
State 1	0.95622 (0.95376,0.95886)	0.02003 (0.01821,0.02186)	0.02376 (0.02190,0.02586)
State 2	0.43800 (0.39680,0.47696)	0.44979 (0.41066,0.48762)	0.11222 (0.08463,0.14774)

To investigate the quality of our model fit, we examined a comparison of the Kaplan-Meier survival curves between Models 1 and 2. The Kaplan-Meier curve is a standard and popular nonparametric method that is used to compare survival curves. Figure 4.4.2 shows the Kaplan-Meier curves resulting from fitting the data assuming Model 1 and separately, assuming Model 2. As displayed in the figures, up until year 10, Model 2 with extended covariates fits the survival better than Model 1 which only

controls for age and gender. However after year 10, Model 1 estimates the survival better than Model 2. Some overestimates can be seen from Model 2 after year 10.

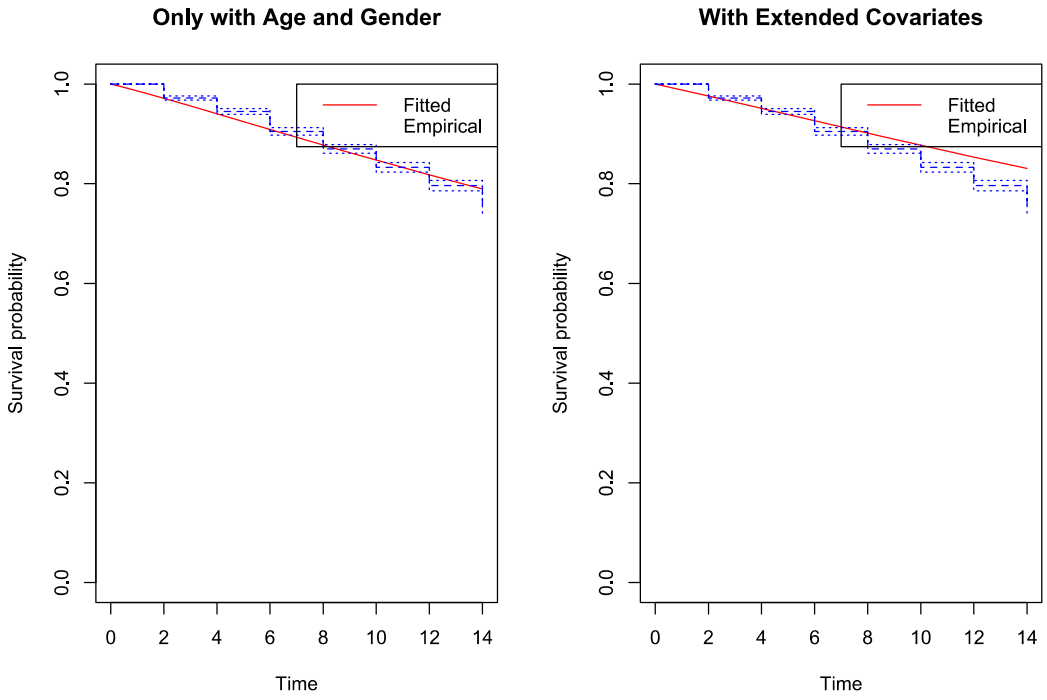


FIGURE 4.4.2: Comparing Survival Curves from Model 1 and Model 2

We also examined test statistics. Pearson-type goodness of fit test proposed by Aguirre-Hernández and Farewell (2002) is calculated to assess the two models. This test sadly indicates that neither Model 1 nor Model 2 sufficiently fit the data. However, when the likelihood ratio test is used to compare between Models 1 and 2, we find that Model 2 significantly performed better than Model 1.

As noted earlier, this is only for illustration purposes. For more details about additional model assessments for panel data like Kaplan-Meier estimates for basic feasibility of the model, local score test which measures the time inhomogeneity of a

transition intensity, general model fit measures by Pearson goodness of fit, and others, please refer to Titman and Sharples (2009).

Chapter 5

Life Insurance Policy Termination and Survivorship

5.1 Introduction

When a life insurance contract terminates due to voluntary non-payment of premiums, there is a possible hidden cost resulting from mortality antiselection. This refers to the tendency of policyholders who are generally healthy to select against the insurance company by voluntarily terminating their policies.

This chapter focus on policy termination together with understanding the survivorship pattern resulting from terminated policies. Our observable data is an extract from a real life data of a portfolio of terminated life insurance policies from an undisclosed insurance carrier which tracked the mortality dates of these policies from the U.S. Social Security Administration office. The primary purpose of obtaining such information is to first understand the relationship between policy termination

and mortality, and later, more importantly, to assess the financial implication of this relationship in the design, pricing and risk management of insurance products. These terminated policies are as of a fixed date, hereby undisclosed to preserve some level of confidentiality. The recorded death date information is also as of this same fixed date which is then considered the censoring date used in our model calibration. Our data file also recorded period around 1920's as the year with the earliest policy issue date in the portfolio. On the aggregate, we have observations totaling to 65,435 terminated single life policies, discarding joint life policies for purposes of our analysis. This set of observations that we use for model calibration in this paper is only a random sub-sample from the insurer's portfolio used in their analysis.

The type of observations in our empirical data is vividly illustrated in Figure (5.1). According to this figure, we observe two distinctly classified policyholders, herewith labeled policyholders 1 and 2, where in both cases, we observe the times when each withdraw their policy out of the insurance company. Policyholder 1 dies before the end of the observation period and therefore we can observe its time from withdrawal until time of death. On the other hand, policyholder 2 is still alive at the end of the observation period and is therefore clearly considered a right-censored observation.

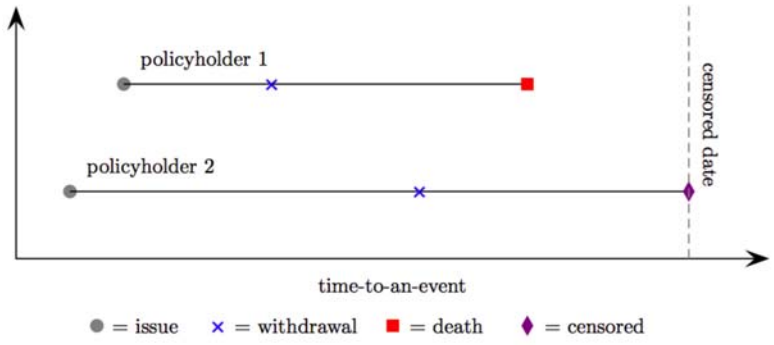


FIGURE 5.1.1: Illustrative diagram of the observed times until withdrawal and death

Given a policy is issued at a fixed and known age, denote this by z, we are interested in estimating the probability distribution of the time-until-withdrawal and the time-until-death from issue. Denote these times to events, respectively, by the random variates T_w and T_d, and define the difference as $T_{wd} = T_d - T_w$. Our data file allows us to observe T_w and the conditional random variate $T_{wd}|T_w$, or effectively $T_d|T_w$ since $(T_{wd}|T_w) = (T_d|T_w) - T_w$. For notation purposes, we can express $T^*_{wd} = T_{wd}|T_w$ and $T_{d|w} = T_d|T_w$. Notice that because not all policies were followed up until their times of death, censoring is therefore present and the observable T_d is therefore calculated as of the censoring date, as previously explained, and a censoring variable is recorded for each of the policies in the portfolio. These are indeed called right-censored observations which are typical in mortality studies. See, for example, Elandt-Johnson and Johnson (1980). Of our entire observations, we found that we have a total of 61,889 right-censored observations. Slightly over 5% of our observations are deaths, something not atypical of mortality follow-up studies.

The nature of our observed data allows us to calibrate models at a micro-level, meaning observations are at the policyholder level. Many actuarial models are developed based on grouped data, but there is more information obtained at a micro-level allowing us to better reflect reality. With micro-level data becoming more available especially to practitioners, there is an increasing trend of developing micro-econometric models, a term used for example by Gourieroux and Jasiak (2007). Next, we use a general class of duration models to specify the parametric distribution of the time-until-withdrawal, T_w, from issue date. This class falls within the general framework of regression models for which the distribution of the error component can in some sense be arbitrarily specified. Not only is this class of models very tractable, but they apparently allows us to incorporate covariate terms within which in our context are

policyholder characteristics such as gender, issue age, and product type. Duration models are commonly used in the field of econometrics, e.g. Gourieroux and Jasiak (2007) and Lancaster (1990). We find that for our data, the most suitable duration model for T_w is one where the error component follows a standard Gamma distribution, and we shall observe later that this indeed results in a Generalized Gamma distribution specification for T_w. Other parametric error distributions were also examined but these models provided weak statistical support to the data.

5.2 Parametric models

5.2.1 A class of duration models for time-until-withdrawal

Consider the time-until-withdrawal random variate, T_w, referring to the duration that the policyholder lapses from date of issue, which clearly has a range of non-negative values. We shall denote its survival, distribution and density functions by S_w, F_w and f_w, respectively. These functions are related, for example, as follows:

$$S_w(t) = P(T_w > t) = 1 - F_w(t) = \int_t^\infty f_w(s)ds.$$

Suppose that we can write T_w as

$$T_w = \exp(\mu)T_0^\sigma, \tag{5.2.1}$$

for some non-negative random variate T_0. By re-writing (5.2.1) through the log-transformation

$$\log(T_w) = \mu + \sigma \log(T_0) = \mu + \sigma\Lambda, \tag{5.2.2}$$

where $\Lambda = \log(T_0)$, we observe that μ is a location parameter and σ is a scale parameter with the restriction that $\sigma \neq 0$ in order to avoid a degenerate distribution for T_w. Because we can write the survival function of T_w as

$$S_w(t) = \begin{cases} S_\Lambda\left(\dfrac{\log(t) - \mu}{\sigma}\right), & \text{for } \sigma > 0 \\ 1 - S_\Lambda\left(\dfrac{\log(t) - \mu}{\sigma}\right), & \text{for } \sigma < 0 \end{cases} \tag{5.2.3}$$

where S_Λ denotes the survival function of Λ, the distribution of T_w indeed belongs to a log-location-scale family of distributions.

Covariates can be introduced through the location parameter μ. Suppose x is a vector of covariates, such as policyholder characteristics, and β, the corresponding vector of linear coefficients. Then we can simply replace $\mu = X'\beta$. For example, (5.2.1) becomes

$$T_w = \exp(X'\beta)T_0^\sigma, \tag{5.2.4}$$

and (5.2.2) becomes

$$\log(T_w) = X'\beta + \sigma \log(T_0) = X'\beta + \sigma\Lambda, \tag{5.2.5}$$

which generalizes the familiar ordinary regression model where the error component has a Normal distribution. The specification in (5.2.4) is also a special case of the Accelerated Failure Time (AFT) model commonly studied in survival models. See

Elandt-Johnson and Johnson (1980).

It is also straightforward to find the distribution of T_w in terms of the distribution of T_0. The survival distribution function of T_w can be expressed as

$$S_w(t) = S_0\big((e^{-\mu}t)^{1/\sigma}\big) \tag{5.2.6}$$

and its density as

$$f_w(t) = \frac{1}{|\sigma|t}(e^{-\mu}t)^{1/\sigma}\, f_0\big((e^{-\mu}t)^{1/\sigma}\big)\,, \tag{5.2.7}$$

where S_0 and f_0 are respectively the survival and density functions of T_0. Within this class of models, it is oftentimes more straightforward to specify the distribution of T_0 rather than of its logarithm.

Example 1. (Log-Normal Distribution) As an illustration, in the case where T_0 has a log-normal distribution with parameters 0 and 1, it can be shown that

$$f_w(t) = \frac{1}{\sqrt{2\pi}\sigma t}\exp\left[-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right] \tag{5.2.8}$$

which also gives T_w a log-normal distribution with parameters μ and σ, where $\sigma > 0$. This distribution is well-studied both in finance and actuarial science.

Example 2. (Generalized Gamma Distribution) Here, we suppose that T_0 has a standard Gamma distribution, i.e. one where the scale parameter is 1 but with a shape parameter m so that its density can be expressed as

$$f_0(y) = \frac{1}{\Gamma(m)}y^{m-1}e^{-y}.$$

It can be shown that

$$f_w(t) = \frac{1}{|\sigma|t}\frac{1}{\Gamma(m)}(e^{-\mu}t)^{m/\sigma}\exp\left[-(e^{-\mu}t)^{1/\sigma}\right]. \tag{5.2.9}$$

This gives a large class of distributions called the *Generalized Gamma* with parameter vector (μ, σ, m). For a member of this class, we shall write it as $T_w \sim \text{GG}(\mu, \sigma, m)$. This family of distributions which is often attributed to Stacy (1962), includes as special cases the Gamma, Exponential, Log-Normal and Weibull distributions. Despite its flexibility, this family of distribution is less-studied in finance and actuarial science.

Example 3. (GB2 Distribution) Here, we suppose that T_0 has a Beta of the second kind (B2) distribution whose density is expressed as

$$f_0(y) = \frac{1}{B(\gamma_1, \gamma_2)}\frac{y^{\gamma_1 - 1}}{(1+y)^{\gamma_1+\gamma_2}}.$$

This type of distribution is sometimes called the standard form of a Pearson Type VI distribution, see Johnson, Kotz, and Balakrishnan (1995). It can be shown that

$$f_w(t) = \frac{1}{|\sigma|t}\frac{1}{B(\gamma_1, \gamma_2)}\frac{(e^{-\mu}t)^{\gamma_1/\sigma}}{\left[1+(e^{-\mu}t)^{1/\sigma}\right]^{\gamma_1+\gamma_2}}. \tag{5.2.10}$$

This gives a large class of distributions called the GB2, or Generalized Beta of the second kind, with parameter vector $(\mu, \sigma, \gamma_1, \gamma_2)$. For a member of this class, we shall write it as $T_w \sim \text{GB2}(\mu, \sigma, \gamma_1, \gamma_2)$. This family of distributions was first studied by McDonald (1984) and has been applied in modeling insurance claims, e.g. Cummins et al. (1990). As pointed out by Sun, Frees, and Rosenberg (2008), it is well suited for fitting heavy-tailed data.

5.2.2 Survival models for the age at death random variable

Denote the (fixed) issue age by z and let X_d be the age at death random variable so that

$$X_d|z = z + T_w + (T_d - T_w) = z + T_w + T_{wd},$$

provided $T_{wd} > 0$. Thus, if T_w is known, then

$$(X_d|z, T_w = t_w) = z + t_w + T_{wd}.$$

Thus, it becomes clear that

$$\begin{aligned} P(T_{wd} > t_{wd}|z, T_w = t_w) &= P(T_d > T_w + t_{wd}|z, T_w = t_w) \\ &= \frac{P(X_d > z + t_w + t_{wd})}{P(X_d > z + t_w)} \\ &= \frac{S_d(z + t_w + t_{wd})}{S_d(z + t_w)}. \end{aligned} \tag{5.2.11}$$

Here, S_d refers to the corresponding survival distribution function of the age at death, X_d, random variable. That is, $S_d(x) = P(X_d > x)$. Driven by the observable data, we need to specify the distribution model for T_{wd}, given the issue age z and the time-until-withdrawal t_w. According to (5.2.11), this is equivalent to specifying the distribution model for the age at death random variable X_d. While we have examined several survival models for the age at death, the two models, both of which are commonly known to actuaries, are the Gompertz and the Weibull distributions.

Example 1. (Gompertz Distribution) For the Gompertz distribution, we write its

survival function in the form

$$S_d(x) = \exp\left[e^{-m^*/\sigma^*}\left(1 - e^{x/\sigma^*}\right)\right], \tag{5.2.12}$$

where $m^* > 0$ is the mode and $\sigma^* > 0$ is a dispersion measure about this mode of the distribution. This reparameterization has been suggested by Carriere (1992) and is being used here both for ease of parameter interpretation and estimation. By re-expressing the parameters with

$$B = \frac{1}{\sigma^*}\exp(-m^*/\sigma^*) \text{ and } c = \exp(1/\sigma^*), \tag{5.2.13}$$

it leads us to the hazard function

$$\mu_x = \frac{f_d(x)}{S_d(x)} = Bc^x.$$

This simple expression is quite familiar to actuaries and has been well-studied in the actuarial literature. See, for example, Gompertz (1825), Carriere (1992), Frees, Carriere, and Valdez (1996), and Bowers et al. (1986). There is an additional interesting property of the Gompertz model that may allow us to take advantage of especially when graphically examining candidate models for survival data. Consider the probability that an individual, now age x, will survive another year:

$$p_x = \frac{S_d(x+1)}{S_d(x)} = \exp\left[e^{(x-m^*)/\sigma^*}(1 - e^{1/\sigma^*})\right]. \tag{5.2.14}$$

It becomes rather straightforward to see that $\log(\log(p_x))$ is linear in age x, that is,

$$\log(\log(p_x)) = a + bx, \tag{5.2.15}$$

where clearly $a = \log(1 - e^{1/\sigma^*}) - (m^*/\sigma^*)$ and $b = 1/\sigma^*$. See Valdez (2000).

Example 2. (Weibull Distribution) Here, we express the Weibull survival distribution as

$$S_d(x) = \exp\left[-(x/m^*)^{m^*/\sigma^*}\right]. \tag{5.2.16}$$

The parameters $m^* > 0$ and $\sigma^* > 0$ are respectively location and dispersion parameters. While this reparameterization has been suggested by Carriere (1992) in the actuarial literature, this distribution is even more widely familiar in survival analysis and reliability theory. Conducting a preliminary investigation of the possible quality of a Weibull model to a survival data is sometimes done visually using the so-called Weibull plot. It can be shown that

$$\log(-\log(S_d(x))) = (-m^*/\sigma^*)\log(m^*) + (-m^*/\sigma^*)\log(x) = c + d\log(x), \tag{5.2.17}$$

where $c = (-m^*/\sigma^*)\log(m^*)$ and $d = -m^*/\sigma^*$, and is clearly linear in the logarithm of x. Thus, the Weibull plot is the scatter plot of $\log(-\log(S_d(x)))$ against $\log(x)$.

For either the Gompertz or Weibull model, we injected observable covariate characteristics through either (or both) the location parameter m^* or the scale parameter σ^*.

In order to investigate the robustness of the models, other parametric survival models were examined but we found that these did not adequately fit our data. For comprehensive purpose, especially for readers who wish to investigate such other

models, we make a list of some in Table 5.2.1. As a matter of fact, it has also been suggested in both Carriere (1992) and Valdez (2000) that mixing some of these survival models provide a better quality fit of survival data over the entire human lifetime.

TABLE 5.2.1: Some other parametric survival distribution models

Distribution	Survival Function $S_d(x)$	Force of Mortality μ_x
Exponential	$\exp(-\mu x)$	μ
Inverse-Gompertz	$\dfrac{1-\exp\left[-e^{(x-m)/\sigma}\right]}{1-\exp(-e^{m/\sigma})}$	$\dfrac{1}{\sigma}\exp\left(\dfrac{x-m}{\sigma}\right)$
Inverse-Weibull	$1-\exp\left[-(x/m)^{-m/\sigma}\right]$	$\dfrac{(1/\sigma)(x-m)^{-(m/\sigma)-1}}{\exp[(x/m)^{-m/\sigma}]-1}$

5.3 Data characteristics

In our empirical investigation in this article, we analyzed data drawn from a major insurer's portfolio of terminated single life insurance contracts with mortality dates tracked from the U.S. Social Security Administration office. On the aggregate, we drew a randomly selected sub-sample of 65,435 such terminated policies; although the sub-sampling algorithm used results in a random sample, we carefully made the draw in order to preserve the overall characteristics of the insurer's portfolio. To begin, we have three main product classifications herewith labeled PlanTypeP, PlanTypeT and PlanTypeO. PlanTypeP consisted of the traditional participating whole life insurance policies and is approximately 42.4% of the entire sample. PlanTypeP

consisted of traditional term insurance products and is approximately 28.0% of the entire sample. For the rest, approximately 29.6%, are PlanTypeO which primarily consisted of conventional Universal Life, although we have very little policies that were term conversion which were grouped into this classification. Term conversion policies are those initially purchased as traditional term contracts that later converted into some form of permanent policies. Similar such proportions for plan types have been observed from the insurer's entire portfolio.

Table 5.3.1 provides for a summary of the policy characteristics in our data together with other interesting observable information that we later find useful predictor variables.

To illustrate, gender and smoker categories are included in our data files. It is well known that there are significant mortality differentials between males and females, with females generally living longer than males. We have significantly more males in our data than females, with roughly a ratio of almost 2 to 1. In addition, because of previous medical studies, it has become an acceptable premise that smoking does affect mortality. These findings are again reinforced in the results of our empirical work and our data have roughly five times more non-smokers than smokers. Finally, we also have 21% of our observations classified as combined smoker and non-smoker; these refer to those policies classified as unismokers. Virtually, unismoker policies refer to those insurance contracts with premiums rated regardless of smoking habits.

For each contract observed, we have the policy effective or issue date, the withdrawal date and the date of death, if applicable. Policies with no observable date of death are considered censored observations, with a fixed and known censoring date, herewith being a withheld information to preserve confidentiality. These dates allow us to measure the duration from issue to policy withdrawal, and given this duration

TABLE 5.3.1: Policy characteristics and other observable information

Categorical variables	Description		Proportions	
PlanType	Type of insurance plan:	PlanTypeP	42.4%	
		PlanTypeT	28.0%	
		PlanTypeO	29.6%	
RiskClass	Insured's assigned risk class:	RiskClass = N	72.0%	
		RiskClass = Y	28.0%	
Sex	Insured's sex:	Male = 1	65.2%	
		Female = 0	34.8%	
Smoker	Smoker class:	Non-smoker = N	66.6%	
		Smoker = S	12.4%	
		Combined = C	21.0%	
Censor	Censoring indicator for death:	Censor = 1	94.6%	
		Censor = 0	5.4%	
Continuous variables		Minimum	Mean	Maximum
IssAge	The policyholder's issue age	0	37.70	89.65
Face Amount	The policy's insured amount	1	213,000	60,000,000
Temp FEAmt	Temporary flat extra amount (per 1000)	0.00	0.08	49.00
Perm FEAmt	Permanent flat extra amount (per 1000)	0.00	0.06	48.00
MEFact	Extra mortality factor	1.00	1.01	4.00
Dates				
IssDate	Policy effective or issue date			
BDate	Insured's date of birth			
WDate	Policy withdrawal or lapse date			
DDate	Insured's date of death, if applicable			

of withdrawal, the time of death, if policy is uncensored, or the time from withdrawal till the censoring date, if policy is censored. Our policy records indicate 61,889 of the total 65,435 observations are censored, representing about 95% of the policies in the data.

When insurance policies are underwritten prior to issue, the insurer may find additional or extra hazard, such as certain lifestyle or past illness, for which the insurer may be willing to assume but for obviously with a premium differential or extra cost. Insurers price for these costs with degrees and methods, and according to our

records, ours increases the mortality assumption in the premium calculation with an extra mortality factor and/or assesses a flat extra premium on either a temporary or permanent basis. Several of our policies were priced with little or no extra hazard. However, for those that were subjected to such premium differentials, the extra mortality factor used in the premium calculation ranged from as little as just slightly 1% to as high as above 400% of that presumably used for standard policies.

TABLE 5.3.2: Number of policies and average face amount by type of plan, sex and issue age

	Issue Age								**Total**
	Males				**Females**				
Plan Type	**$\leq$ 30**	**30-50**	**50-70**	**$>$ 70**	**$\leq$ 30**	**30-50**	**50-70**	**$>$ 70**	
PlanTypeP									
Count	6,461	8,476	2,300	100	4,401	4,545	1,374	119	27,776
Face Amount	46,766	152,345	139,624	213,028	35,611	103,401	150,228	213,891	100,605
PlanTypeT									
Count	1,130	9,557	1,963	20	964	4,262	434	3	18,333
Face Amount	323,955	475,092	653,320	1,461,250	168,350	251,603	408,421	425,833	416,264
PlanTypeO									
Count	2,076	7,314	3,091	188	1,516	3,789	1,103	249	19,326
Face Amount	124,896	193,958	203,519	445,704	79,893	133,510	310,929	604,947	181,690

Finally, Table 5.3.2 provides an interesting summary of the number of policies together with the average face amounts according to type of plan, gender and issue age. For this purpose, it was meaningful to partition issue age according to 4 groups: less than or equal to 30, between 30 and 50, between 50 and 70, and above 70. Of our total 65,435 policies, we find that the overall average face amount is $212,992. As earlier noted, roughly a bulk of our data are PlanTypeP policies. However, it is interesting to note that PlanTypeT policies tend to have much larger face amounts, with its average more than 4 times the average of that for PlanTypeP policies. The over-

all average face amount for PlanTypeP is 100,606 while it is 416,264 for PlanTypeT policies. For all types of plan, most issue ages are in the range of 30 through 50 years old, especially so for PlanTypeT. This could either be the result of the choice of the policyholder or that of the insurer. By design, term life insurance products tend to have premiums that exponentially increase with age so that it is not surprising to find fewer policies in the above 50 age categories. What is a little bit surprising is to find fewer policies in the younger age categories; such may be the result of a marketing strategy by our insurer.

Mortality studies for insurance contracts tend to account for the impact of policy face amount by using them as weights. For our purposes, because our observations are at the policyholder level, we used counts but we do recognize the effect of face amount using them as a covariate characteristic in our parametric models. This is much more flexible as it allows us to directly quantify the effect of any increases in face amount on the survivorship of policyholders.

5.4 Model calibration results

5.4.1 Time-until-withdrawal

Prior to fitting the various duration models discussed in section (5.2.1), we performed preliminary investigation of the observed distributions of the time-until-withdrawal according to the various available classifications (e.g. Plan Type, Sex, etc.). While it becomes too cumbersome and possibly even overwhelming to show the results of this preliminary investigation for all possible classifications, at best we present this analysis by Plan Type. The time-until-withdrawal has been measured in years from

policy issue.

FIGURE 5.4.1: A frequency histogram of the time-until-withdrawal

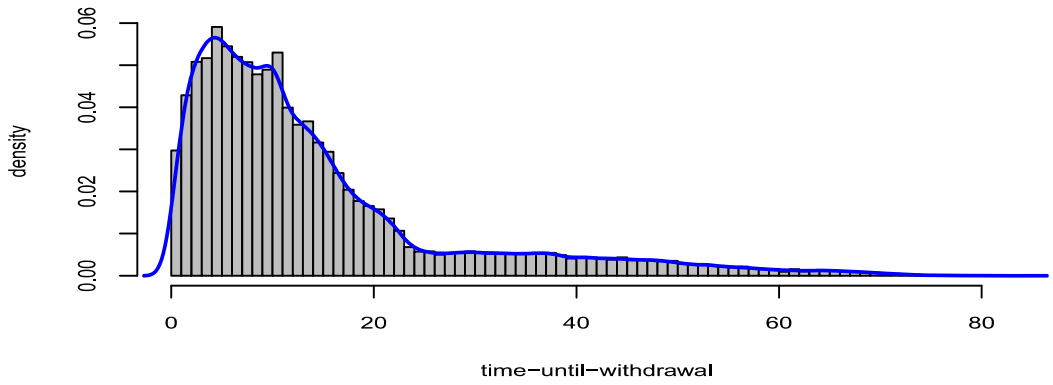


Figure 5.4.1 provides a graphical display of the frequency histogram of the observed time-until-withdrawal for all 65,435 policies. Broadly speaking, we find that policyholders do voluntarily terminate their contracts following the early duration from policy issue. However, upon inspection by policy type as shown in Figure 5.4.2, there is apparently a wide variation. First, a greater proportion of term insurance contracts tends to lapse during the early duration from policy issue; a possible explanation is the exponential increase in premium for such contracts. Second, more permanent contracts also follow the same pattern but at a much relatively lower rate than term contracts. There is a greater proportion, though, of such contracts to lapse in later years; a possible explanation is the cash value component usually associated with such contracts. Finally, for other types of contracts which primarily consist of Universal Life or similar products, there tends to be more a relatively flat stable proportion of policy lapses across duration; a possible explanation for this is the tendency of these products to be more viewed as savings or investment-type products with relatively less important insurance component.

FIGURE 5.4.2: A frequency histogram of the time-until-withdrawal by Plan Type

(A) PlanTypeP (B) PlanTypeT (C) PlanTypeO

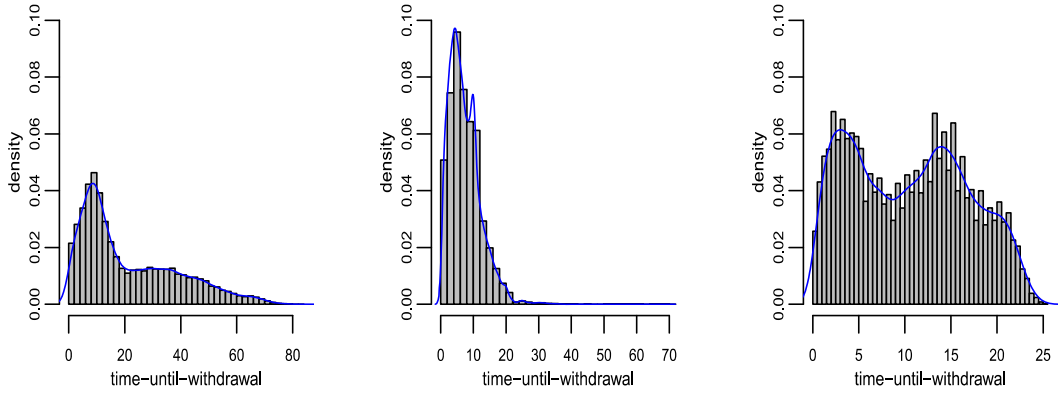


Table 5.4.1 provides basic summary statistics of the time-until-withdrawal according to Plan Type as well as on the aggregate. On the aggregate, the earliest policy termination happened to be about 0.01 of a year, or roughly one week from issue. On the other hand, the latest policy lapse happen after 83.75 years since policy issue. Observe the variation of the summary statistics by type of plan.

TABLE 5.4.1: Summary statistics of the time-until-withdrawal

Plan Type	Number	Min	Mean	Median	Max	Std Dev
PlanTypeP	27,776	0.08	21.46	14.80	83.75	17.24
PlanTypeT	18,333	0.01	7.34	6.42	70.15	4.83
PlanTypeO	19,326	0.08	10.51	10.62	25.01	6.36
Aggregate	65,435	0.01	14.27	10.01	83.75	13.57

In estimating the model parameters, we use maximum likelihood techniques with the log-likelihood function following the form of

$$\log \mathrm{L}(\beta, \theta; t_{w,i}) = \sum_{i=1}^{65{,}435} \log f_w(t_{w,i}),$$

where f_w refers to the density function applicable for the time-until-withdrawal random variable and $t_{w,i}$ refers to the time-until-withdrawal for the i-th observation. Here the vector β refers to the set of parameters corresponding to the coefficients in the regression equation for the location parameter μ while θ is the vector of the rest of the parameters applicable to the fitted model.

We fitted three types of parametric models as discussed in section (5.2.1): the Log-Normal, the Generalized Gamma and the GB2 distribution models. All three models provide enough flexibility so as to capture the observed long tailness as visually demonstrated in Figures 5.4.1 and 5.4.2. The estimation of the parameters has been coded using R and is a straightforward procedure. The calibration results for the time-until-withdrawal is numerically summarized in Table 5.4.2.

The interpretation of the regression coefficients is rather straightforward; many of these results also do not vary much by the choice of the distribution model. For example, upon inspection of the GB2 model, PlanTypeP and PlanTypeT policies tend to relatively have earlier policy terminations, males tend to lapse later, older issue ages tend to lapse earlier and those policies that are subjected to extra hazard with extra mortality cost tend to lapse earlier.

Figure 5.4.3 provides a graphical display of assessing the quality of the model fit of the various distribution models. For each of the three models considered, we display the histogram together with the parametric fit of the observable errors after taking into account policy characteristics that make the observations heterogeneous. To reinforce the quality of this fit, we provide additionally the corresponding probability-probability (P-P) plots of the observed residuals from each model considered. We find that both Generalized Gamma and GB2, according to these figures, provide reasonably excellent fit; however, the GB2 appears to be a marginally better fit and

TABLE 5.4.2: Maximum likelihood estimates for the various duration models of time-until-withdrawal

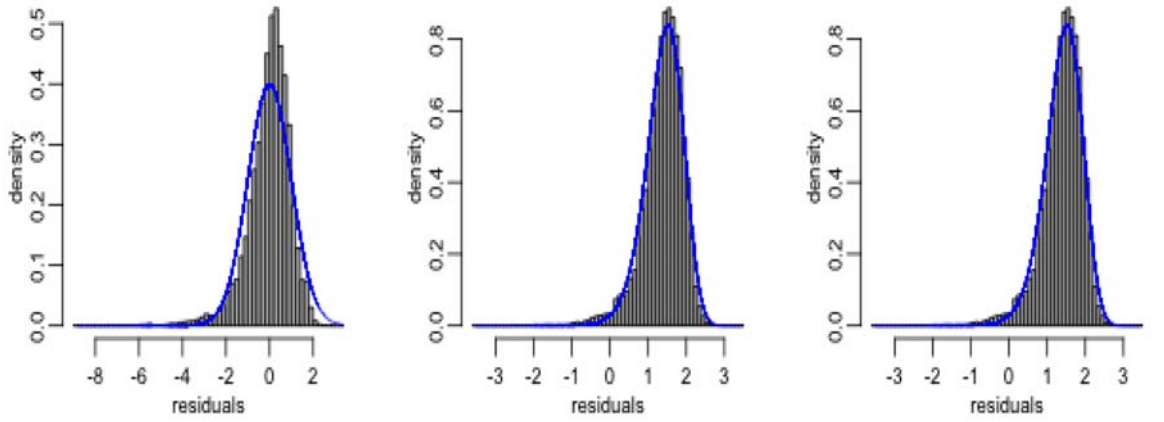
Parameter	Log-Normal		Generalized Gamma	GB2
Regression coefficients				
β_0 (intercept)	2.5534 (0.0263)		1.2138 (0.0419)	3.0034 (0.0238)
β_1 (PlanTypeP)	-0.4022 (0.0071)		-0.1604 (0.0061)	-0.1956 (0.0054)
β_2 (PlanTypeT)	-0.2808 (0.0068)		-0.1422 (0.0060)	-0.2805 (0.0055)
β_5 (RiskClassY)	-0.9787 (0.0063)		-0.6593 (0.0056)	-0.8199 (0.0060)
β_6 (Male)	0.0582 (0.0053)		0.0297 (0.0047)	0.0326 (0.0041)
β_7 (SmokerN)	0.2388 (0.0079)		0.3641 (0.0065)	0.1258 (0.0063)
β_8 (SmokerC)	1.6988 (0.0099)		1.7042 (0.0086)	1.2458 (0.0079)
β_{10} (Face Amount)	-0.0003 (0.0004)	*	-0.0027 (0.0003)	-0.0089 (0.0004)
β_{11} (Temp FEAmt)	0.0157 (0.0026)		0.0287 (0.0027)	-0.0258 (0.0020)
β_{12} (Perm FEAmt)	-0.0104 (0.0028)		-0.0167 (0.0023)	-0.0306 (0.0024)
β_{13} (MEFact)	-0.1168 (0.0240)		-0.6373 (0.0162)	-0.1553 (0.0216)
β_{14} (IssAge)	-0.0060 (0.0002)		-0.0092 (0.0002)	-0.0030 (0.0002)
Model specific parameters				
σ	0.6464 (0.0018)		1.2089 (0.0130)	0.2190 (0.0065)
m	-		4.5774 (0.0966)	-
γ_1	-		-	0.4303 (0.0168)
γ_2	-		-	1.2020 (0.0486)
Model fit statistics				
Number of observations	65,435		65,435	65,435
Log-likelihood	-209,054.1		-206,010.2	-201,199.5
Number of parameters	13		14	15
Akaike information criterion	418,134.19		412,048.47	402,428.96

Notes:
a. Face amount is re-scaled in 100,000.
b. Standard errors are in parenthesis.
c. An asterisk * identifies 'not significant' at the 5% level.

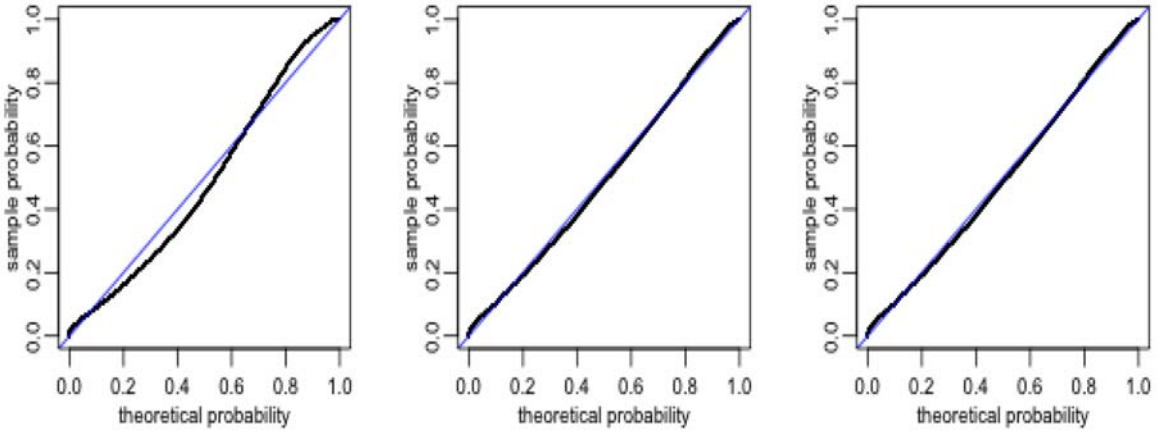
this is further bolstered by the slightly lower AIC criterion measure displayed in Table 5.4.2.

FIGURE 5.4.3: Comparing the quality of fit of the various duration models of time-until-withdrawal

(A) Log-Normal (B) Generalized Gamma (C) GB2



(D) Log-Normal (E) Generalized Gamma (F) GB2



5.4.2 Age at death

Unlike the time-until-withdrawal, preliminary analysis of the duration from withdrawal till death is much more difficult to perform because of the presence of the censoring of the observations. Table 5.4.3 provides the frequency distribution of the mortality status of the policies in the portfolio according to issue age and gender. As shown on this table, there is a strong presence of censoring on the observations. For example, out of the total 42,676 males in the data, we observe only 2,480 actual deaths as at the end of the observation period; this represents only about 5.8% of the all males in the data. Similarly, out of the 22,759 females observed, we have 1,066 deaths as at the end of the observation period, and this represents less than 4.7% of all females in the data. Furthermore, on the aggregate, we therefore observe only 3,546 deaths out of the total 65,435 observations in the data. This is just about 5.4% observed deaths in the data.

TABLE 5.4.3: Number of policies by issue age, sex and mortality status

	Mortality status		
Issue Age	**Survive**	**Death**	**Total**
Males			
$\leq$ 30	8,995	672	9,667
30-50	24,341	1,006	25,347
50-70	6,621	733	7,354
$>$ 70	239	69	308
Total	40,196	2,480	42,676
Females			
$\leq$ 30	6,532	349	6,881
30-50	12,202	394	12,596
50-70	2,653	258	2,911
$>$ 70	306	65	371
Total	21,693	1,066	22,759

Maximum likelihood techniques were used to estimate the parameters in the distribution models for the age at death. As discussed in subsection(5.2.2), while we investigated several other parametric models, our analysis resulted in a conclusive decision between the Gompertz and Weibull survival models, both distribution models of which are familiar distributions to actuaries. Our observable data, $(z_i, t_{w,i}, t_{wd,i}, \delta_i)$, consists of the age at issue, the time of withdrawal, the time of death from withdrawal (if applicable), and a censoring variable. The censoring variable δ_i has a value of 1 if censored, that is, the policyholder survived to reach the end of the observation period. Otherwise, it has a value of 0 if the policyholder died during the observation period. Based on this observable data, we constructed the log-likelihood using the result in equation (5.2.11) as follows:

$$\log \mathrm{L}(m^*, \sigma^*; z_i, t_{w,i}, t_{wd,i}, \delta_i) =$$

$$\sum_{i=1}^{65,435} \left[(1-\delta_i) \log \frac{f_d(z_i + t_{w,i} + t_{wd,i})}{S_d(z_i + t_{w,i})} + \delta_i \log \frac{S_d(z_i + t_{w,i} + t_{wd,i})}{S_d(z_i + t_{w,i})} \right], \quad (5.4.1)$$

where f_d is the corresponding density function for the age at death random variable.

We attempted to fit covariate information to account for policyholder heterogeneity similar to that of the time to withdrawal. However, we found that several of these heterogeneous characteristics did not significantly affect the pattern of mortality once the policy lapsed. In addition, we find that there was no differential between male and female for the location parameter m^*, but such was not the case for the variability parameter σ^*. The estimate for m^* is about 94 years old, both for Gompertz

and Weibull models and was not affected by gender. At first glance, we thought that the location estimates appear to be quite high may be a little bit of a surprising result. However, this has been largely a result of the censoring of our observations. As initially indicated, we have only approximately about 5% of our observations that were not censored; the rest were censored. For the uncensored observations, that is where deaths were observed, the median age at death is 81 years old and that the 75th percentile is 88 years old. The maximum age at death observed is slightly above 106 years old. For the censored observations, on the other hand, we found that the median age at the time of censoring is approximately 57 years old and the 75th percentile is 65 years old, with a maximum of 108 years old. All these high ages both for censored and uncensored observations contributed to the high location estimates.

TABLE 5.4.4: Maximum likelihood estimates for the various survivorship models

Parameter	Gompertz	Weibull
m^*	93.6031(0.1428)	94.2095 (0.1811)
σ^*	6.8420 (0.0975)	8.3039 (0.1337)
$\sigma^*\times$ Male	0.5206 (0.1161)	0.7507 (0.1481)
Model fit statistics		
Number of observations	65,435	65,435
Log-likelihood	-18,264.55	-18,433.82
Number of parameters	3	3
Akaike information criterion	36,535.11	36,873.63

The quality of the fit between the Gompertz and the Weibull models can be visualized in Figure 5.4.4. These figures compare the nonparametric Kaplan-Meier type survival curves against corresponding parametric survivorship curves with parameters calibrated from the data. Kaplan-Meier survival curves do account for the censoring of the observations as in our data. The comparison here is not only between models

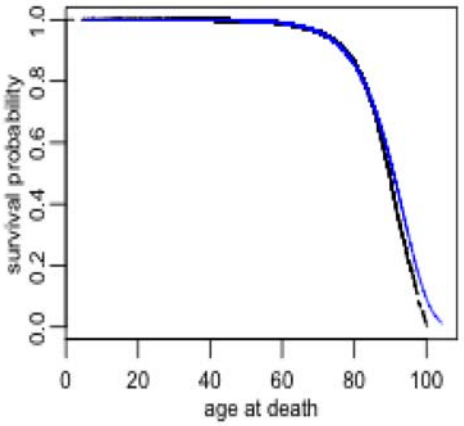
but also between males and females. Broadly speaking, we find that the Gompertz model, for either the male or female, slightly outperform the Weibull model. This is not at all surprising considering the nature of our observed data; it is well known that the Gompertz model explains a large part of the exponentially increasing pattern of mortality at very high ages. Our observed data is derived from an insured group which consisted, in large sense, of policyholders issued at very high ages and observed for a long time duration. While it is true that we have negligibly few policyholders with very young issue ages, it is quite uncommon to have an insurance coverage at early ages. However, our average issue age in the data has been about 38 years old with a 75th percentile of 46 years old. Indeed, surprisingly, we even have a maximum issue age of approximately 90 years old. Except possibly under special circumstances, insurers typically do not actively seek insurance sale within the very old age market. As can be deduced from Table 5.4.3, more than half of the policies have had issue ages around the range of 30 to 50 years old. Insurance is generally viewed as a financial product that provides economic security against early and premature death particularly for the head of the household; hence, it is not surprising to find the range of issue ages in our data.

5.5 Implications

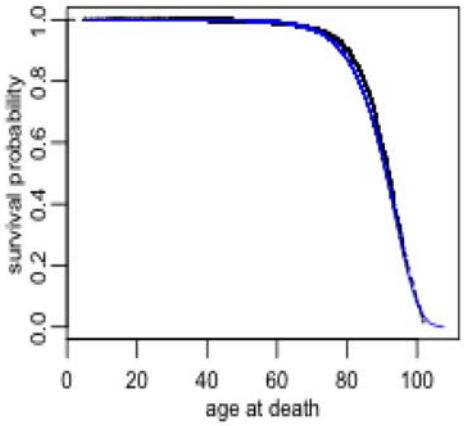
In order to understand the consequences of our calibrated models, we examined two material aspects that may be of importance to actuaries. The first one is a deduction of the presence of mortality antiselection. The second one is the financial cost of insurance policy terminations. In this section, we consider these two implications

FIGURE 5.4.4: Kaplan-Meier versus fitted survival curves.

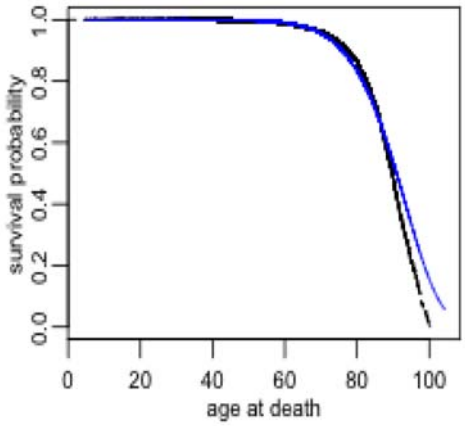
(A) Gompertz - Male



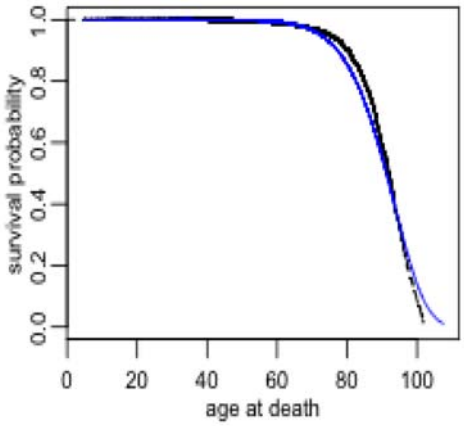
(B) Gompertz - Female



(C) Weibull - Male



(D) Weibull - Female



separately.

5.5.1 Evidence of the presence of mortality selection

In the life insurance industry, mortality antiselection refers to the adverse consequences of the imbalance in the portfolio that may result because policies that do terminate are those with better survival pattern. When there is a presence of antis-

election, the insurance company may end up with the spiraling effect of a worsening mortality pattern as a result of policy terminations. Insurance portfolios with worse mortality pattern may have consequences that can negatively impact both the company's balance sheet as well as income statements. Thus, we look for evidence of the presence of such selection.

In analyzing mortality selection, we define what we meant by antiselection. We follow the definition, which are equivalent, considered both in Carriere (1998) and Valdez (2001). We say that there is presence of antiselection at withdrawal in life insurance if

$$S_{d|w}(t_d|t_w) > S_d(t_d), \quad \text{for every } t_d \geq t_w. \tag{5.5.1}$$

To interpret definition (5.5.1), antiselection is evidently present when the survival pattern of those terminated policies, conditional on all periods of termination, have generally better unconditional survival pattern. To look for evidence in our data, we consider a specific type of a policyholder with the following characteristics: issue age 35, permanent whole life, a non-smoker, male, face amount of 250,000, and not-so-risky with no flat extra charges. We compare the conditional and unconditional survivorship curves for this policyholder if he terminates his contract for different years from issue; here we consider withdrawals for years 2, 4, 6, 8, 10, 15, 20 and 30. The result of this comparison is graphically displayed in Figure 5.5.1.

According to this figure, we clearly find some evidence of mortality antiselection since survivorship of terminated policies are always above those unconditional survivorship curves for all duration of policy termination being consider here. While we certainly can consider an infinitely many more years of policy termination than those being considered, it would be too overwhelming to the reader to see this evident. Suf-

fice to say that we investigated many more years of policy termination, and similar pattern has been observed.

An additional interesting observation that can be made from this figure is the increase in magnitude of the difference between the two curves as policy withdrawals occur in later years. This translates to an even stronger presence of antiselection when policy terminates in later years. This is not at all counterintuitive because policyholders who have had their insurance contracts for a longer period would have to seriously consider their greater risk of mortality as the probability of dying exponentially increases with age; this is even more evident in the parametric form of the Gompertz survival curve.

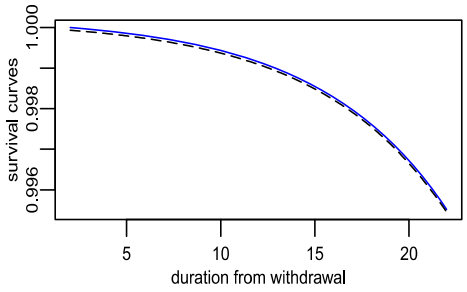
Finally, it is natural to ask whether the differences between the two survivorship curves we saw in Figure 5.5.1 are statistically significant. To further convinced ourselves, we conducted a formal hypothesis tests based on the Wilcoxon signed-rank sum test to compare the significance of these differences. This is a nonparametric hypothesis tests for examining evidence of significant differences between two related samples, applicably so in this situation. We found that for early withdrawals, in particular for years 2, 4 and 6, the differences are not statistically significant at level $\alpha = 5\%$. This could also be visualized from the graphs. Nevertheless, for policy terminations in later years, specifically for years 8 and beyond, we found the differences to be statistically significant at level $\alpha = 5\%$.

5.5.2 The financial impact of policy termination

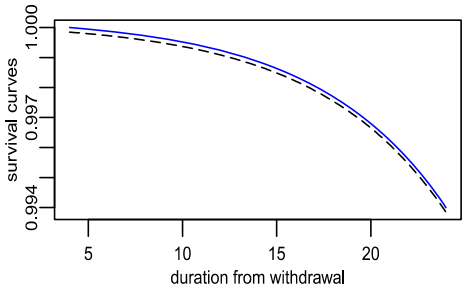
To illustrate how we can quantify the financial consequences of policy terminations, we illustrate this by considering a policyholder with the same characteristics that was

FIGURE 5.5.1: Comparing survival curves after policy withdrawal for an issue age 35. The smooth curve gives the survival probabilities conditional from policy termination. The dashed curve gives the unconditional survival probabilities. The curves are provided for 20 years from policy termination.

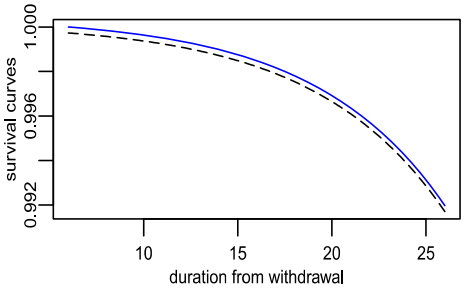
(A) Withdrawal 2 years



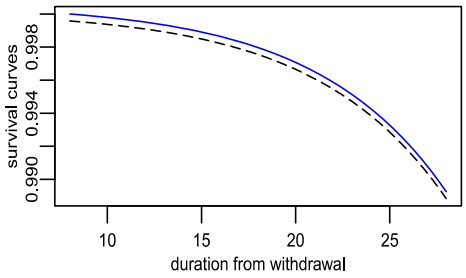
(B) Withdrawal 4 years



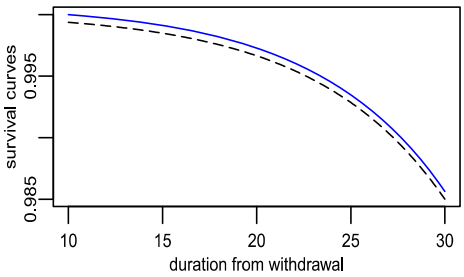
(C) Withdrawal 6 years



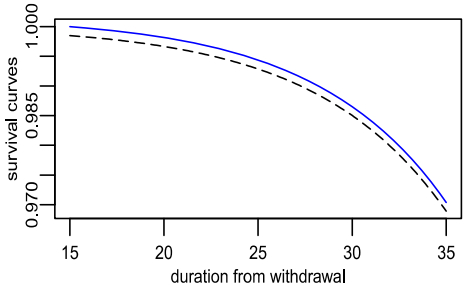
(D) Withdrawal 8 years



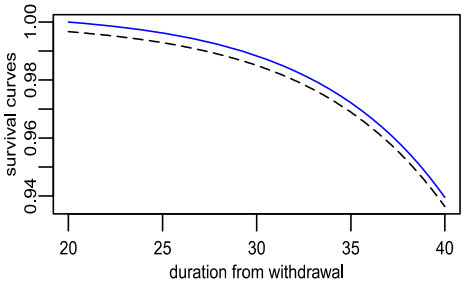
(E) Withdrawal 10 years



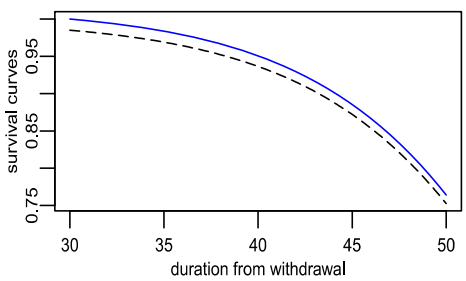
(F) Withdrawal 15 years



(G) Withdrawal 20 years



(H) Withdrawal 30 years



considered in the previous section: issue age 35, permanent whole life, a non-smoker, male, face amount of 250,000, and not-so-risky with no flat extra charges. Two types of expenses were assumed in the calculations:

- acquisition expenses: 80 plus 4.5 per 1,000 of death benefit; and
- maintenance expenses: 60 plus 3.5 per 1,000 of death benefit.

These assumptions have been somewhat drawn from the expense study done in Segal (2002) where he estimated both "the acquisition and maintenance costs associated with life policies". We refer the reader to this article for details if interested. For simplification, we assume that death benefit is paid at the moment of death while premiums, with expenses, occur at a continuous rate throughout each year. Finally, interest rate used for discounting has been set at the constant rate of 5% per year.

In order to investigate the financial impact, first we calculated the premium payable for this policy. We based this calculation on the actuarial equivalence principle, something typically learned in a mathematics for life contingencies course. In this case, the premium has been calculated at the rate of 2,010 per annum.

All stochastic components in the calculation process have been done using simulation. The time-until-withdrawal random variables were simulated based on the Generalized Gamma family of distributions. While we said earlier that the GB2 distribution models appear to be marginally better, for simulation purpose, the Generalized Gamma family provides more ease in simulation. The age at death random variables were simulated based on the Gompertz model. To demonstrate for example how to simulate from our Gompertz model as specified in equation (5.2.12), we can use the inverse transform method. Here, we start with a random number, say U, and

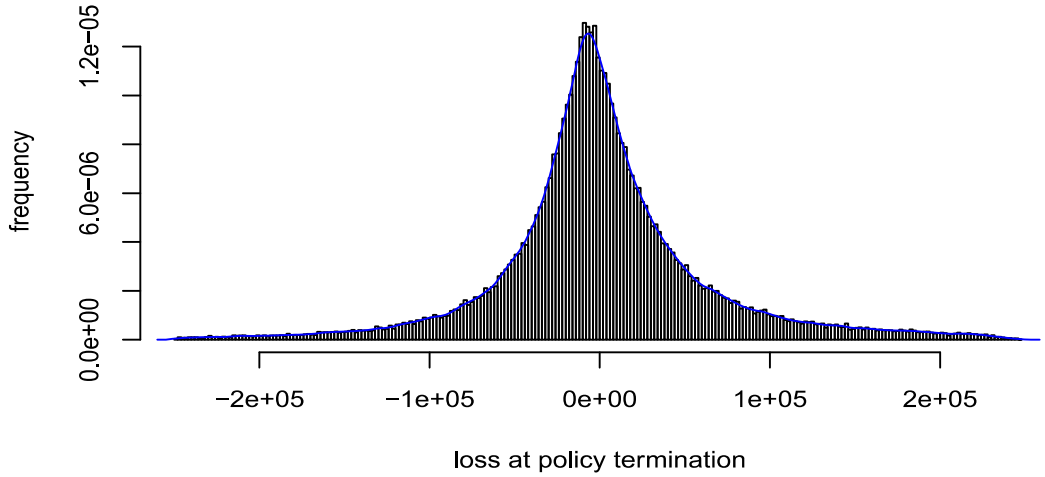
generate a Gompertz lifetime, say T, from the following equation:

$$T = \sigma^* \log\left[1 - (\log(U))e^{m^*/\sigma^*}\right] \tag{5.5.2}$$

The financial impact is the loss incurred when policy terminates. These loss calculations have been done based on retrospective principles. In effect, we define the loss at policy termination to be the accumulated values of all past expenses incurred, plus policy reserves, reduced by the accumulated value of all past premiums paid. Once simulations of the random components are done, this process of loss calculation is rather straightforward. We coded the calculations using the R package. This loss is best summarized, first, with a frequency distribution as depicted in Figure 5.5.2.

According to these simulation results, the largest negative loss is -249,500 and the largest positive loss is 248,000. The mean and median losses are, respectively, 1,223 and -3,128. The 25th percentile -s -26,440 while the 75th percentile is 25,610. There is about 54% chance that the loss will be negative and about 46% that the loss is positive. Finally, there is a very slim chance that the loss will be larger than 200,000 but there is a 3% chance it will be above 150,000.

FIGURE 5.5.2: The frequency histogram of the loss at policy termination



Chapter 6

Concluding Remarks

This chapter concludes the thesis by listing the key findings of the work we completed in this dissertation. Suggestions for future research and possible industrial and practical uses are also presented.

Chapter 3 investigated the effect of demographic, health, lifestyle, and financial risk factors on mortality. We find that these factors provide additional helpful information in predicting survivorship so that we further evaluate the financial impact of these factors related to pension annuities. This investigation used a Cox regression model with extended covariates, many of which were time-varying. With the given set of variables, in addition to age and gender, parent's average age, census division, marital status, health indices , self-reported health, mental health Index, total household income, physical activity or exercise, alcohol drinking status, smoking status, and body mass index are the variables selected in the final model. As expected, increasing age increases the hazard for mortality and males have higher hazard rates than females. Some other key findings with the extended covariates are:

- Increasing parent's average age decreases the hazard for mortality significantly.
- Mid Atlantic, ES Central, WS Central, and the Pacific census divisions have significantly lower hazards than EN Central. Please refer to the appendix for details of these census divisions.
- Single persons have higher mortality risk than persons who are married or have partners.
- The greater the number of problematic health conditions lead to a greater mortality risk.
- The better you feel about your health lowers your mortality risk.
- People with higher income generally have lower mortality risk.
- Not surprisingly, current smokers have higher mortality risk than non-smokers or former smokers.

In the appendix, we also summarized how many of these key extended variables we find significant in our study compare to the results of many other previous studies.

As an application of the estimated model in projecting pension liabilities particularly when it relates to annuity value calculations, we find that annuity values can be more precisely calculated with the addition of information related to the significant factors we find in our model. In general, we find that the more information we know about an individual, the more we are able to give precision in our mortality prediction and this leads generally to a higher annuity values. These values were compared to calculations that simply used the RP-2000 mortality table with Scale BB improvement rates. The Society of Actuaries has recently published new mortality tables

Retirement Plans 2014 (RP-2014) and mortality improvement scale called Mortality Projection (MP-2014) which can be used to measure the retirement liabilities in the United States. For more details, see *RP-2014 Mortality Tables* (2014) and *Mortality Improvement Scale MP-2014* (2014). It is definitely worth exploring how these new tables now compare to the mortality projections and annuity value calculations based on our Cox regression model.

It is important to note that the use of extended variables, other than the conventional age and gender, do indeed contribute to a more accurate prediction of pension liabilities. However, for certain pension plans, such extended variables may not be readily available or may be difficult to obtain, especially those related to possibly health and lifestyle. The use of proxy variables is highly suggested in such instances. To illustrate, some key variables may be derived from published industry sources such as those related to census divisions and changing marital or partnership status. As yet another example, most jobs are generally known to be classified according to Blue, White or Mixed collar categories. These collar categories are a useful information about the level of income of the worker. See appendix for this occupation class categorization. The use of proxy variables is commonly used in data analysis. It will be interesting to note how well these proxy variables perform as substitutes in predicting mortality.

For pension plans who may not wish to use complicated models like what we proposed because extended variables may be unavailable, it is worth continuing to compare results of our model to the *Mortality Improvement Scale BB* (2012) work by the Society of Actuaries. This work provides a two-dimensional improvement scale designed to reflect recent mortality improvement experiences. Scale BB reflects age as well as the cohort effect using the year-of birth. In our primary modeling work, we

have included the cohort effect as a covariate but later found that this cohort effect was not statistically significant. To some degree, this may be the result of our key extended variables are contributing to the cohort effect, that is, for example, income, smoking habits, and physical activities may have had contributed to the cohort effect. Henceforth, we suggest to continue to explore the use of Scale BB improvement rates as a valuable information to use in the absence of the extended variables.

For companies estimating pension liabilities, it appears to be most appropriate to incorporate one's own mortality experience as a baseline in our Cox regression model with extended variables. This baseline can be further improved by blending the company's own experience with that of a comparable industry using a credibility-weighted baseline mortality approach that is commonly used in practice. As described in the *Credibility Practice Note* (2008), limited fluctuation credibility or greatest accuracy credibility are two simple possible approaches that can be used to develop the suitable credibility factors, and these blended mortality hazards can then be used to define the baseline function in the Cox regression model. Furthermore, because the industry usually has a huge amount of available information for data analysis, it is a common practice to classify groups such as age intervals provided they share some homogeneous characteristics; in this case, one possible future work is to re-evaluate the estimation of the Cox regression model using grouped data.

Chapter 4 is a complex illustration of the use of multiple state models for long term care. Multiple state models are becoming an important tool for actuaries who are interested in pricing and reserving many life-contingent contracts. In our long term care study, the states are defined depending on the number of ADL difficulties together with death as the absorbing state. We find that based on the likelihood ratio test, Model 2 with extended covariates performed statistically better than Model 1

which is based only on age and sex. The resulting estimated hazard ratios illustrate that:

- Worse health indices show higher hazard rates to transition between states.
- Increasing the number of ADL difficulties can be seen as a result of increasing BMI while lowering the risk for mortality.
- The less educated people generally tend to be at risk of increasing the number of ADL difficulties and lowering ADL recovery rates.
- Weak mental health increases the number of ADL difficulties.
- As average parents age increases, lower hazard can be observed.

Other test statistics we performed indicate the lack of adequacy of our Model 2. We partly attribute this to the interpretation of the data we used itself. For one, we were unable to capture precise long term care information in our data set, and hence, we used the number of ADL difficulties as a proxy to the level of long term care required by the individuals in our dataset. It will be interesting to examine the use of our multiple state model based on a dataset that captures precise long term care information. We suggest this for future work.

For additional possible future work, we suggest improvements in the model by possibly accounting for recovery from state to state. This will additionally complicate the multiple state framework particularly in the estimation where you now have to account for the likelihood of those individuals that are transitioning within these states.

Projecting the cost of long term care per se is an integral part of any actuarial analysis. However, we find that within a multiple state model, the calculations of

expected present value of future long term care benefits are no longer straightforward. A complex development of a program code is suitable but requires some serious effort. Therefore, we defer this work for future research.

In Chapter 5, we conducted an empirical investigation of the mortality pattern of terminated policies. We drew a random sample from a follow-up study conducted by a major insurance company which tracked the death pattern of their portfolio of insurance policies. We examined and modeled these life insurance policy termination together with their survivorship pattern. We used parametric distribution duration models to calibrate the data observed on time-until-withdrawal and found that the log-location scale class of distributions fit our data very well. This class of distributions is very flexible and is able to accommodate covariate information, through the location parameter, in order to account for the apparent heterogeneity in our data.

The more interesting aspect of our work is studying the survivorship pattern of terminated contracts. The censored nature of mortality data typically presents a challenge when calibrating the age at death data. While we investigated several classes of parametric mortality distribution models, we narrowed our choice between the Gompertz and the Weibull survival models. We found that while both provide quality fit to the data, the Gompertz model appeared to marginally outperform the Weibull model. Just as with the class of duration models we investigated for the time-until-withdrawal, we injected covariate information directly through the location and scale parameters; however, we found very little statistical evidence of heterogeneity in the mortality pattern. Indeed, even surprisingly, gender did not statistically affect the average age at death, but it did so on the variation of this age at death.

Furthermore, we examined the actuarial implications of our model calibration. We found that the data provides support for evidence of mortality antiselection. This

indicates that policies that do terminate generally tend to have better survivorship pattern than those who do not. However, we also discovered that the difference in the survivorship pattern is affected by when the policy terminates. Generally speaking, there is little or no statistical significance for early policy terminations, in particular, for terminations before 6 years from policy issue. Beyond 6 years, we found stronger evidence of antiselection and this antiselection increases with later terminations.

Finally, we quantified the financial effect of policy terminations by examining the loss that would have been incurred when the policy terminates. The loss usually consists of unrecoverable expenses incurred at policy issue (it is well known that at policy issue, acquisition expenses are too large relative to the premium collected), net premium reserve that is usually released when the policy terminates, and the loss of future uncollected premiums. To perform the investigation, we examined a particular policy with specified characteristics as a case study. We found that there is about a 50-50 chance of a negative and positive loss when policy terminates.

The results of our study in Chapter 5 has the limitation that we lack the information associated with those policies that were in-force. This precluded us from developing a more precise pattern of mortality associated with in-force. This is more a result of the lack of data rather than the model development and estimation itself. Furthermore, if data is available, a similar analysis of the effect of termination on mortality can be done for pension plans as well as long term care insurance products. This would indeed provide for an additional interesting actuarial analysis.

Bibliography

Aalen, O. 1978. "Nonparametric Inference for a Family of Counting Processes." *The Annals of Statistics* 6 (4): 701-726.

Aguirre-Hernández, R., and V.T. Farewell. 2002. "A Pearson-type Goodness-of-fit Test for Stationary and Time-continuous Markov Regression Models." *Statistics in Medicine* 21 (13): 1899–1911.

Allison, D.B., K.R. Fontaine, J.E. Manson, J. Stevens, and T.B. VanItallie. 1999. "Annual Deaths Attributable to Obesity in the United States." *The Journal of the American Medical Association* 282 (16): 1530–1538.

Allison, P.D. 2012. *Survival Analysis Using SAS: A Practical Guide.* SAS Institute.

Andersen, P.K., and N. Keiding. 2002. "Multi-state Models for Event History Analysis." *Statistical Methods in Medical Research* 11 (2): 91–115.

Andersen, P.K., and R.D. Gill. 1982. "Cox's Regression Model for Counting Processes: A Large Sample Study." *The Annals of Statistics* 10: 1100-1120.

Bandeen-Roche, K., Q.L Xue, L. Ferrucci, J. Walston, J.M. Guralnik, P. Chaves, S.L. Zeger, and L.P. Fried. 2006. "Phenotype of Frailty: Characterization in the Women's Health and Aging Studies." *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 61 (3): 262–266.

Banks, J., M. Marmot, Z. Oldfield, and J.P. Smith. 2006. "Disease and Disadvantage in the United States and in England." *The Journal of the American Medical Association* 295 (17): 2037-2045.

Barsky, R.B., F.T. Juster, M.S. Kimball, and M.D. Shapiro. 1997. "Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Study." *The Quarterly Journal of Economics* 112 (2): 537-579.

Bassuk, S.S., L.F. Berkman, and B.C. Amick. 2002. "Socioeconomic Status and Mortality among the Elderly: Findings from Four US Communities." *American Journal of Epidemiology* 155 (6): 520-533.

Bentler, S.E., L. Liu, M. Obrizan, E.A. Cook, K.B. Wright, J.F. Geweke, E.A. Chrischilles, C.E. Pavlik, R.B. Wallace, R.L. Ohsfeldt et al. 2009. "The Aftermath of Hip Fracture: Discharge Placement, Functional Status Change, and Mortality." *American Journal of Epidemiology* 170 (10): 1290–1299.

Beyersmann, J., A. Allignol, and M. Schumacher. 2012. *Competing Risks and Multistate Models with R.* Springer.

Black, B.S., P.V. Rabins, and P.S. German. 1999. "Predictors of Nursing Home Placement among Elderly Public Housing Residents." *The Gerontologist* 39 (5): 559–568.

Blau, D.M., and D.B. Gilleskie. 2006. "Health Insurance and Retirement of Married Couples." *Journal of Applied Econometrics* 21 (7): 935-953.

Blustein, J., S. Chan, and F.C. Guanais. 2004. "Elevated Depressive Symptoms among Caregiving Grandparents." *Health Services Research* 39 (6p1): 1671-1690.

Booth, H., and L. Tickle. 2008. "Mortality Modeling and Forecasting: A Review of Methods." *Annals of Actuarial Science* 3 (1-2): 3-43.

Bowers, N.L., H.U. Gerber, J.C. Hickman, D.A. Jones, and C.J. Nesbitt. 1986. *Actuarial Mathematics.* Schaumburg, Illinois: Society of Actuaries.

Branch, L.G., and A.M. Jette. 1982. "A Prospective Study of Long-term Care Institutionalization among the Aged." *American Journal of Public Health* 72 (12): 1373–1379.

Brown, R.L., and J. McDaid. 2003. "Factors Affecting Retirement Mortality." *North American Actuarial Journal* 7 (2): 24-43.

Byrne, D., M. S. Goeree, B. Hiedemann, and S. Stern. 2009. "Formal Home Health Care, Informal Care, and Family Decision Making." *International Economic Review* 50 (4): 1205-1242.

Carriere, J.F. 1992. "Parametric Models for Life Tables." *Transactions of the Society of Actuaries* 44: 77-99.

Carriere, J.F. 1998. "Withdrawal Benefits under a Dependent Double Decrement Model." *ASTIN Bulletin* 28: 49-57.

Chemerinski, E., R.G. Robinson, and J.T. Kosier. 2001. "Improved Recovery in Activities of Daily Living Associated with Remission of Post Stroke Depression." *Stroke* 32 (1): 113–117.

Cigolle, C.T., K.M. Langa, M.U. Kabeto, Z. Tian, and C.S. Blaum. 2007. "Geriatric Conditions and Disability: The Health and Retirement Study." *Annals of Internal Medicine* 147 (3): 156-164.

Cox, D.D.R., and H.D. Miller. 1977. *The Theory of Stochastic Processes.* Vol. 134 CRC Press.

Cox, D.R. 1972. "Regression Models and Life-tables." *Journal of the Royal Statistical Society. Series B (Methodological)* 34: 187-220.

Cox, D.R. 1975. "Partial Likelihood." *Biometrika* 62: 269-276.

Cramér, H., and H. Wold. 1935. "Mortality Variations in Sweden: A Study in Graduation and Forecasting." *Scandinavian Actuarial Journal* 1935 (3-4): 161-241.

Credibility Practice Note. 2008.

Crimmins, E.M. 2011. *International Handbook of Adult Mortality.* Vol. 2 Springer.

Crimmins, E.M., M.D. Hayward, A. Hagedorn, Y. Saito, and N. Brouard. 2009. "Change in Disability-free Life Expectancy for Americans 70 Years Old and Older." *Demography* 46 (3): 627–646.

Cummins, J.D., G. Dionne, J.B. McDonald, and M.B. Pritchett. 1990. "Applications of the GB2 Family of Distributions in Modeling Insurance Loss Processes." *Insurance: Mathematics and Economics* 9: 257-272.

Day, J.C., and E.C. Newburger. 2002. *The Big Payoff: Educational Attainment and Synthetic Estimates of Work-life Earnings.* US Department of Commerce, Economics and Statistics Administration, US Census Bureau.

de Leon, C.F.M., T.A. Glass, L.A. Beckett, T.E. Seeman, D.A. Evans, and L.F. Berkman. 1999. "Social Networks and Disability Transitions Across Eight Intervals of Yearly Data in the New Haven EPESE." *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 54 (3): S162–S172.

de Moivre, A. 1725. *Annuities Upon Lives.* printed by WP and sold by Francis Fayram; and Benj. Motte; and W. Pearson.

DESA, UN. 2013. "World Population Prospects, The 2012 Revision." *New York: Department for Economic and Social Affairs.*

DeSalvo, K.B., N. Bloser, K. Reynolds, J. He, and P. Muntner. 2006. "Mortality Prediction with a Single General Self-Rated Health Question." *Journal of General Internal Medicine* 21 (3): 267-275.

Dickson, D.C.M., M.R. Hardy, and H.R. Waters. 2009. *Actuarial Mathematics for Life Contingent Risks.* Cambridge University Press.

Dunlop, D.D., J.S. Lyons, L.M. Manheim, and R.W. Song, J.and Chang. 2004. "Arthritis and Heart Disease as Risk Factors for Major Depression: The Role of Functional Limitation." *Medical Care* 42 (6): 502-511.

Elandt-Johnson, R.C., and N.L. Johnson. 1980. *Survival Models and Data Analysis.* New York: John Wiley and Sons.

Fitzpatrick, D. 2014. "Rising U.S. Life Spans Spell Likely Pain for Pension Funds." *The Wall Street Journal.*

Fleming, T.R., and D.P. Harrington. 2005. *Counting Processes and Survival Analysis.* Wiley. com.

Frees, Edward W. 2004. *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences.* Cambridge University Press.

Frees, E.W., J.F. Carriere, and E.A. Valdez. 1996. "Annuity valuation with Dependent Mortality." *Journal of Risk and Insurance* 63: 229-261.

Friedberg, L., and A. Webb. 2005. "Retirement and the Evolution of Pension Structure." *Journal of Human Resources* 40 (2): 281-308.

Gallo, W.T., H.M. Teng, T.A. Falba, S.V. Kasl, H.M. Krumholz, and E.H. Bradley. 2006. "The Impact of Late Career Job Loss on Myocardial Infarction and Stroke: a 10 year Follow Up sing the Health and Retirement Survey." *Occupational and Environmental Medicine* 63 (10): 683-687.

Gill, R.D., and S. Johansen. 1990. "A survey of Product-integration with a View toward Application in Survival Analysis." *The Annals of Statistics* 18 (4): 1501-1555.

Gompertz, Benjamin. 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.

Gourieroux, C., and J. Jasiak. 2007. *The Econometrics of Individual Risk: Credit, Insurance, and Marketing.* Princeton, New Jersey: Princeton University Press.

Haberman, S. 1996. "Landmarks in the History of Actuarial Science (up to 1919)." *Dept. of Actuarial Science and Statistics, City University, London. Actuarial Research Paper* 84.

Håkansson, K., S. Rovio, E.L Helkala, A.R. Vilska, B. Winblad, H. Soininen, A. Nissinen, A.H. Mohammed, and M. Kivipelto. 2009. "Association between Midlife Marital Status and Cognitive Function in Later Life: Population based Cohort Study." *British Medical Journal* 339.

Heeringa, S.G., and J.H. Connor. 1995. "Technical Description of the Health and Retirement Survey Sample Design." *Ann Arbor: University of Michigan.*

Hjelmborg, J.vB., I. Lachine, A. Skytthe, J.W. Vaupel, M. McGue, M. Koskenvuo, J. Kaprio, N.L. Pedersen, and K. Christensen. 2006. "Genetic Influence on Human Lifespan and Longevity." *Human Genetics* 119 (3): 312-321.

Hougaard, P., and P. Hougaard. 2000. *Analysis of Multivariate Survival Data.* Vol. 564 Springer New York.

Hougaard, Philip. 1999. "Multi-state Models: A Review." *Lifetime Data Analysis* 5 (3): 239–264.

Idler, E.L., and Y. Benyamini. 1997. "Self-rated Health and Mortality: A Review of Twenty-seven Community Studies." *Journal of Health and Social Behavior* 38: 21-37.

Jackson, C.H. 2011. "Multi-state Models for Panel Data: the msm Package for R." *Journal of Statistical Software* 38 (8): 1-29.

Johnson, N.L., S. Kotz, and N. Balakrishnan. 1995. *Continuous Univariate Distributions.* New York: John Wiley and Sons.

Juster, F. T., and R. Suzman. 1995. "An Overview of the Health and Retirement Study." *Journal of Human Resources* 30: S7-S56.

Jylhä, M. 2009. "What is Self-rated Health and Why Does it Predict Mortality? Towards a Unified Conceptual Model." *Social Science & Medicine* 69 (3): 307-316.

KAISER. 2011. "Medicaid and Long-Term Care Services and Supports.". Available at `http://www.kff.org/medicaid/upload/2186-08.pdf`.

KAISER. 2013. "Five Key Facts About the Delivery and Financing of Long Term Services and Supports.".

Kalantar-Zadeh, K., G. Block, M.H. Humphreys, and J.D. Kopple. 2003. "Reverse Epidemiology of Cardiovascular Risk Factors in Maintenance Dialysis Patients." *Kidney International* 63 (3): 793–808.

Kalbfleisch, J.D., and J.F. Lawless. 1985. "The Analysis of Panel Data Under a Markov Assumption." *Journal of the American Statistical Association* 80 (392): 863-871.

Kaplan, E.L., and P. Meier. 1958. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association* 53 (282): 457-481.

Kay, R. 1986. "A Markov Model for Analyzing Cancer Markers and Disease States in Survival Studies." *Biometrics* 42 (4): 855-865.

Kemper, P., H.L. Komisar, and L. Alecxih. 2005. "Long-Term Care Over an Uncertain Future: What Can Current Retirees Expect?" *INQUIRY: The Journal of Health Care Organization, Provision, and Financing* 42 (4): 335-350.

Kramarow, E., J. Lubitz, H. Lentzner, and Y. Gorina. 2007. "Trends in the Health of Older Americans, 1970–2005." *Health Affairs* 26 (5): 1417–1425.

Kwon, H.S., and B.L. Jones. 2006. "The Impact of the Determinants of Mortality on Life Insurance and Annuities." *Insurance: Mathematics and Economics* 38 (2): 271-288.

Lancaster, T. 1990. *The Econometric Analysis of Transition Data*. Cambridge, United Kingdom: Cambridge University Press.

McDonald, J.B. 1984. "Some Generalized Functions for the Size Distribution of Income." *Econometrica* 52: 647-663.

McGarry, K., and R.F. Schoeni. 1995. "Transfer Behavior in the Health and Retirement Study: Measurement and the Redistribution of Resources within the Family." *Journal of Human Resources* 30: S184-S226.

Mitchell, O.S., J. Piggott, M. Sherris, and S. Yow. 2006. "Financial Innovation for an Aging World.".

Moore, K. 2011. "An Overview of the US Retirement Income Security System and the Principles and Values It Reflects." *Comparative Labor Law & Policy Journal* 33 (1).

Mortality Improvement Scale BB. 2012. Society of Actuaries.

Mortality Improvement Scale MP-2014. 2014. Society of Actuaries.

Note, Pension Committee Practice. 2011. "Selecting and Documenting Mortality Assumptions for Pensions." Revised October 2011.

Ortman, J.M., V.A. Velkoff, and H. Hogan. 2014. "An Aging Nation: The Older Population in the United States." *U.S. Census Bereau*.

Pappas, G., S. Queen, W. Hadden, and G. Fisher. 1993. "The Increasing Disparity in Mortality between Socioeconomic Groups in the United States, 1960 and 1986." *New England Journal of Medicine* 329 (2): 103–109.

Pitacco, E. 2004. "Survival Models in a Dynamic Context: a Survey." *Insurance: Mathematics and Economics* 35 (2): 279-298.

Quadagno, Jill, and JoEllen Pederson. 2012. "Has Support for Social Security Declined? Attitudes toward the Public Pension Scheme in the USA, 2000 and 2010." *International Journal of Social Welfare* 21 (s1): S88-S100.

Radloff, L.S. 1977. "The CES-D Scale a Self-report Depression Scale for Research in the General Population." *Applied Psychological Measurement* 1 (3): 385–401.

Ridsdale, B., and A. Gallop. 2010. "Mortality by Cause of Death and by Socioeconomic and Demographic Segmentation.". International Congress of Actuaries.

Rogers, R.G., R.A. Hummer, and C.B. Nam. 1999. *Living and Dying in the USA: Behavioral, Health, and Social Differentials of Adult Mortality*. Elsevier.

Rogowski, J., and L. Karoly. 2000. "Health Insurance and Retirement Behavior: Evidence from the Health and Retirement Survey." *Journal of Health Economics* 19 (4): 529-539.

Rosner, B., C. Raham, F. Orduña, M. Chan, L. Xue, Z. Benjazia, and G. Yang. 2013. "Literature Review and Assessment of Mortality Improvement Rates in the U.S. Population: Past Experience and Future Long-Term Trends.". SOA & Ernst & Young LLP.

RP-2014 Mortality Tables. 2014. Society of Actuaries.

Schachter, F., L. Faure-Delanef, F. Guénot, H. Rouger, P. Froguel, L. Lesueur-Ginot, and D. Cohen. 1994. "Genetic Associations with Human Longevity at the APOE and ACE loci." *Nature Genetics* 6 (1): 29–32.

Scientific Productivity of HRS. 2013. `http://hrsonline.isr.umich.edu/modules/biblio/CumulativeWorkformV2.htm`.

Segal, D. 2002. "An Economic Analysis of Life Insurance Company Expenses." *North American Actuarial Journal* 6: 81-94.

Singer, Judith D., and John B. Willett. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* Oxford University Press.

Sorlie, P.D., E. Backlund, and J.B. Keller. 1995. "US Mortality by Economic, Demographic, and Social Characteristics: the National Longitudinal Mortality Study." *American Journal of Public Health* 85 (7): 949–956.

Spillman, B. 2012. "Financial Preparedness for Long-Term Care Needs in Old Age." *International Series on Consumer Science* Part 3: 239-253.

Stacy, E.W. 1962. "A Generalization of the Gamma Distribution." *Annals of Mathematical Statistics* 33: 1187-1192.

Sun, J., E.W. Frees, and M.A. Rosenberg. 2008. "Heavy-tailed Longitudinal Data Modeling using Copulas." *Insurance: Mathematics and Economics* 42: 817-830.

Thun, M.J., R. Peto, A.D. Lopez, J.H. Monaco, S.J. Henley, Clark W. Heath Jr, and R. Doll. 1997. "Alcohol Consumption and Mortality among Middle-aged and Elderly US Adults." *New England Journal of Medicine* 337 (24): 1705–1714.

Titman, A.C., and L.D. Sharples. 2009. "Model Diagnostics for Multi-state Models." *Statistical Methods in Medical Research.*

U.S.Census Bureau. 2014. "65+ in the United States: 2010." *U.S. Government Printing Office.*

Valdez, E.A. 2000. "Developing a General Law of Mortality for Singapore." *Singapore International Insurance and Actuarial Journal* 4(1): 1-17.

Valdez, E.A. 2001. "Bivariate Analysis of Survivorship and Persistency." *Insurance: Mathematics and Economics* 29: 357-373.

Vincent, G.K., and V.A. Velkoff. 2010. "The Next Four Decades The Older Population in the United States: 2010 to 2050.". U.S. Census Bureau.

Yashin, A.I., G. De Benedictis, J.W. Vaupel, Q. Tan, K.F. Andreev, I.A. Iachine, M. Bonafe, S. Valensin, M. De Luca, L. Carotenuto et al. 2000. "Genes and Longevity Lessons from Studies of Centenarians." *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 55 (7): B319-B328.

Appendix A

A.1 Additional variables description

Occupation class categorization: Below are descriptions of industries allocated to each occupation class.

Blue Collar class

- Agriculture/Forestry/Fishing
- Mining and Construction
- Services: Food Preparation
- Durable Manufacturing
- Transportation
- Services: Private Houses/Buildings clean

White Collar class

- Finance/Insurance/Real Estate
- Entertainment/Recreation
- Public Administration
- Managerial Specialty Operation

- Prof Specialty Operation/Tech
- Clerical/Admin Support
- Services: Protection
- Health Services

Mixed Collar class

- Non-Durable Manufacturing
- Wholesales/Sales
- Retail
- Business/Repair Services
- Personal Services

HRS census region: The following are descriptions of the various regions used by HRS.

Northeast Region

- New England Division (Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut)
- Middle Atlantic Division (New York, New Jersey, Pennsylvania)

Midwest Region

- East North Central Division (Ohio, Indiana, Illinois, Michigan, Wisconsin)
- West North Central Division (Minnesota, Iowa, Missouri, North Dakota, South Dakota, Nebraska, Kansas)

South Region

- South Atlantic Division (Delaware, Maryland, District of Columbia, Virginia, West Virginia, North Carolina, South Carolina, Georgia, Florida)
- East South Central Division (Kentucky, Tennessee, Alabama, Mississippi)
- West South Central Division (Arkansas, Louisiana, Oklahoma, Texas)

West Region

- Mountain Division (Montana, Idaho, Wyoming, Colorado, New Mexico, Arizona, Utah, Nevada,)
- Pacific Divison (Washington, Oregon, California, Alaska, Hawaii)

A.2 SAS codes for Cox regression

```
DATA FINALCOUNTING_new;
SET CONTING_D_new;
T1=0;T2=2;T3=4; T4=6; T5=8; T6=10; T7=12; T8=14;
T9=16; T10=18;
ARRAY T(*) T1-T10;
array WSTAT(*);
ARRAY CENRE(*);
ARRAY CENDI(*);
ARRAY MARRY(*) ;
ARRAY MRST (*);
ARRAY KID(*) ;
ARRAY RGE(*);
ARRAY SGE(*) ;
ARRAY BM(*);

...


DEAD=0;
event=0;

IF RAGENDER=1 THEN RAGENDER=0;
ELSE IF RAGENDER=2 THEN RAGENDER=1;

/*combine blue hazard, blue non-harzard and mixed
do I=1 TO DIM(RCO);
IF (RCO(I)=0 or RCO(I)=1 OR RCO(I)=2) THEN RCO(I)=0;
ELSE IF RCO(I)=3 THEN RCO(I)=1;
END;*/

do I=1 TO DIM(RCO);
IF (RCO(I)=0) THEN RCO(I)=1;
END;

DO I=1 TO DIM(RLBFF);
IF RLBFF(I)=.A THEN RLBFF(I)=5;
END;
```

```
DO I=1 TO DIM(ADLN);
IF ADLN(I)>=4 THEN ADLN(I)=4;
END;


...

IF 0<= DURANEW_NDI <2 THEN
DO;
START=T1;
STP=DURANEW_NDI;
DEAD= RSS08_NDI;
RAGE=RGE(1);
SAGE=SGE(1);
RMARRY=MARRY(1);

...

OUTPUT;
end;

ELSE DO J=1 TO 9 WHILE ( DURANEW_NDI >= T(J)  AND DEAD=0);

START=T(J);
RAGE=RGE(J);
SAGE=SGE(J);
 RMARRY=MARRY(J);
 RMASTW=MRST(J);

 /* all missing in ADL ADL=ADLN(J);
IADL=IADLN(J);*/
AGED=AGEDN(J);
CESD=CESN(J);
CONDE=COND(J);
SHLT=SHL(J);
RPENY=RPNY(J);
SPENY=SPNY(J);

...


STP=T(J+1);
```

```
if (duranew_ndi=t(j+1) AND (wstat(j+1) NE 1 OR
wstat(j+1) NE 4 OR wstat(j+1) NE 7  )) then do;
dead=RSS08_NDI;
end;
else  IF DURANEW_NDI <T(J+1) THEN DO;
STP=DURANEW_NDI;
IF STP=START THEN STP=DURANEW_NDI+.01;
DEAD=RSS08_NDI;
  END;
output;
END;
RUN;

PROC PHREG DATA=D_ING.FINALCOUNTING_new;
WEIGHT RAWTSAMP;
CLASS RAGENDER CENDIV(Ref="3") RMARRY CONDE SHLT CESD
 HTINC VIGACT DRINKR SMOKEV;
MODEL(START,STP)*DEAD(0)= RAGE RAGENDER PAVAGE CENDIV
RMARRY CONDE SHLT CESD HTINC VIGACT
DRINKR SMOKEV BMI/TIES=EFRON ;
run;
```

A.3 Comparison with previous studies

The following tables compare the results of our variable selections in the survival model of our pension plan study to those used in earlier studies appearing in the literature.

Health variables	Agree ($\surd$) or not ($\times$)	Literature
CONDE	$\surd$	National Vital Statistics Report, 2009
SHLT	$\surd$	Idler, E.L., et al., 1997
CESD	$\surd$	Gump, B.B., et al., 2004

Lifestyle variables	Agree ($\surd$) or not ($\times$)	Literature
SMOKEV	$\surd$	Doll R. et al., 2004; Lantz et al., 1998
DRINKR	$\surd$	Thun M.J. et al., 1997; Paula M.L. et al., 2010
	$\times$	Valliant G.E., & K.Mukamal, 2001
VIGACT	$\surd$	Doll R. et al., 2004; Steven N.B., 1996
BMI	$\surd$	Campos et al., 2006; Sui et al., 2007
	$\times$	Wei et al., 1999

Financial variables	Agree ($\surd$) or not ($\times$)	Literature
RCOLLAR	$\surd$	Hurd, M. 1990, Sorlie, P.D. et al., 1995
HTINC	$\surd$	Know, D. et al., 1997, Rogers, R.G. et al., 2000
HTOTW	$\surd$	Attanasio O.P. et al., 2000; Menchik Paul 1993
HITOT	$\surd$	Moulton B.E. et al., 2012; Krieger N. et al., 2005
	$\times$	Blakely T. et al., 2003

Demographic variables	Agree ($\surd$) or not ($\times$)	Literature
AGE	$\surd$	Horuchi S. et al.,2010; Brown R.L., 1988
RAGENDER	$\surd$	Rogers R.G., 1995; Travato, F., & N. K. Lalu, 1998
RAEDUC	$\surd$	Paula M.L. et al.,2010; Sorlie P.D. et al., 1995
	$\times$	Attanasio O.P., & and C. Emmerson, 2001
RMARRY	$\surd$	Hui Liu, 2009 ; Kaplan R.M., & Richard H.K., 2006
	$\times$	Attanasio O.P., & and C. Emmerson, 2001; Rogers R.G.,1995
CENDIV	$\surd$	Greene, S.K., et al., 2005
	$\times$	Purushotham M., et al., 2011
HKIDS	$\surd$	Kotler P., & D.L.Wingard, 1989
PAVAGE	$\surd$	Hjelmborg, J.vB., et al., 2006

A.4 R codes

The following are some of the R codes used to fit the Model 1 and Model 2 with msm package. These codes relate to the long term care model in Chapter 4.

```
> rm(list=ls())
> MSMD <- read.csv("....csv")
> attach(MSMD)
> library(msm)
> InQ.q <- rbind(c(0, 0.25,  0.25), c(0.166,0, 0.166),
c(0, 0, 0))
> rownames(InQ.q) <- colnames(InQ.q) <- c("S1", "S2", "Death")
> msm_M1 <- msm(state ~ START, subject = HHIDPN, data = MSMD,
covariates = ~ AGE+RAGENDER, qmatrix =InQ.q , death = 3,
method = "BFGS", control = list(fnscale = 23000, maxit = 10000,
reltol=1e-16))

>msm_M2 <- msm(state ~ START, subject = HHIDPN, data = MSMD,
covariates = ~ AGE+RAGENDER+CONDE+BMI+RAEDUC+CESD, qmatrix =InQ.q ,
death = 3, method = "BFGS", control =
list(fnscale = 23000, maxit = 10000, reltol=1e-16))
>irtest.msm(msm_M1, msm_M2)
  -2 log LR df p
msm_M2  173.8667  4 0
```

The following is a sample code used to fit the GB2 model related to the policy termination in Chapter 5.

```
rm(list=ls())
Lapse_SMP <- read.csv("....csv")
Lapse_SMP<- subset(Lapse_SMP,twd>0)
PlanType<-ifelse(PlanType=="PAR","PLTYP",
ifelse(PlanType=="TERM","PLTYT","PLTYO" )  )
CvgSmkrCd<-ifelse(CvgSmkrCd=="S","S", ifelse(CvgSmkrCd=="N","N","C"))
attach(Lapse_SMP)
# define PlanType variables
PlanTypeP<-numeric(length(PlanType));
PlanTypeP[which(PlanType=="PLTYP")]<-1
PlanTypeT<-numeric(length(PlanType));
PlanTypeT[which(PlanType=="PLTYT")]<-1
# define PolStatCd variables
# define CvgSmkrCd variables
CvgSmkrCdC<-numeric(length(PlanType));
CvgSmkrCdC[which(CvgSmkrCd=="C")]<-1
CvgSmkrCdN<-numeric(length(PlanType));
CvgSmkrCdN[which(CvgSmkrCd=="N")]<-1
```

```
# define CvgSexCd variables
CvgSexCdM<-numeric(length(PlanType));
CvgSexCdM[which(CvgSexCd=="M")]<-1
#define rclass variables
rclassY<-numeric(length(PlanType));
rclassY[which(rclass=="Y")]<-1
# re-scale face amount
CvgFaceAmtS <- CvgFaceAmt/100000
# density function of beta distribution of the 2nd kind
"dB2" <- function(x,gamma1,gamma2)
{
num <- x^(gamma1-1)
temp <- (1+x)^(gamma1+gamma2)
den <- beta(gamma1,gamma2)*temp
result <- num/den
return(result)
}
# cumulative distribution function of beta of the 2nd kind
"pB2" <- function(x,gamma1,gamma2)
{
result <- pbeta(x/(1+x),gamma1,gamma2)
return(result)
}
# quantile function of beta of the 2nd kind
"qB2" <- function(q,gamma1,gamma2)
{
f <- function(x,q.q,gamma1.q,gamma2.q){
temp <- pbeta(x/(1+x),gamma1.q,gamma2.q)
result <- temp-q.q
return(result)
}
temp <- rep(0,length(q))
for (i in 1:length(q)){
tmpx <- ifelse(q[i]==0,0,ifelse(q[i]>0.9999999,72000000,
uniroot(f, c(0,1000000), tol = 0.0001,
q.q=q[i],gamma1.q=gamma1,gamma2.q=gamma2)$root))
temp[i] <- tmpx
}
return(temp)
}
```

```
# density function of GB2
"dGB2" <- function(x,mu,sigma,gamma1,gamma2)
{
z <- (exp(-mu)*x)^(1/sigma)
ftemp <- dB2(z,gamma1,gamma2)
temp <- z/(x*abs(sigma))
result <- ftemp*temp
return(result)
}
# cumulative distribution function of GB2
"pGB2" <- function(x,mu,sigma,gamma1,gamma2)
{
z <- (exp(-mu)*x)^(1/sigma)
result <- pB2(z,gamma1,gamma2)*
(sigma>0)+(1-pB2(z,gamma1,gamma2))*(sigma<0)
return(result)
}
# negative log-likelihood of the GB2 distribution
 for the time-until-withdrawal
"negll.GB2" <- function(parm, x) {
int <- parm[1]
bPlanTypeP <- parm[2]
bPlanTypeT <- parm[3]
bCvgSmkrCdC <- parm[4]
bCvgSmkrCdN <- parm[5]
bCvgSexCdM <- parm[6]
biss.age<-parm[7]
bCvgFaceAmtS<-parm[8]
bCvgFeUpremAmt<-parm[9]
bCvgPFeUpremAmt<-parm[10]
brclassY<-parm[11]
bCvgMeFct<-parm[12]
reg_eqn <-  int + bPlanTypeP*PlanTypeP + bPlanTypeT*PlanTypeT +
 bCvgSmkrCdC*CvgSmkrCdC + bCvgSmkrCdN*CvgSmkrCdN +
 bCvgSexCdM*CvgSexCdM + biss.age*iss.age +
 bCvgFaceAmtS*CvgFaceAmtS + bCvgFeUpremAmt*CvgFeUpremAmt
  + bCvgPFeUpremAmt*CvgPFeUpremAmt + brclassY*rclassY
  + bCvgMeFct*CvgMeFct
reg_eqn
```

```
sigma <- parm[13]
gamma1 <- parm[14]
gamma2 <- parm[15]
temp <- log(dGB2(x,reg_eqn,sigma,gamma1,gamma2))
result <- -sum(temp)
return(result)
}
# now find the parameter estimates using (un)constrained optimization
# first set the initial parameter estimates
# use linear model to find initial estimates for
the regression parameters
lm1 <- lm(log(tw) ~  PlanTypeP +  PlanTypeT +  CvgSmkrCdC +
  CvgSmkrCdN + CvgSexCdM + iss.age + CvgFaceAmtS +
  CvgFeUpremAmt + CvgPFeUpremAmt + rclassY + CvgMeFct)
init.est <- c(lm1$coeff,1,0.5,0.5)
fit.GB2 <- optim(init.est, negll.GB2, NULL, hessian=T, x=tw)
parm.hat <- fit.GB2$par
loglik.GB2 <- -fit.GB2$value
# next estimate the standard errors.
inv.GB2.Hess <- solve(fit.GB2$hessian)
parm.se <- sqrt(abs(diag(inv.GB2.Hess)))
# put together the model with the est, se, t, pval, AIC, BIC
dfe <- length(tw)-length(parm.hat);
t_ratio<-parm.hat/parm.se;
#test if diff. from 1 t_ratio[1:3]<-(parm.hat[1:3]-1)/parm.se[1:3];
pval <- pf(t_ratio*t_ratio,df1=1,df2=dfe,lower.tail=F);
output <- cbind(parm.hat,parm.se,t_ratio,pval)
output <- round(output,digits=4)
rownames(output)<- c("int",
"bPlanTypeP",
"bPlanTypeT",
"bCvgSmkrCdC",
"bCvgSmkrCdN",
"bCvgSexCdM",
"biss.age",
"bCvgFaceAmtS",
"bCvgFeUpremAmt",
"bCvgPFeUpremAmt",
"brclassY",
"bCvgMeFct",
```

```
"sigma",
"gamma1",
"gamma2")
colnames(output)<- c("estimate", "std error", "t-val","Pr>|t|");
cat("",fill=T)
print(output)
AIC<- 2*negll.GB2(parm.hat,tw) + 2*length(parm.hat);
BIC<- 2*negll.GB2(parm.hat,tw) + log(length(tw))*length(parm.hat);
cat("",fill=T)
cat(paste("AIC estimated at ",round(AIC,2)),fill=T);
cat(paste("BIC estimated at ",round(BIC,2)),fill=T);
# the SBC (Schwarz Bayesian Criterion) test
numb.par <- length(parm.hat)
SBC.stat <- loglik.GB2 - (numb.par/2)*log(length(PlanType))
out <- rbind(loglik.GB2,SBC.stat)
colnames(out) <- c("value")
rownames(out) <- c("Neg Log Likelihood","SBC criterion")
print(out)
```