DEGREE PROJECT IN MATHEMATICS,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2017

Premium influencing factors in life assurance

Study of an income parameter in mortality analysis

PONTUS WALTRÉ

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

Premium influencing factors in life assurance

PONTUS WALTRÉ

Degree Projects in Financial Mathematics (30 ECTS credits)
Degree Programme in Engineering Physics
KTH Royal Institute of Technology year 2017
Supervisor at Folksam: Anders Munk
Supervisor at KTH: Boualem Djehiche
Examiner at KTH: Boualem Djehiche

TRITA-MAT-E 2017:39
ISRN-KTH/MAT/E--17/39--SE

Royal Institute of Technology
School of Engineering Sciences
KTH SCI
SE-100 44 Stockholm, Sweden
URL: www.kth.se/sci

Abstract

The expected lifetime is steadily increasing in Sweden and the World. As the increase will eventually level out the question is: at what level will the expected lifetime be then? The outlook for or searching after factors and assumptions that can influence mortality in life insurance, are principally missing. This despite the awareness that lifestyle and sickness greatly affects the lifespan. There were two aims for this thesis. The first aim was to look into the hypothesis that there were more parameters than just gender and age that could be important to consider when doing mortality studies. The second goal was to analyse the spread of mortality to get a better understanding within a group how the mortality behaves and how low it could get. For realize these goals I've analysed the deaths within a certain population during three year, 2010-2012, depending on an income parameter. The least square approach was used to calculate estimated parameters of the Makeham mortality model. The non-parametric bootstrap model was then further used to estimate the accuracy. The results show that with a confidence level of 1 %, there is a lower expected lifetime at the age of 30 for the first income quantile than the fourth.

Premiepåverkande faktorer inom livförsäkring

Den förväntade livslängden ökar ständigt i Sverige och även i Världen. Då ökningen någon gång kommer att plana ut är frågan: när och på vilken nivå kommer medellivslängden att ligga på då? Inom livförsäkring saknas huvudsakligen någon utblick över eller sökande efter andra faktorer och antaganden som kan påverka dödlighetsantagandet. Detta trots att medvetenheten om att livsstil och sjuklighet påverkar dödligheten är hög. Det fanns två syften med denna uppsats. Det första syftet var att undersöka hypotesen att det fanns fler parametrar än bara kön och ålder som är viktiga att ta hänsyn till när man gör en dödlighetsundersökning. Det andra målet var att analysera spridningen i dödlighet inom en grupp för att få en bättre förståelse för hur dödligheten beter sig och hur låg dödligheten skulle kunna bli. För att komma underfund med detta har jag analyserat dödsfall inom ett bestånd under tre år 2010-2012, med hänsyn taget till en inkomstparameter. Minsta kvadratmetoden användes för att beräkna uppskattade parametrar till Makehams dödlighetsmodell. Därefter användes den icke parametriserade Bootstrap modellen för att bestämma noggrannheten i skattningen. Resultaten visar att med en konfidensnivå på 1 % så är det en längre förväntad medellivslängd vid 30 års ålder för den första inkomstkvantilen än för den fjärde.

Acknowledgements

I would first like to thank my supervisor at Folksam, Anders Munk, for the valuable discussions and guidance during the process. Furthermore I would like to thank Björn Nilsson for the introducing conversations around the topic and for the support along the way. I also would like to thank my supervisors at KTH, prof. Filip Lindskog and Boualem Djehiche for advise and patience. Not to mention the continuous support of Patrik Engström, Meliha Aktas, Emily Säldebring, Anders Karlsson and Jan Söderström for much needed help in all kind of areas, without whom the thesis would not have been possible. Additionally I would like to express my gratitude towards my family, who have supported me throughout the process.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Even though legislation and European directives often prevents the use of different mortality assumptions for men and women at the determination of premiums, these are still used in life insurance for determining reserves. In this area there exist a relatively large certainty in the models and the parameters that are used. In non-life insurance, there exits an innovation mentality which takes its appearance through the searching for and utilizing of new parameters as well as inventing new work procedures. In life insurance however, as opposed of non-life insurance, the outlook for or searching after other factors and assumptions that can influence the mortality assumptions are principally missing. This despite the awareness that lifestyle and sickness greatly affects the lifespan. It has been shown in several reports, for instance [2] and [4], that there are significant lifespan differences between geographical areas, civil status, living and education. These reports however have been written from an demographical point of view and not an insurance related angle.

1.2 Aim of thesis

The aim with this study is to look into the hypothesis that there are more parameters than just the gender that could be important to consider when doing mortality studies. As there are some or large correlation between income and education or living, this parameter have been chosen for further study in this analysis. Furthermore the prognoses that are used to determine the future mortality parameters are based on historical statistics, which gives an uncertainty, as the factors that affected an improved life span historically not necessarily needs to affect the improved life span in the future. The longer average life expectancy of today and the nearby future derives from people living longer after retirement. At the same time evidence suggest that at the latest stages of life, the mortality remains the same as before. By analysing the spread that exist within a certain group, it would be possible to see how low the mortality curve could get, with the medicine and health care that exist today.

1.3 Earlier research

There are several studies both in Sweden and internationally that illustrate that there are more parameters then just sex that affects the life span, for example [2] and [4] mentioned earlier. But these have been done from a population demographical approach and have not studied how this could affect an insured population and the parameters and models that an insurance company are using.

1.4 Definitions

1.4.1 Base amount

Price base amount

The price base amount is based on the consumer index price. It has several uses, which all work to ensure that a value do not decline because of an increase of the inflation. This amount is adjusted after the general price development in society. The Swedish government decides the amount one year at a time. The price base amount is used within the social insurance and tax systems, for example to decide the guaranteed minimum retirement pension and ensuring that sickness benefits and study support do not decline. The price base amount was 44,400 in 2014 [16] [17] [18].

Increased price base amount

The increased price base amount, like the price base amount above, evolves with the inflation and is settled by the Swedish government It is used for calculating pension points for supplementary pensions for those receiving a pension based on the older regulations. The increased price base amount was 45,300 in 2014. [17].

Income price base amount

Some functions previously served by the price base amount have now been transferred to the income index and the income price base amount. The Swedish government determines the income price base amount based on the salary development of the society, which is governed by the income index. So as to ensure that pension balances and pension rights earned follow general income development instead of the inflation development, pensions are adjusted upward each year by the annual change in the income index. This amount is more precisely used to calculate the income roof for retirement pension. It is generally also used to decide the size of defined-benefit pension in occupational pension. Furthermore it is used to decide the premium the employer shall pay in a defined-contribution pension plan. Additionally it is used to compute the maximum pensionable income. The present ceiling is set to 7.5 income base amounts. The increase base amount was 56,900 in 2014 [16] [17] [18].

1.4.2 Yearly average point

The yearly average point is used in this research as an income parameter. A yearly point is calculated each calender year, that consists of the ratio between the employees pensionable salary and the income base amount the corresponding year. The yearly average point is defined according to [6]. It is determined from the employees yearly points during the seven calendar years that closest precede the actual calendar year the calculation occurs. The yearly average point constitute of the average of the five highest yearly points. If the yearly points cannot be calculated for all the above-mentioned seven years the yearly average point shall be calculated according to the following rules:

1. If the pension plan hasn't applied to the employee during an entire calender year, that year will disregarded.
2. If at least four of the above-mentioned seven yearly points exists, the two lowest will be disregarded.
3. If two or three of the above-mentioned seven yearly points exist, the lowest will be disregarded.
4. If there is only one of the above-mentioned seven yearly points, then it will constitute the yearly average point.
5. If the pension plan has not applied during any entire year of the above-mentioned seven yearly points, the yearly average point is determined from the pre-settled yearly salary of the employment, including addition of annual leave.

1.4.3 Normal amount

The normal amount is an income parameter used before 1985 for pension calculations, and is calculated as

$$normal\ amount = yearly\ salary \cdot time\ factor \cdot 0.65$$

and is recalculated to yearly average points using the price base amount of the retirement year of the individual.

Chapter 2

Theory

2.1 Stochastic model

Let us consider a population with individuals aged x years. We denote the future lifetime of a randomly chosen individual by $T(x)$, implicating that the age at death of the individual will be $T + x$ years. Let the future lifetime T be a non-negative continuous stochastic variable with the probability distribution function

$$F_x(t) = P(T_x \leq t),\ t \geq 0.$$

The function $F_x(t)$ is thus the probability that an individual will die within t years. The probability density function f is defined by

$$f_x(t) = F_x'(t),\ t \geq 0. \tag{2.1}$$

We also introduce the survival function

$$l_x(t) = 1 - F_x(t) = P(T_x > t),\ t \geq 0. \tag{2.2}$$

l_x is the probability that a x year old individual survives for at least t more years. Furthermore

$$l_x(t) = \frac{l_0(x+t)}{l_0(x)},\ t \geq 0. \tag{2.3}$$

(2.3) shows the important connection between the survival function of an individual aged x years and the survival function of a newborn. Note that from now on, the denotations T_0, $F_0(t)$, $f_0(t)$ and $l_0(t)$ will be written as T, $F(t)$, $f(t)$ and $l(t)$. The relation between the probability density function and the survival function can be shown using (2.1) and (2.2) as

$$l'(t) = \frac{d(1-F(t)}{dt} = -f(t). \tag{2.4}$$

Now regard the age interval $(x, x+dx)$. The probability to pass away during this interval, given that the individual lives at the age x and that dx is small, is approximately equal to $\mu_x dx$. We denote μ_x as the mortality intensity and define it as

$$\mu_x = \frac{f(x)}{1-F(x)}.$$

Using (2.2) and (2.4) the equation above can also be written as

$$\mu_x = -\frac{l'(x)}{l(x)} = -\frac{d(\ln l(x))}{dx}.$$

Now applying that $l(0) = 1$ this can be reordered and written as

$$l(x) = \mathrm{e}^{-\int_0^x \mu_s \, \mathrm{d}s}.$$

Let us now define the expected remaining lifetime e_x as

$$e_x = E(T_x).$$

By using the definition of expectation we get

$$e_x = E(T_x) = \int_0^\infty (1 - F_x(t)) \, \mathrm{d}t = \int_0^\infty \frac{l(x+t)}{l(x)} \, \mathrm{d}t. \tag{2.5}$$

A fair approximation of the expected remaining lifetime of the equation (2.5) can be calculated using the trapezoidal rule. The trapezoidal rule states that $\int_a^b f(t) \, \mathrm{d}t$ can be approximately calculated as

$$\int_a^b f(t) \, \mathrm{d}t \approx \frac{h}{2} \cdot \sum_{i=0}^{n-1} (f_i + f_{i+1}). \tag{2.6}$$

Using equation (2.6), letting h be equal to 1 and $n \to \infty$, the equation (2.5) can be calculated as

$$e_x \approx \frac{h}{2} \cdot \sum_{i=0}^{\infty} \left(\frac{l(x+i)}{l(x)} + \frac{l(x+i+1)}{l(x)} \right) = \left[\sum_{i=0}^{\infty} \frac{l(x+i)}{l(x)} \right] - \frac{1}{2}, \; x = 0, 1, 2...,$$

as $l(\infty) = 0$. It is possible to use the Euler-Maclaurin summation formula to attain an even better accuracy in the calculations above. The Euler-Maclaurin summation formula is as follows

$$\frac{h}{2} \cdot \sum_{i=0}^{n-1} (f_i + f_{i+1}) = \int_a^b f(t)\,\mathrm{d}t + \frac{h^2}{12}(f'(b) - f'(a)) + R(h),$$

where $R(h)$ is a remainder term that contains terms of the fourth order of h and higher. Using that $f(t) = l(x+t)/l(x)$ we have

$$\frac{df(t)}{dt} = \frac{d[l(x+t)/l(x)]}{dt} = \frac{-l(x+t) \cdot \mu_{x+t}}{l(x)}.$$

Using that the upper integration limit $b \to \infty$ the term $f'(b)$ equals 0. We can now calculate e_x as

$$e_x \approx \left[\sum_{i=0}^{\infty} \frac{l(x+i)}{l(x)}\right] - \frac{1}{2} - \frac{1}{12} \cdot \mu_x, \ x = 0, 1, 2... \tag{2.7}$$

2.2 Life expectancy models

There are several well-known mathematicians who have presented important breakthroughs to establish lifespan tables in life insurance, among others deMoivre (1729), Gompertz (1825), Makeham (1860), Sang(1868), Weibull (1939) and Lee-Carter (1992). One of the prominently individuals in Sweden was Pehr Wargentin (1717-1783), whose work laid the foundation to the Swedish statistical government agency, Statistiska centralbyrån. In this chapter some of the historically most commonly used mortality functions will be presented, along with their respective flaws and strengths. Alongside these mortality functions, some of the historically important mortality intensity parameter sets that have been used the last one hundred years in Sweden will be introduced.

2.2.1 Gompertz mortality function

Benjamin Gompertz presented 1825 his law of mortality in his article [20], where he assumed that the mortality intensity was exponentially age dependent according to the following equation

$$\mu(x) = \beta \cdot \mathrm{e}^{(\gamma \cdot x_i)}, \ x \geq 0, \tag{2.8}$$

with $\beta > 0$ and $\gamma > 0$. This equation were used in several countries and with relatively good results. The problem was that this model didn't catch the infant mortality, the fatalities at younger ages nor the mortality for the very old, where the phenomenon of late-life mortality deceleration occurs, where the death rates increase at a decreasing rate than this model predicts.

2.2.2 Makeham law of mortality

The life model that have been, and still is in use in large parts of Scandinavia and particularly in Sweden is called the Makeham law of mortality or the Makeham distribution. The model was first presented 1860 by William Makeham, in his article "On the Law of Mortality and the Construction of Annuity Tables" [21]. The model builds upon three parameters instead of Gompertz two, where Makeham uses an age independent constant together with Gompertz age dependent ones. The Makeham formula for the mortality intensity looks like

$$\mu(x) = \alpha + \beta \cdot \mathrm{e}^{(\gamma \cdot x)}, x \geq 0 \tag{2.9}$$

where $\mu(x)$ is the mortality intensity for an individual at the age of x and $\alpha > 0$, $\beta > 0$ and $\gamma > 0$. It better takes into account the risk to die because of an accident or of other causes that are not age dependent, such as infant mortality and men in their twenties. Today the Makeham model is often used in combination with other models to even better catch different aspects, like an improvement factor catching trends in mortality or a generation model, where different generations receive various Makeham parameters. Another improvement for the Makeham model is to use a linear function for very old individuals over a certain age, normally somewhere between 95 and 100 years old, that is

$$\mu(x) = \begin{cases} \alpha + \beta \cdot \mathrm{e}^{(\gamma \cdot x)}, & \text{for } 0 \leq \text{x} \leq \text{w} \\ \mu(w) + k(x - w) & \text{for x} > \text{w.} \end{cases}$$

E.g. the Swedish Pension Agency uses an age parameter w of 97 when determining the linear trend, according to [23].

2.2.3 Lee-Carter model

The Lee-Carter model was presented by Ronald D Lee and Lawrence Carter in 1992 and is a numerical algorithm that is used in mortality and life expectancy forecasting. The idea with the model is to find a univariate time series vector κ_t which might capture up to 80-90 % of the mortality trend. The model uses singular value decomposition to achieve this. Let $m(x, t)$ be the central death rate for the age x in the year t. The matrix of death rates are fitted by the model according to

$$\ln(m(x,t)) = a_x + b_x k_t + \epsilon_{x,t},$$

or

$$m(x,t) = \exp\left(a_x + b_x k_t + \epsilon_{x,t}\right),$$

for appropriately chosen sets of a_x, b_x and $k_{x,t}$. Here the ϵ-term is an error term with mean 0 and variance σ_ϵ^2.

2.2.4 Sum at risk and economical mortality

Sum at risk is defined as the reserve just after a death minus the reserve just after according to the equation 2.10

$$RS = S - V \tag{2.10}$$

where S = just after the death and V = the reserve just before. When doing an economical mortality the weights and the stochastic variables in subchapter "Estimation of Makeham parameters" the estimations are based on the sum at risk instead of the amount of individuals.

2.2.5 Historical distributions used in Sweden

Sweden was the first country publishing a national lifespan table, as far back as 1755. Other countries followed, for example Netherlands (1816), France (1817), Norway (1821), England (1843), Germany (1871), Switzerland (1876) and USA (1900), though many of these countries had produced regional tables for a long time. Because the population longevity is constantly increasing, the parameters have been changing with time. Some of the most important sets of parameters that have been used in Sweden are:

1. L39
2. L55
3. G64
4. M64
5. M90
6. DUS06
7. DUS14

The first sets of parameters 1-5 have used the Makeham model presented in 2.2.2. During the years 1989 and 1990 the Swedish committee, Grundkommittén, was working with the mortality among the insured population of Sweden. They published the set of Makeham parameters labelled M90, which are still in use in some companies today. M90 use the base 10 instead of e as well as using four parameters α, β, γ and f according to

$$\mu(x) = \alpha + \beta \cdot 10^{\gamma \cdot (x-f)}, x \geq 0,$$

where $\alpha = 0.001$, $\beta = 0.000012$, $\gamma = 0.44$ and

$$f = \begin{cases} 0, & \text{for men} \\ 6 & \text{for women.} \end{cases}$$

Here the f is used as an age dislocation parameter. Insurance Sweden, Svensk försäkring, have then used the Lee-Carter model presented in 2.2.3 in both their studies DUS06 [2] and DUS14 [3].

2.3 Estimation of the Makeham parameters

Observe a randomly chosen individual from a population of n individuals. Let L_i be the remaining lifetime of the individual i. Look at the age interval $(x,\ x+h)$. Defining the stochastic variables R_i and D_i as

$$R_i = \min\left(L_i, h\right)$$

and

$$D_i = \begin{cases} 1 & L_i \leq h \\ 0 & L_i > h \end{cases}$$

for $i = 1, 2, ..., n$. R_i denotes the risk time in the time interval $(x,\ x+h)$ and D_i denotes whether the individual i passes away in that interval. Calculates the observed mortality intensity as

$$\hat{\mu}(t) = \frac{D_i}{R_i}. \tag{2.11}$$

The distribution of $\hat{\mu}$ is very complex. According to Beyer, Keiding, Simonsen [5] the estimated $\hat{\mu}$ is asymptotically normal distributed with mean μ and variance σ^2, that is

$$\sqrt{n}(\hat{\mu} - \mu) \sim asN(0, \sigma^2)$$

Regard an individual at the age of x at the observation time t. Defining $N_x(t)$ as the number of people living at the end of calender year t and that turns x years old during calender year t. Similarly defining $D_x(t)$ as the number of individuals that passes away during calender year t and turned or would turn x years old during calender year t. Finally the exposure $E_x(t)$ is defined as the fraction of days that the passed away individual lived during the calender year t and turned or would turn x years old during the calender year t.

In this survey, the observed mortality intensity has been calculated as

$$\hat{\mu}(t) = \frac{D_x(t)}{N_x(t) + D_x(t) \cdot E_x(t)}, \tag{2.12}$$

Want to use the Makeham model from chapter 2.2.2, that is

$$\mu(x) = \alpha + \beta \cdot \mathrm{e}^{(\gamma \cdot x)}, x \geq 0 \tag{2.13}$$

where $\alpha + \beta > 0$, $\beta > 0$ and $\gamma \geq 0$. As a further condition α is set to 0 when it became negative. The opposite might have given negative mortality probabilities at lower ages. Using the least squares method to calculate a fitted curve for our observed $\hat{\mu}$ values. The least squares is a standard approach to the approximate solution of sets of equations in which there are more equations than unknowns. The method means that the overall solution minimizes the sum of the squares of the errors made in the results of every single equation. That is solving

$$\min Q = \min \sum_{i=1}^{n} r_i^2,$$

where r_i is equal to the difference between our observed $\hat{\mu}$ from equation (2.12) and the μ we want to fit to the Makeham model in equation (2.13). But as our observed mortality intensities have different uncertainties, a weighted least squares method must be used. Observations with a higher precision, that is a lower variance, will thus have a higher efficiency than observations with a lower precision. Given the above we want to minimize

$$Q = \sum_{i=1}^{n} w_{x_i} \cdot (\hat{\mu}_{x_i} - (\alpha + \beta \cdot \mathrm{e}^{\gamma \cdot x_i}))^2 \tag{2.14}$$

where w_{x_i} is an appropriate weight. The weights are chosen as $1/\sigma^2(\hat{\mu}_{x_i})$, as we want observations with higher precision, that is a lower variance, to have a larger weight in the calculations. According to [1], for large n $\hat{\mu}_{x_i}$ has an expected value of μ_{x_i} and a variance of $\mu_{x_i}^2/D_{x_i}$. Given this, we can use (2.11) and that according [1] $R_{x_i} \approx N_{x_i}$ to get

$$w_{x_i} = \frac{D_{x_i}}{\hat{\mu}_{x_i}^2} = \frac{R_{x_i}}{\hat{\mu}_{x_i}} \approx \frac{N_{x_i}}{\hat{\mu}_{x_i}}. \tag{2.15}$$

Fixating c as a constant and solving the equation system

$$\frac{\partial Q}{\partial \alpha} = 0 \tag{2.16}$$

$$\frac{\partial Q}{\partial \beta} = 0 \tag{2.17}$$

Beginning with (2.16)

$$\begin{aligned} \frac{\partial Q}{\partial \alpha} &= 2 \cdot \sum_{i=1}^{n} w_{x_i} \cdot (\hat{\mu}_{x_i} - (\alpha + \beta \cdot \mathrm{e}^{(\gamma \cdot x_i)})) \cdot (-1) \\ &= \sum_{i=1}^{n} w_{x_i} \cdot \hat{\mu}_{x_i} - \alpha \cdot \sum_{i=1}^{n} w_{x_i} - \beta \cdot \sum_{i=1}^{n} w_{x_i} \cdot \mathrm{e}^{(\gamma \cdot x_i)}. \end{aligned}$$

Which, by reordering the equation give α as

$$\alpha = \frac{\sum_{i=1}^{n} w_{x_i} \cdot \hat{\mu}_{x_i} - \beta \cdot \sum_{i=1}^{n} w_{x_i} \cdot \mathrm{e}^{(\gamma \cdot x_i)}}{\sum_{i=1}^{n} w_{x_i}} \tag{2.18}$$

Continuing with (2.17)

$$\begin{aligned}\frac{\partial Q}{\partial \beta} &= 2 \cdot \sum_{i=1}^{n} w_{x_i} \cdot (\hat{\mu}_{x_i} - (\alpha + \beta \cdot \mathrm{e}^{(\gamma \cdot x_i)}) \cdot \mathrm{e}^{(\gamma \cdot x_i)}) \cdot (-1) \\ &= \sum_{i=1}^{n} w_{x_i} \cdot \hat{\mu}_{x_i} \cdot \mathrm{e}^{(\gamma \cdot x_i)} - \alpha \cdot \sum_{i=1}^{n} w_{x_i} \cdot \mathrm{e}^{(\gamma \cdot x_i)} - \beta \cdot \sum_{i=1}^{n} w_{x_i} \cdot \mathrm{e}^{(2 \cdot \gamma \cdot x_i)}. \qquad (2.19)\end{aligned}$$

Inserting equation (2.18) into (2.19) and solving (2.19) for β gives

$$\beta = \frac{\sum_{i=1}^{n} w_{x_i} \cdot \hat{\mu}_{x_i} \cdot \mathrm{e}^{(\gamma \cdot x_i)} \cdot \sum_{i=1}^{n} w_{x_i} - \sum_{i=1}^{n} w_{x_i} \cdot \hat{\mu}_{x_i} \cdot \sum_{i=1}^{n} w_{x_i} \cdot \mathrm{e}^{(\gamma \cdot x_i)}}{\sum_{i=1}^{n} w_{x_i} \cdot \mathrm{e}^{(2 \cdot \gamma \cdot x_i)} \cdot \sum_{i=1}^{n} w_{x_i} - (\sum_{i=1}^{n} w_{x_i} \cdot \mathrm{e}^{(\gamma \cdot x_i)})^2}$$

With

$$\begin{aligned} w &= \sum_{i=1}^{n} w_{x_i}, \\ m_{11} &= \sum_{i=1}^{n} w_{x_i} \cdot \hat{\mu}_{x_i} \cdot \mathrm{e}^{(\gamma \cdot x_i)}, \\ m_{01} &= \sum_{i=1}^{n} w_{x_i} \cdot \hat{\mu}_{x_i}, \\ m_{20} &= \sum_{i=1}^{n} w_{x_i} \cdot \mathrm{e}^{(2 \cdot \gamma \cdot x_i)}, \\ m_{10} &= \sum_{i=1}^{n} w_{x_i} \cdot \mathrm{e}^{(\gamma \cdot x_i)}, \end{aligned}$$

α can be written as

$$\alpha = \frac{m_{01} - \beta \cdot m_{10}}{w} \qquad (2.20)$$

and β can be written as

$$\beta = \frac{w \cdot m_{11} - m_{10} \cdot m_{01}}{w \cdot m_{20} - m_{10}^2}. \tag{2.21}$$

Now we can use (2.15), (2.20) and (2.21) in (2.14) and minimizing Q with a varying γ to get the fitted three Makeham parameters for the observed data.

2.4 Regression analysis

2.4.1 Simple Linear Regression

The core of regression analysis is to explain every observation of the dependent variable y with two parts; a systematic component and a random component. A simple linear regression model is thus a mathematical relationship between two variables and can be written as

$$y = \beta_1 + \beta_2 x + \epsilon. \tag{2.22}$$

The systematic component of y is its conditional mean, E(y|x) $= \beta_1 + \beta_2 x$. The random component is the difference between y and its conditional mean and is called the random error term and denoted by ϵ. The expected value of the error term given x is

$$E(\epsilon|x) = E(y|x) - \beta_1 - \beta_2 x = 0. \tag{2.23}$$

As the dependent variable y and its random error term ϵ differ only by a constant term, their variance must be homoscedastic with an identical and equal to a finite σ^2, that is

$$var(\epsilon) = \sigma^2 = var(y). \tag{2.24}$$

This means that the probability density functions of y and ϵ have the same shape, even though their locations differ. (2.22), (2.23) and (2.24) are known as the first, second and third assumption of the simple linear regression model. The fourth assumption states that the covariance between any pair of random errors ϵ_i and ϵ_j is

$$cov(\epsilon_i, \epsilon_j) = cov(y_i, y_j) = 0.$$

The fifth assumption states that the variable x is not random and must take at least two different values. There exist a sixth optional assumption stating that the values of ϵ are normally distributed with

$$\epsilon \sim N(0, \sigma^2)$$

if the values of y are normally distributed, and vice versa.

Estimating the Regression Parameters

The observations are denoted by y_i and we assume they follow a simple linear regression

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i$$

where the errors ϵ_i are independent and identically distributed with zero mean and variance σ^2. The parameters β_1 and β_2 of the true regression line will be estimated by use of the least squares principle. To fit a line to the data values y_i we want to minimize the sum of the squares of the vertical distances from each point to the line. The estimated intercept $\hat{\beta_1}$ and slope $\hat{\beta_2}$ of the line are the least squares estimates of β_1 and β_2. The fitted line has the shape

$$\hat{y_i} = \hat{\beta_1} + \hat{\beta_2} x_i.$$

The differences between the observed and predicted values of y are called the least squares residuals and are given by

$$\hat{\epsilon_i} = y_i - \hat{y_i} = y_i - (\hat{\beta_1} + \hat{\beta_2} x_i). \tag{2.25}$$

Calculating

$$Q = min \sum_{i=1}^{n} \hat{\epsilon_i}^2$$

gives the least squares estimators as

$$\begin{aligned}
\hat{\beta}_1 &= \overline{y} - \hat{\beta}_2\overline{x} \\
\hat{\beta}_2 &= \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}
\end{aligned}$$

where $\overline{y} = \sum y_i/n$ and $\overline{x} = \sum x_i/n$.

Estimation of the Error

If the model assumptions hold, the expected value of $\hat{\beta}_1$ is β_1 and of $\hat{\beta}_2$ equal to β_2. The variances and covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$ are calculated as

$$\begin{aligned}
var(\hat{\beta}_1) &= \sigma^2 \frac{\sum x_i^2}{n\sum(x_i - \overline{x})^2} \\
var(\hat{\beta}_2) &= \frac{\sigma^2}{\sum(x_i - \overline{x})^2} \\
cov(\hat{\beta}_1, \hat{\beta}_2) &= \sigma^2 \frac{-\overline{x}}{n\sum(x_i - \overline{x})^2}.
\end{aligned}$$

Now we only have to estimate the variance of the random error term, σ^2. The variance is

$$var(\epsilon_i) = \sigma^2 = E[\epsilon_i^2] - E[\epsilon_i]^2 = E[\epsilon_i^2]$$

as $E[\epsilon_i] = 0$. We will estimate this, by using the average of squared errors. Instead of using the random errors ϵ_i, whom are unobservable, we will use the least squares residuals $\hat{\epsilon_i}$, recall (2.25). Thus the variance can be calculated as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\epsilon_i}}{n - 2}.$$

To make the estimator $\hat{\sigma}^2$ unbiased, the number "2" is subtracted in the denominator. This is the number of regression parameters, our β_1 and β_2.

2.5 The Non-Parametric Bootstrap

Bootstrapping is a method to estimate and to measure accuracy of sample estimates, such as a confidence interval. The idea is to create new samples from the original set and to then calculate approximative measures of accuracy. The most common reason to apply the Bootstrap model is when the form of the underlying distribution from which a sample is taken is unknown. Suppose we have the observations $x_1, ..., x_n$ of independent and identically distributed random variables $X_1, ..., X_n$ and that we have an unknown distribution F of the $X_k s$. Bootstrapping allow the possibility to gather alternative versions from the observed data sample. This is done by assuming that the random sample data set from a population has the characteristics that roughly match that of the source population. By repeatedly re-sampling the observed sample itself, bootstrapping enables estimates that are distribution independent. We can still use the sample mean $\overline{x}$ as a point estimate for μ. The bootstrap method is roughly based on the law of large numbers. The re-sampling is done by randomly selecting the same number n as in the original observation, but with replacement, with many of the original sample repeated while others would be excluded. The probability that none of the $x_k s$ are drawn twice among n tries is $n!/n^n$, thus that $X_k^* \neq X_j^*$ for all $j \neq k$ is very small for a larger n. By doing this several times, we create a large number N of data sets that we might have seen. This will produce a new sample $X_1^{*(j)}, ..., X_n^{*(j)}$ that is uniformly distributed on the set of the original observations $x_1, ..., x_k$, with j in the set $1, ..., N$. The empirical distribution of $X_1^{*(j)}, ..., X_n^{*(j)}$ is written as $F_n^{*(j)}$. The bootstrap principle states that $F_n^{*(j)} \approx F_n$. Even though $X_1^{*(j)}, ..., X_n^{*(j)}$ are not samples from F, they will have most of the characteristics of the real sample, as long as n and N are sufficiently large. It is now possible to use the probability distribution $\hat{\theta}^*$ to form an approximate confidence interval. Calculating the estimated probability function as

$$\hat{\theta}_j^* = \theta(F_n^{*(j)})$$

and the residuals as

$$R_j^* = \hat{\theta}_{obs} - \hat{\theta}_j^*.$$

We can now use this to form the approximated confidence interval

$$I_{\theta,q} = (\hat{\theta}_{obs} + R^*_{[N(1+q)/2]+1,N}, \hat{\theta}_{obs} + R^*_{[N(1-q)/2]+1,N}).$$

Chapter 3

Data

The study in this report have covered the years 2010, 2011 and 2012. To get a better statistical foundation the three years are calculated together, under the assumption that there are no changes in the mortality during the three years. A restriction is that those that passed away during the year must have been policy holders the year before. Furthermore the income is fetched from the previous year, as the deceased does not have any income registered the year of death. To make it statistically accurate both the living and the deceased must have been alive and assured the year before. That is, the income data is fetched from between 2009 to 2011.

The individuals that are examined are between 30 and 100 years old. As the underlying data comes from working individuals, there are very few individuals under the age of 20. Furthermore people under the age of 30 have little importance in life insurance as most doesn't contribute with premiums until late twenties, and usually with only small amounts. At the same time there are very few deaths at lower ages, which makes the data rather poor for lower ages. To get a satisfactory amount of statistical data, the lowest age included in this rapport will be of 30 year old. At the same time, the Makeham model is badly correlated with observed data past the age of 100. Moreover the data input is too fragile to make a statistical analysis for this group. [1]

Some of the data have not been used, as those with an income of 0 have not been included in this research. These consists part of individuals that did not actually have an income the targeted year. The larger group however includes people where it was not possible to fetch income information. Some groups with older collective agreements used instead of ÅMP another type of income parameter to calculate their benefits. Other groups, for example retired, simple misses this information. In these cases the information was as much as possible fetched and complemented from other data systems and older files.

The underlying material for this thesis comes from individuals that are or have been working for the municipalities and counties of Sweden. These municipalities and counties are indebted to these individuals, and are being helped by KPA to continuously calculate their debt. There are information of about 3 000 000 distinct assurances and about 2 000 000 distinct individuals in the database. The individuals that are included are composed only of assured with current assurance. This data have later been completed with other income data from older files, among other things the normal amount for individuals retiring before 1985.

In order to get one distinct income to each individual, the data was cleared of doublets in the order as follows:

1. When a policy holder have had several entries with different incomes in the data, the post using the latest date have been used. 2. If the policy holder have had both premium paying and non-premium paying (paid-up or paid-out) life-insurances, then premium paying entries have been used before the paid up but after the paid out entries. 3. In cases when the individual have been working in several municipalities or counties, have several pension types or retired at different times, the entry with the highest income have been used.

The total exposed population is displayed in Figure 3.1. There are 4 042 286 women and 1 360 985 men that are exposed during the three years. As can be seen in the diagram, there are about 110 000 to 120 000 individuals per year up to the age of about 65 years of age, when the numbers begins to decline, leaving a very small population over the age of 95. Another remark is that there are 74.8 % women in the sample population. Furthermore this percentage increases with the ageing population.

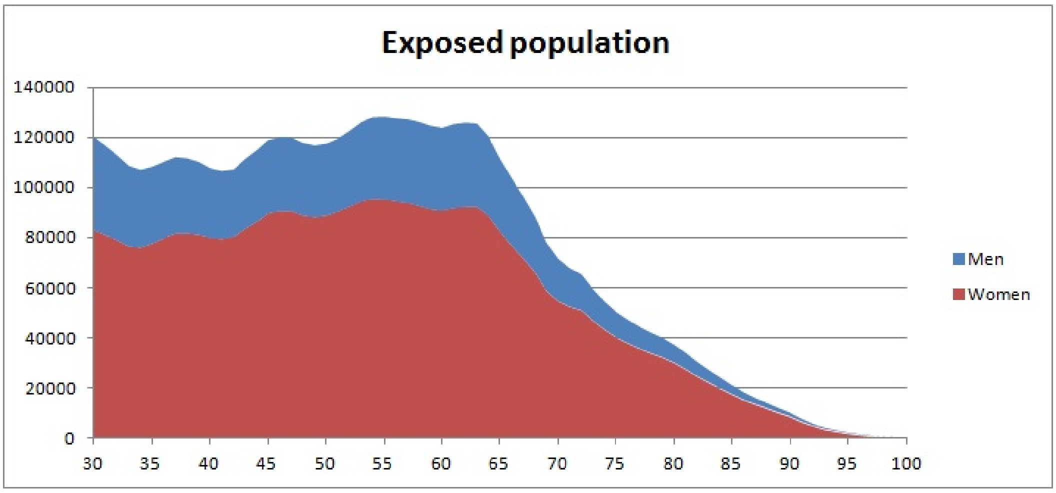

Figure 3.1. The exposed population in the study, divided in women, men and the total amount.

If we instead look at the deceased part of the sample population, see Figure 3.2, the numbers range between 37 that passed away at the age of 30. After 30 the number of deceased raises up until the age of 87, where 1 497 individuals passed away. After this point, the number of deaths dwindle to 57 at the age of 100.

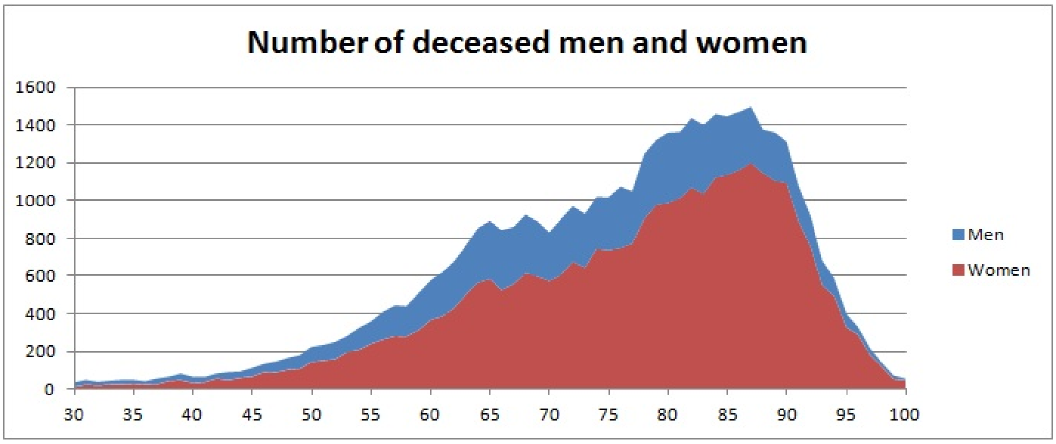

Figure 3.2. The number of deceased individuals in the study, divided in women, men and the total amount.

Chapter 4

Methods and results

In this chapter the methods that have been used will be presented, as well as the findings of the results. Excel and VBA coding have been used as the main tool of calculations and graphics.

4.1 Estimation of the Makeham parameters

We start by looking at the entire sample of the population frpm 2010 to 2012. We can calculate the observed and fitted μ as described in 2.3. The fitted parameters α, β and γ are given as in Table 4.1.

The resulting graphs of the fitted Makeham curves together with the observed data are given in the figure 4.1.

The y-axes are log normally scaled, giving the observed and fitted Makeham functions an almost linear shape. As can be seen, the functions fit the observed values rather well, even though there are some small volatility at the lower ages, which of course is explained by fewer deaths in these ages. In the figure 4.2 can be shown how Q is minimized by varying the γ with different values, in this case with a step of 0.00005, with Q being calculated from function (2.14). In this case, it is the minimized Q for the women in our sample.

	α	β	γ	Q
women	0.000237	$2.50 \cdot 10^{-6}$	0.1198	292.03
men	0.000439	$6.00 \cdot 10^{-6}$	0.1136	85.93
total	0.000303	$3.53 \cdot 10^{-6}$	0.1166	313.10

Table 4.1. Makeham parameters for women, men and of the total sample

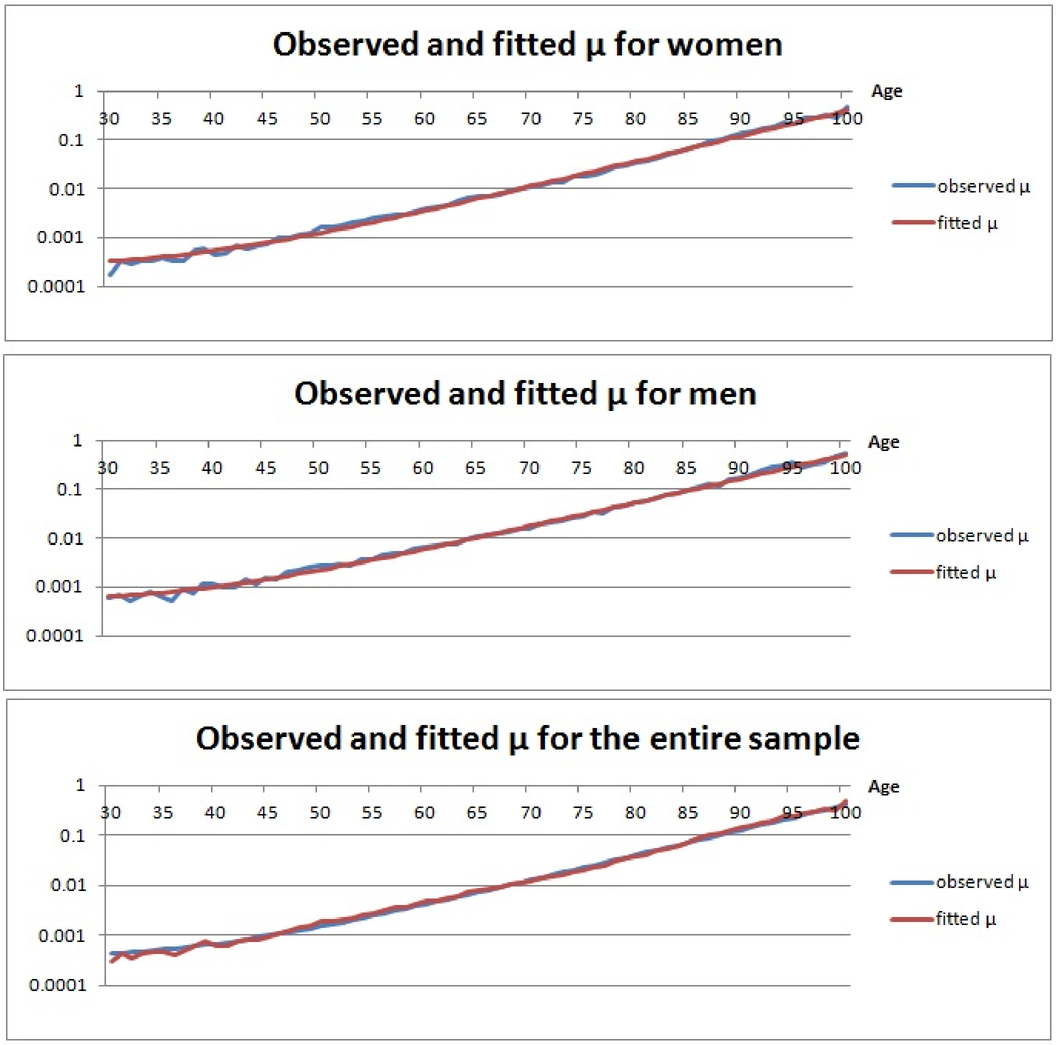

Figure 4.1. The observed and the fitted μ calculated for the women, men and the entire sample

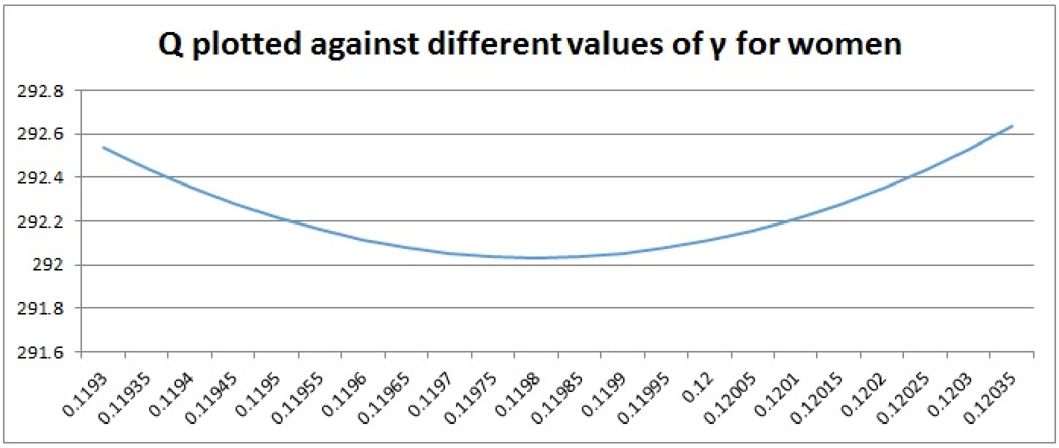

Figure 4.2. Calculating different Q's by varying γ resulting in an equation of the second degree.

4.1.1 The Makeham distribution with an income parameter

Considering that the population will be divided into four different income groups. The exposed population will have about 30 000 individuals in each age group and quantile, with about 22 500 women and 7 500 men. This could of course affect the results, as the calculations will be much more accurate for the women. That there are three times more women than men might play a role in how large the variance will be, as every death will have a larger impact for the men then the women. The calculations have been done with regards to sex, which is standard in similar mortality studies. But they have also been done on the total sample populous, as insurance companies often have to use sex independent premiums by law. Another reason is to be able to compare an income parameter to a sex parameter.

Quantiles have then been used to divide the population into four different groups of equal size. A quantile is the value of a variable under which a certain 25-percentage of the observations of the variable occur. That is, the third income quantile is the income value where 75 % of the population have their incomes. For a 60 year old individual the quantiles looked like in Table 4.2, based on incomes from 2009 to 2011. As the incomes are based on yearly average points that are to be multiplied with the income base amount, see 1.4.1 and 1.4.2. In the table 4.2, the income base amount from 2017 have been used.

The first quantile for the sample population has an upper yearly income of 242 000, and thus includes everyone with a yearly incomes between 0 and 242 000.

Next the passed away was similarly examined how many had died in each age, sex and income quantile group, where the quantiles were determined by the living individuals, as well as their respective exposure as defined in 2.3.

By using smaller groups, as quantiles, the variability of the deaths increases. To

	Women		**Men**		**Total**	
	lower	upper	lower	upper	lower	upper
Quantile 1	0	235 000	0	270 000	0	242 000
Quantile 2	235 000	310 000	270 000	359 000	242 000	322 000
Quantile 3	310 000	390 000	359 000	478 000	322 000	413 000
Quantile 4	390 000	∞	478 000	∞	413 000	∞

Table 4.2. Yearly income (in SEK) for respective quantile group for men, women and entire sample, based on income data from 2009 to 2011

somewhat counter this effect when calculating the adjusted Makeham parameters, the mortality intensity used in the calculation of the weight is the fitted mortality intensity parameter of the entire sample population, thereby somewhat decreasing the effects of variances in the data, with the weight being calculated as in equation (2.15). γ was now being varied, to find a minimum of Q, according to equation (2.14). This then gave different sets of parameters of α, β and γ for each sex and quantile. γ was varied with a step of 0.0001.

The resulting Makeham parameters are given as in the tables 4.3, 4.4 and 4.5. As can be seen in the tables, the first and second quantiles are very similar, a part from their α-parameters. The third and fourth quantiles have lower β- and higher γ-parameters then the first two in all three sets of tables, giving them a lower but steeper shape of the curve.

Women	α	β	γ	Q
Quantile 1	0.000334	$4.24 \cdot 10^{-6}$	0.1139	158.65
Quantile 2	0.000156	$4.30 \cdot 10^{-6}$	0.1139	145.51
Quantile 3	0.000270	$2.29 \cdot 10^{-6}$	0.1214	83.58
Quantile 4	0.000245	$7.26 \cdot 10^{-7}$	0.1332	84.25

Table 4.3. Makeham parameters for womens quantiles

Men	α	β	γ	Q
Quantile 1	0.000309	$1.70 \cdot 10^{-5}$	0.1032	98.75
Quantile 2	0.000576	$1.16 \cdot 10^{-5}$	0.1065	103.80
Quantile 3	0.000461	$2.94 \cdot 10^{-6}$	0.1212	54.63
Quantile 4	0.000354	$9.36 \cdot 10^{-7}$	0.1333	46.13

Table 4.4. Makeham parameters for mens quantiles

Entire sample	α	β	γ	Q
Quantile 1	0.000391	$7.61 \cdot 10^{-6}$	0.1069	221.71
Quantile 2	0.000251	$6.66 \cdot 10^{-6}$	0.1090	194.00
Quantile 3	0.000268	$2.97 \cdot 10^{-6}$	0.1194	68.57
Quantile 4	0.000246	$9.91 \cdot 10^{-7}$	0.1314	74.34

Table 4.5. Makeham parameters for the entire samples quantiles

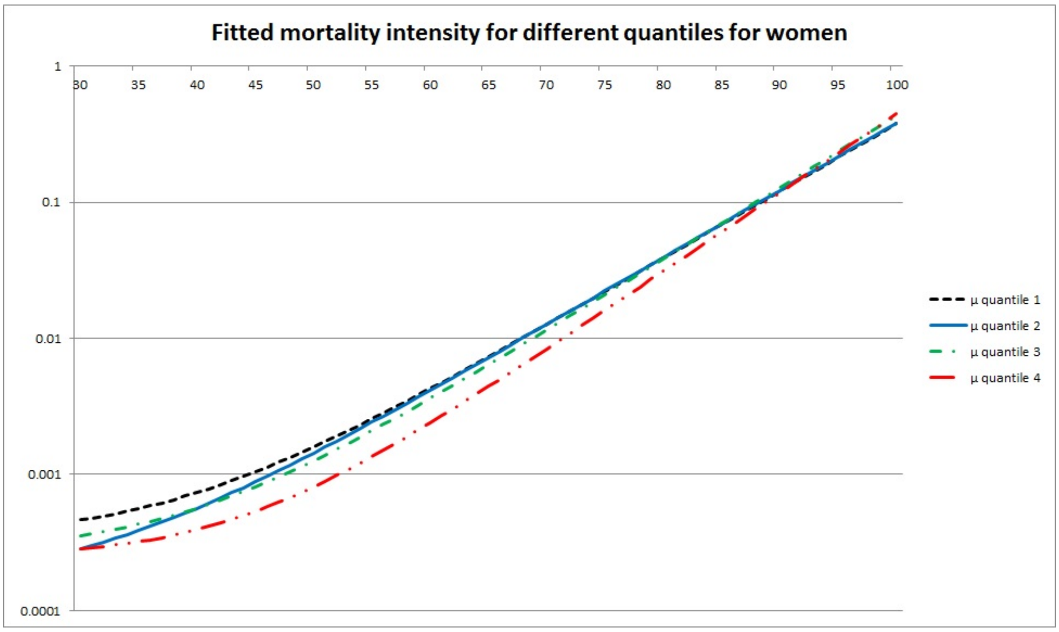

Figure 4.3. Fitted mortality intensities for different quantiles for women

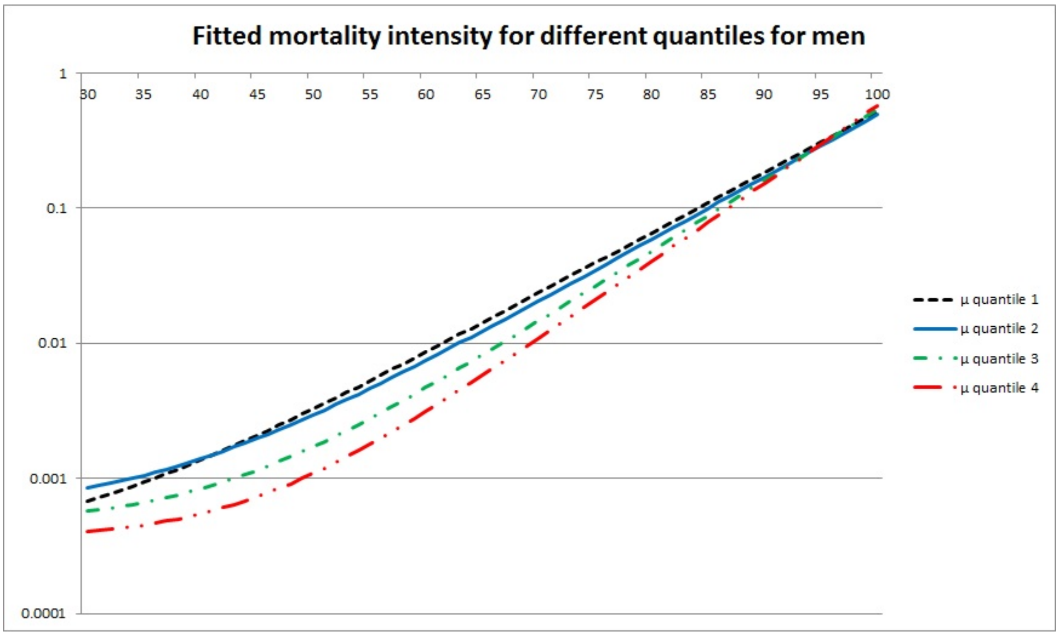

Figure 4.4. Fitted mortality intensities for different quantiles for men

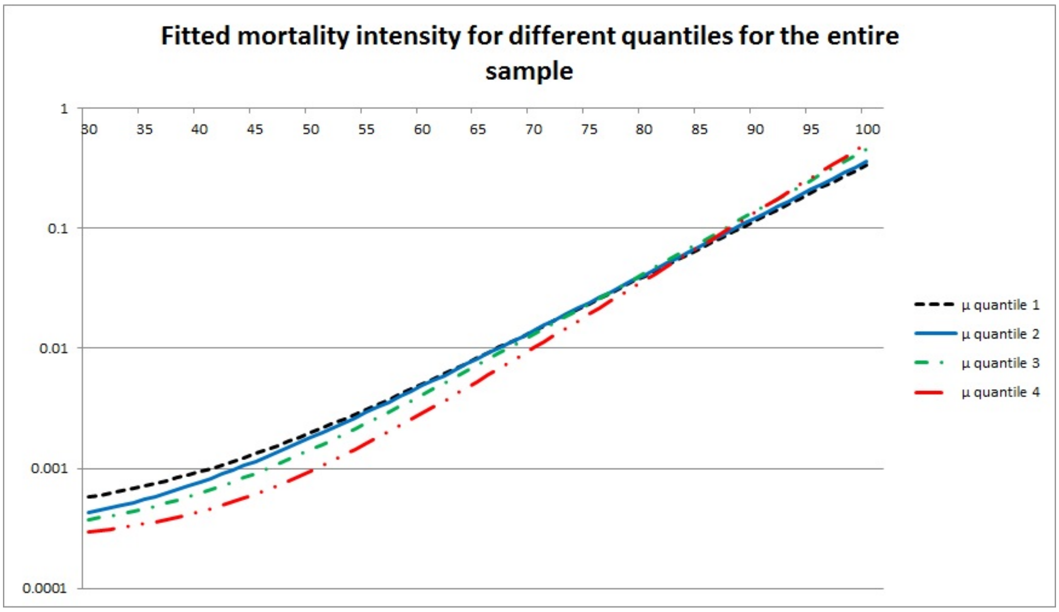

Figure 4.5. Fitted mortality intensities for different quantiles for the total population

What we previously discussed and concluded from the tables of the Makeham parameters of the four quantiles for the different sets, we can also see in the graphs shown in figures 4.3 to 4.5. For the women, the first and second quantile are wide apart, showing how much the α-parameter matter in low ages. In all three sets, both the third and fourth quantile have deeper and steeper shapes of the curves, especially the fourth. For women the curves intersect around the age of 92. For men, the effect is even more obvious, and the intersection point is not until the age if 96. This is because of the smaller β-parameter and the higher γ-parameter of the higher quantiles.

In the figures 4.6 to 4.8, the observed mortality intensities are plotted against the fitted Makeham curves of the mortality intensities for the four quantiles. There are some volatility in the lower ages in all twelve graphs, and especially in the fourth quantiles, indicating few deaths below 40. In some of the curves we notice notches, representing an age group of a quantile lacking any deceased. Remember that we are using log-scale and that it doesn't allow for an outcome of zero.

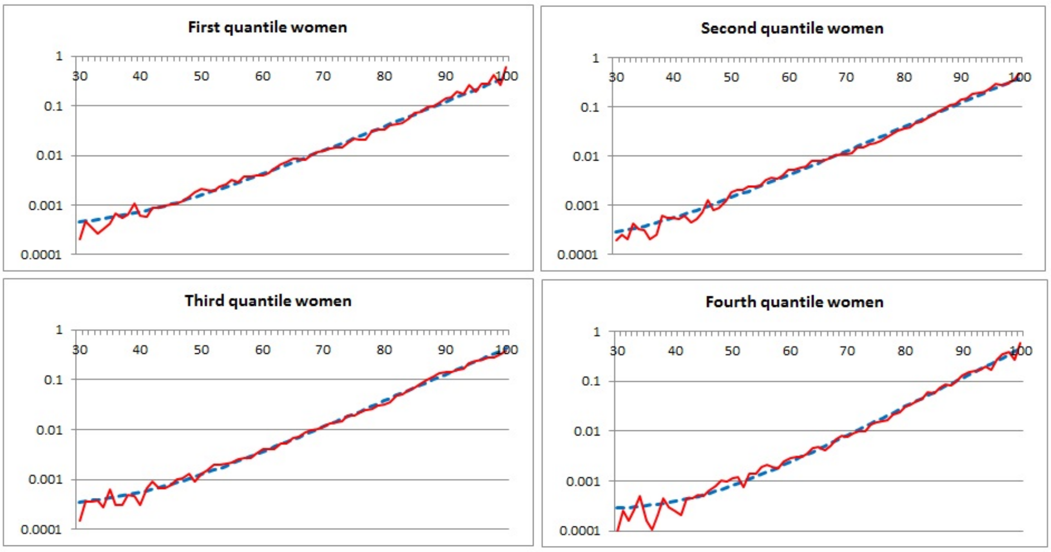

Figure 4.6. Fitted against observed mortality intensities for different quantiles for women

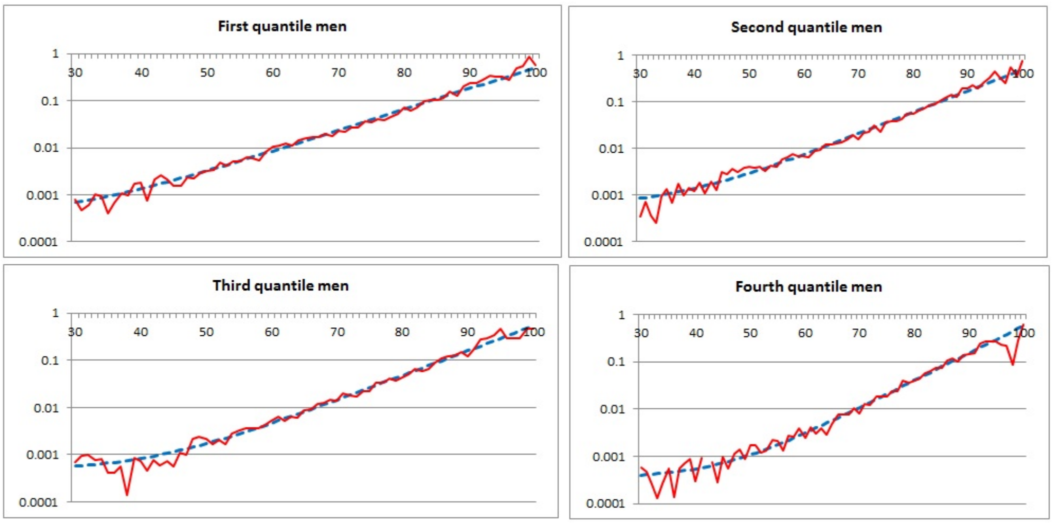

Figure 4.7. Fitted against observed mortality intensities for different quantiles for men

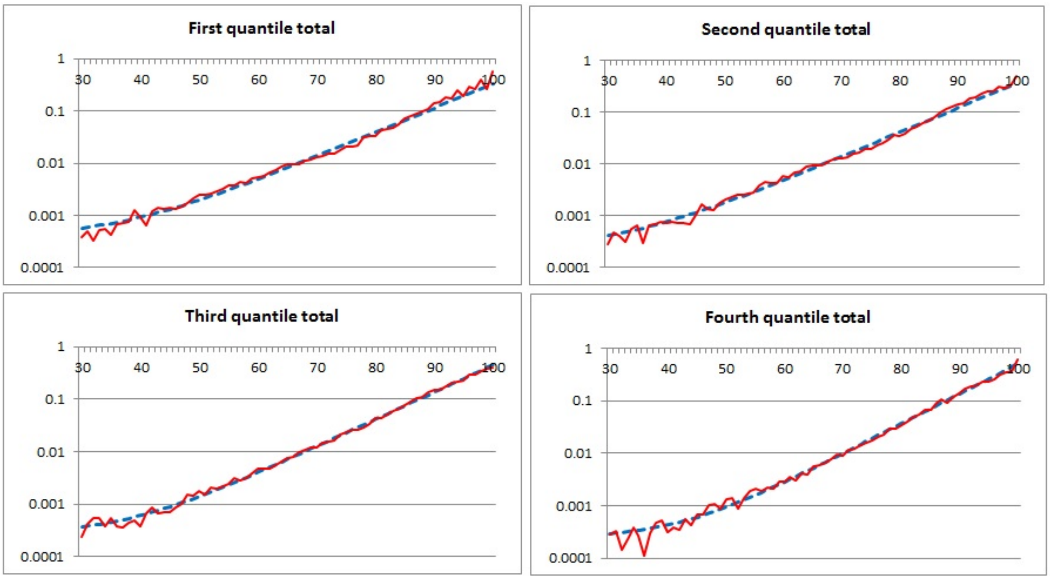

Figure 4.8. Fitted against observed mortality intensities for different quantiles for the total sample

Another way to interpret the Makeham parameters are to use them for calculating the expected remaining lifetime at different ages. Here the approximative formula of e_x is used, using equation (2.7). The result is shown in the figure 4.9.

The lowest curve show the remaining expected lifetime at birth, and the other three curves show the expected lifetime conditioned on being alive at a certain age, that is

$$e_x^* = E[T_x|x] + x$$

As can be seen in the three different sets, there is a rather large gap between the fourth quantile and the other three. The gap is very distinct looking at the expected remaining lifetime at birth, but becomes less apparent as we begin conditioning on older ages. For women, the fourth quantile have more than two and a half year longer expected lifetime at birth than the lowest quantile and for men the difference is almost six years. Another remark is that the first and second quantile for women show 0.75 years difference in remaining expected lifetime at birth, but less than a half year later at the ages 30, 50 and 65, indicating how the α-parameter mostly makes an impact at younger years. We can also remark that the older the conditioning age, the greater the effect of the γ-parameter becomes clear.

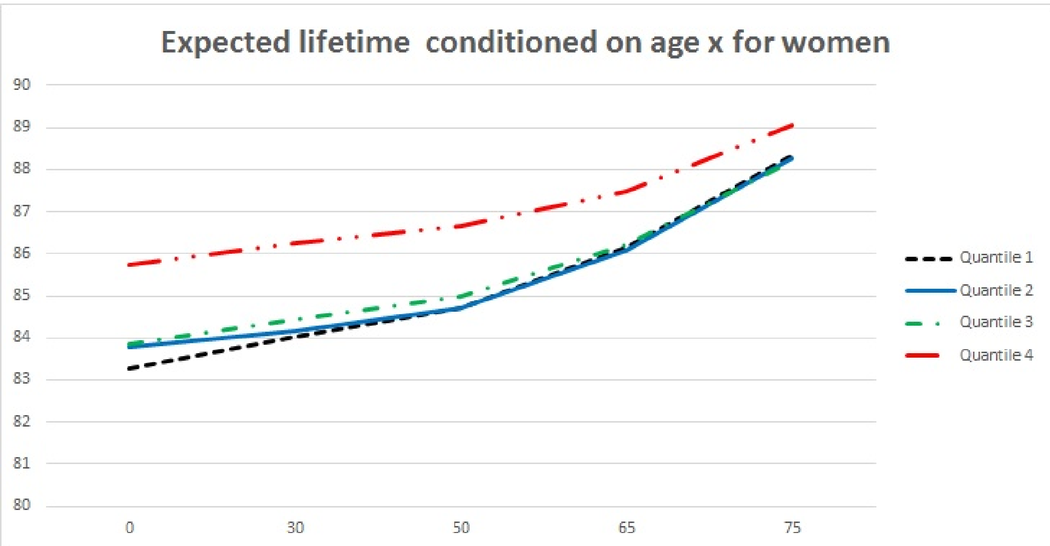

Figure 4.9. Expected remaining lifetime for different quantiles and ages for women

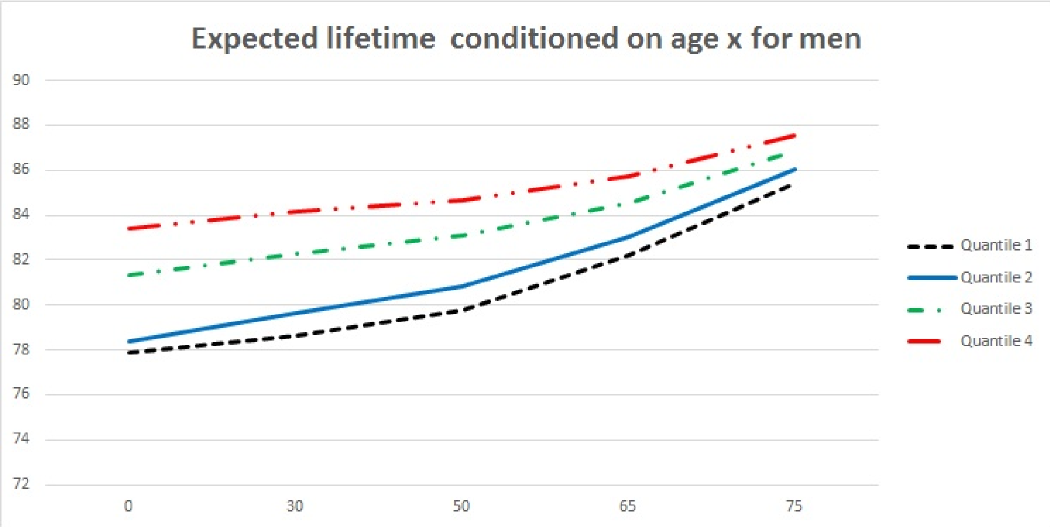

Figure 4.10. Expected remaining lifetime for different quantiles and ages for men

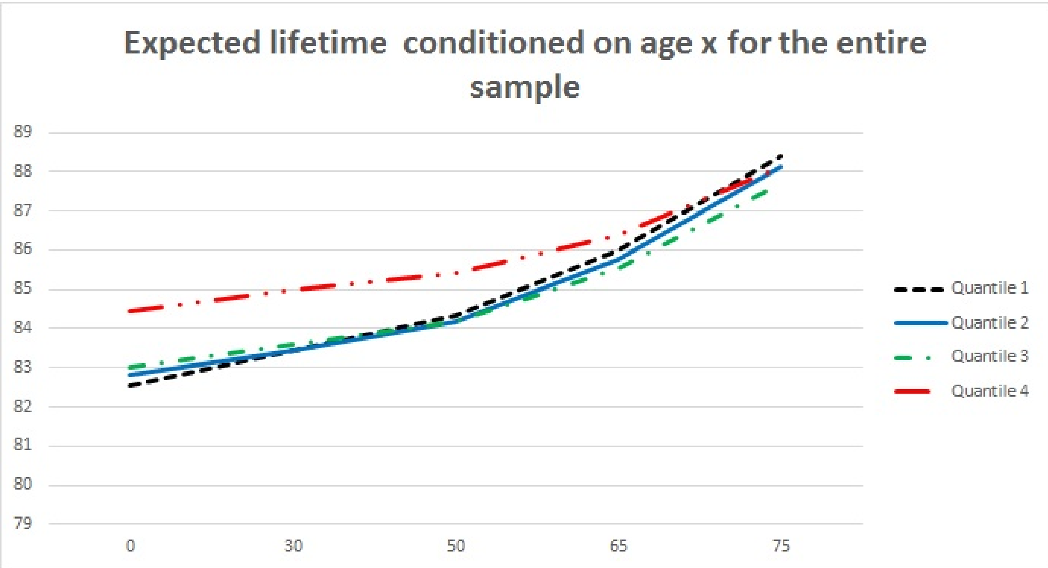

Figure 4.11. Expected remaining lifetime for different quantiles and ages for women

4.2 Non-parametric bootstrapping

To estimate the accuracy of the remaining life expectancy as well as the Makeham parameters in the tables 4.3, 4.4 and 4.5, we use the non-parametric bootstrap method explained in chapter 2.5. Bootstrapping is a practice used to measure and determine the properties of a set when sampling from an approximating distribution. For this we divide the population into groups by age, gender and quantile. We put the original number of survivors and deceased of each of these groups in different boxes. We then draw with replacement the sum of survivors and deceased for each group a thousand times, thus giving us 1000 new data samples for each quantile and gender. These samples can further be used to calculate new sets of Makeham parameters. For each set of parameters, estimated remaining lifetime can then be calculated. The α, β and γ parameters for the different sets are given in the figures 4.12 to 4.20. We'll talk to each group of parameters below.

4.2.1 Bootstrapped Makeham parameters

The alpha parameter

First of all, we will consider the α-parameter. Remember that the α-parameter is age independent and gives the mortality at age 0. It has a large importance to the mortality at younger ages, but will play an insignificant role at older ages. There are a few things to be said. First of all, we only see some less than obvious tendencies that the different quantiles display a similar behavior for the three populations. The fourth quantiles generally have a lower value. Secondly, for the men in the first quantile, some simulations have a negative α-value. To make things a bit easier while doing the bootstrap method, the condition that the α-parameter had to be greater than 0 was removed. As we looked at a least squares approximation that ranged from the age of 30 to 100, it's is not a strange thing, even though impossible in real life. All the parameters have an expectancy lower than that of M90.

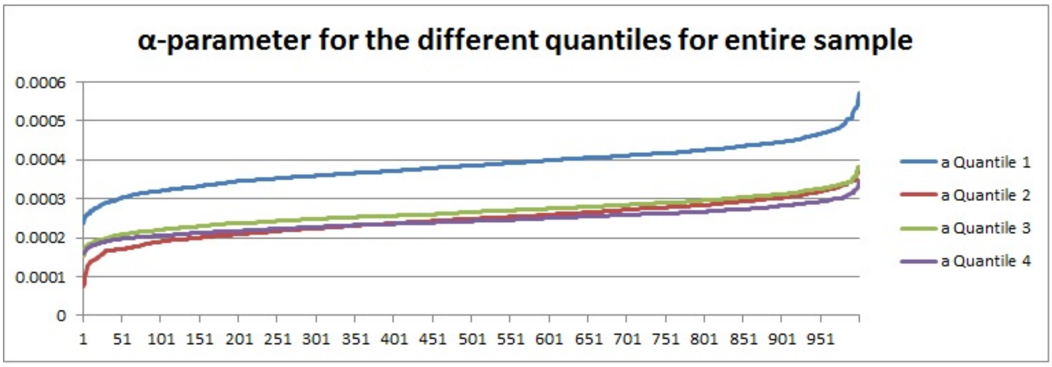

Figure 4.12. The Makeham α-parameters after Bootstrapping the original sample 1000 times

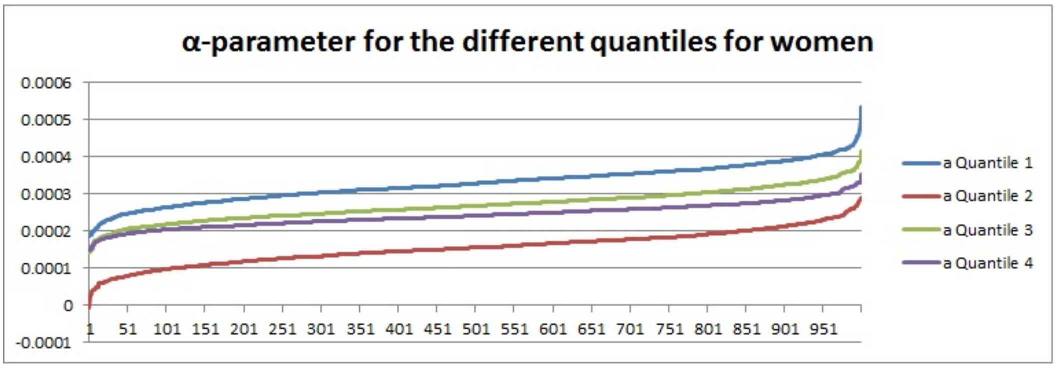

Figure 4.13. The Makeham α-parameters after Bootstrapping the womens sample 1000 times

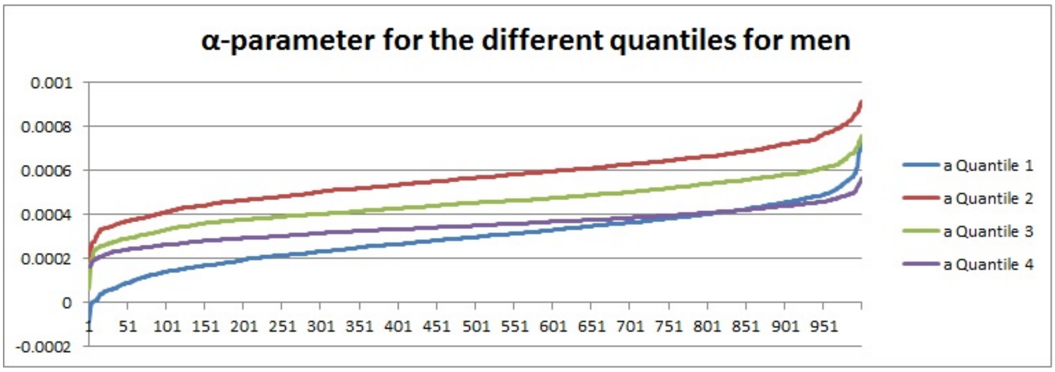

Figure 4.14. The Makeham α-parameters after Bootstrapping the mens sample 1000 times

		μ	σ
Total	Quantile 1	0.000385	$5.02 \cdot 10^{-5}$
	Quantile 2	0.000247	$4.49 \cdot 10^{-5}$
	Quantile 3	0.000266	$3.55 \cdot 10^{-5}$
	Quantile 4	0.000243	$2.94 \cdot 10^{-5}$
Women	Quantile 1	0.000328	$4.88 \cdot 10^{-5}$
	Quantile 2	0.000156	$4.48 \cdot 10^{-5}$
	Quantile 3	0.000270	$4.12 \cdot 10^{-5}$
	Quantile 4	0.000243	$3.15 \cdot 10^{-5}$
Men	Quantile 1	0.000297	$1.23 \cdot 10^{-4}$
	Quantile 2	0.000564	$1.18 \cdot 10^{-4}$
	Quantile 3	0.000453	$9.65 \cdot 10^{-5}$
	Quantile 4	0.000350	$6.70 \cdot 10^{-5}$

Table 4.6. The expected value and standard deviation of the α-parameters

The beta parameter

Secondly, let us examine the β-parameter. Here we see a trend in all three populations, where the lower quantiles experience a much higher β. For men, the expected value of the parameter for the first quantile is more than ten times as high as the fourth. We can also observe that the lower quantiles have a larger spread, at the same time as the fourth quantile show comparably a very small variance.

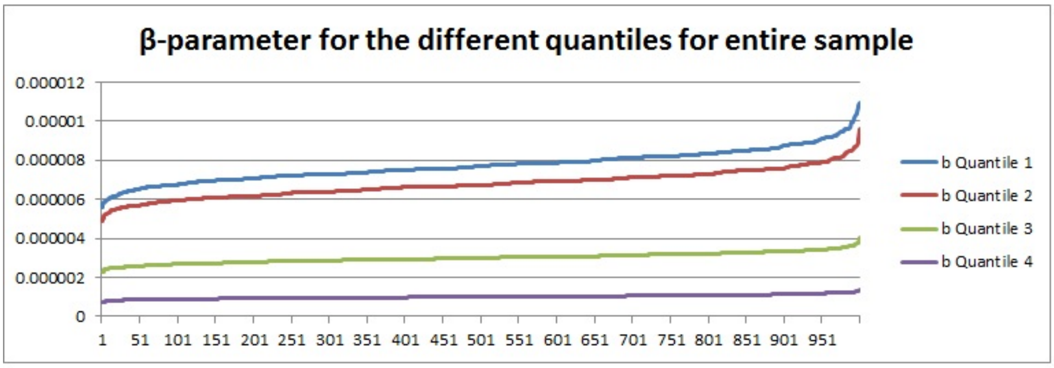

Figure 4.15. The Makeham β-parameters after Bootstrapping the original sample 1000 times

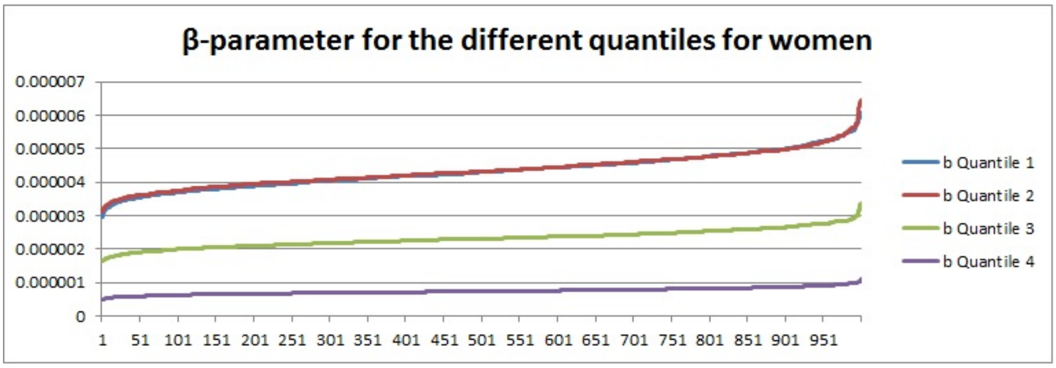

Figure 4.16. The Makeham β-parameters after Bootstrapping the womens sample 1000 times

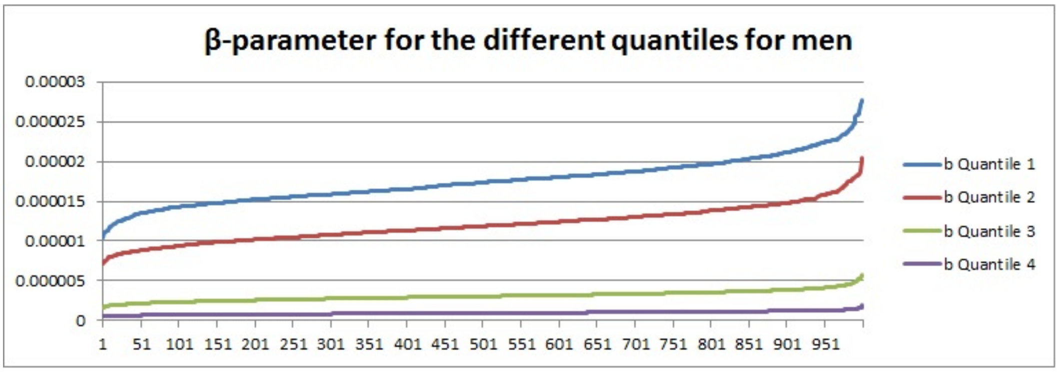

Figure 4.17. The Makeham β-parameters after Bootstrapping the mens sample 1000 times

		μ	σ
Total	Quantile 1	$7.73 \cdot 10^{-6}$	$7.86 \cdot 10^{-7}$
	Quantile 2	$6.77 \cdot 10^{-6}$	$6.72 \cdot 10^{-7}$
	Quantile 3	$3.00 \cdot 10^{-6}$	$2.55 \cdot 10^{-7}$
	Quantile 4	$1.01 \cdot 10^{-6}$	$9.69 \cdot 10^{-8}$
Women	Quantile 1	$4.33 \cdot 10^{-6}$	$5.09 \cdot 10^{-7}$
	Quantile 2	$4.36 \cdot 10^{-6}$	$4.92 \cdot 10^{-7}$
	Quantile 3	$2.32 \cdot 10^{-6}$	$2.59 \cdot 10^{-7}$
	Quantile 4	$7.45 \cdot 10^{-7}$	$9.46 \cdot 10^{-8}$
Men	Quantile 1	$1.75 \cdot 10^{-5}$	$2.75 \cdot 10^{-6}$
	Quantile 2	$1.20 \cdot 10^{-5}$	$2.14 \cdot 10^{-6}$
	Quantile 3	$3.10 \cdot 10^{-6}$	$6.12 \cdot 10^{-7}$
	Quantile 4	$9.68 \cdot 10^{-7}$	$1.93 \cdot 10^{-7}$

Table 4.7. The expected value and standard deviation of the β-parameters

The gamma parameter

For the last Makehamparameter, γ, we can further see as clear trend as for the β-parameter, though introverted.

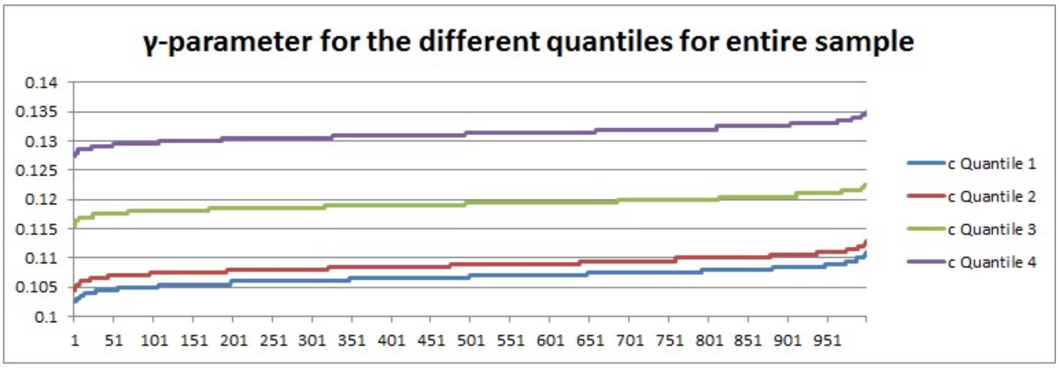

Figure 4.18. The Makeham γ-parameter after Bootstrapping the original sample 1000 times

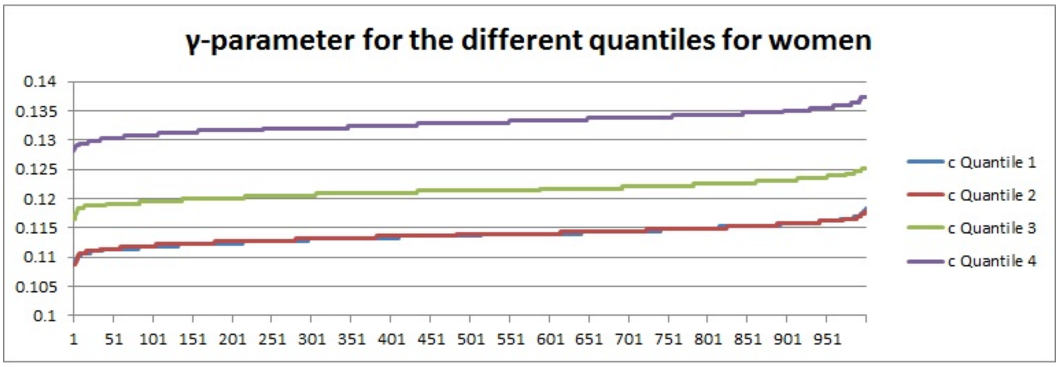

Figure 4.19. The Makeham γ-parameter after Bootstrapping the original sample 1000 times

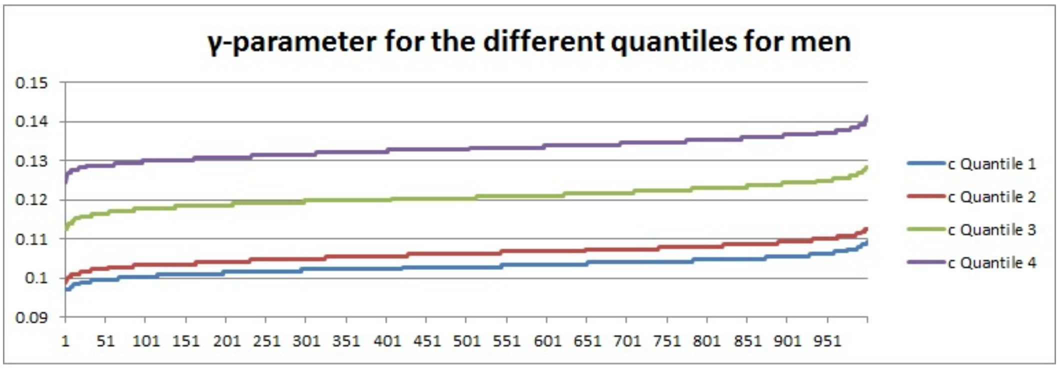

Figure 4.20. The Makeham γ-parameter after Bootstrapping the original sample 1000 times

		μ	σ
Total	Quantile 1	0.1068	0.001278
	Quantile 2	0.1088	0.001251
	Quantile 3	0.1193	0.001063
	Quantile 4	0.1313	0.001183
Women	Quantile 1	0.1137	0.001464
	Quantile 2	0.1138	0.001401
	Quantile 3	0.1213	0.001385
	Quantile 4	0.1330	0.001550
Men	Quantile 1	0.1029	0.002029
	Quantile 2	0.1062	0.002308
	Quantile 3	0.1207	0.002516
	Quantile 4	0.1331	0.002537

Table 4.8. The expected value and standard deviation of the γ-parameters

4.2.2 Scatter plot of the Makeham parameters

An interesting thing to consider would be how the three Makeham parameters interact with one another. In the figures 4.21, 4.22 and 4.23 we can see scatter plots for α/β, α/γ and β/γ interact respectively. These scatter plots are done on the different quantiles of the total sample, but the scatter plots for the men and women looks very similar in shape, if not in numbers. As can be observed, there are some shapes and trends to consider. First thing we can examine is the very defined β/γ trend in figure 4.23. As we have remarked before, the β- and γ-parameters have a negative correlation, which becomes very obvious in the figure. The correlation of nearly negative one is shown in table 4.9. It is not however a surprising finding, given how the parameters interact. In the scatter plots 4.22 and 4.23 we can also see the interval of 0.0005 when fixating γ when minimizing Q. Furthermore, there is also a negative correlation between α and β. Seeing how it is mainly these two parameters that coexist in explaining the deaths of younger and middle age groups, this comes as no big revelation. Trivially we have a positive correlation between the last two parameters α and γ, which can be deducted from above.

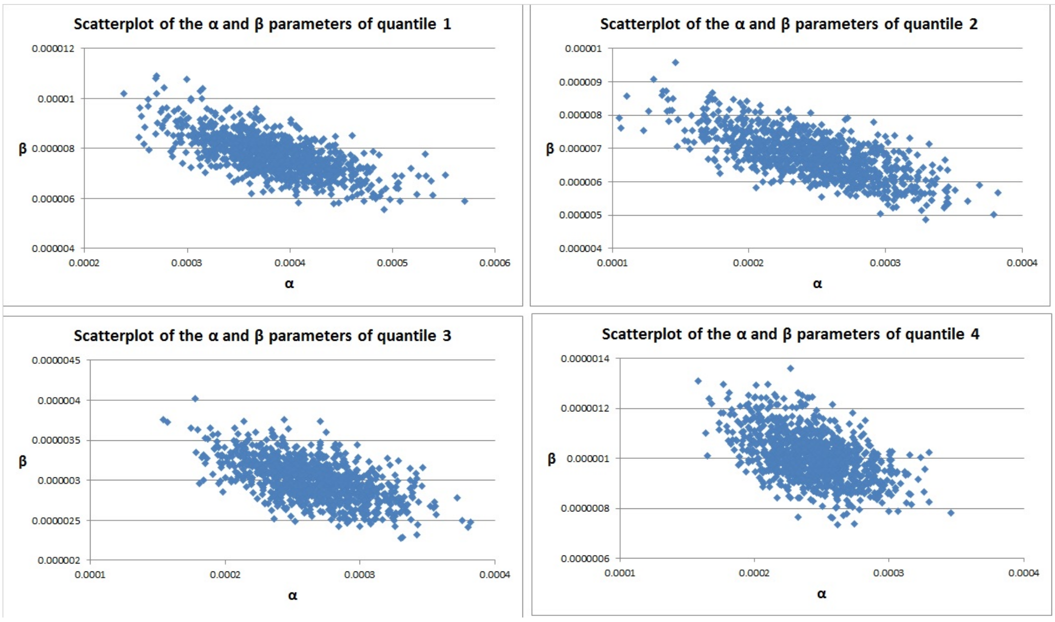

Figure 4.21. Scatter plot of the β- and α-parameters after Bootstrapping the quantiles of the total sample 1000 times

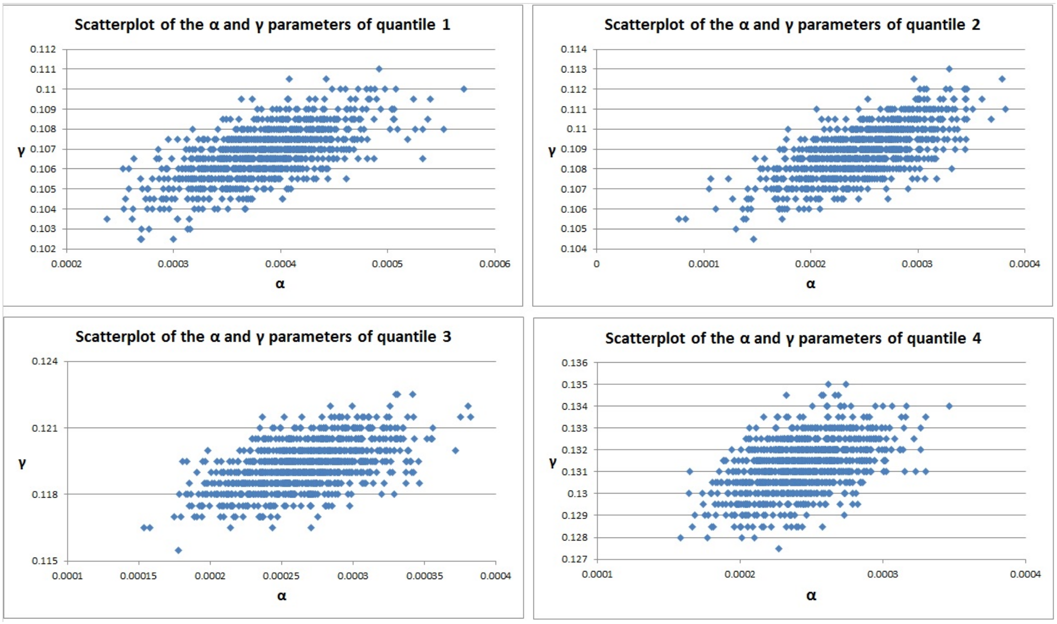

Figure 4.22. Scatter plot of the γ- and α-parameters after Bootstrapping the quantiles of the total sample 1000 times

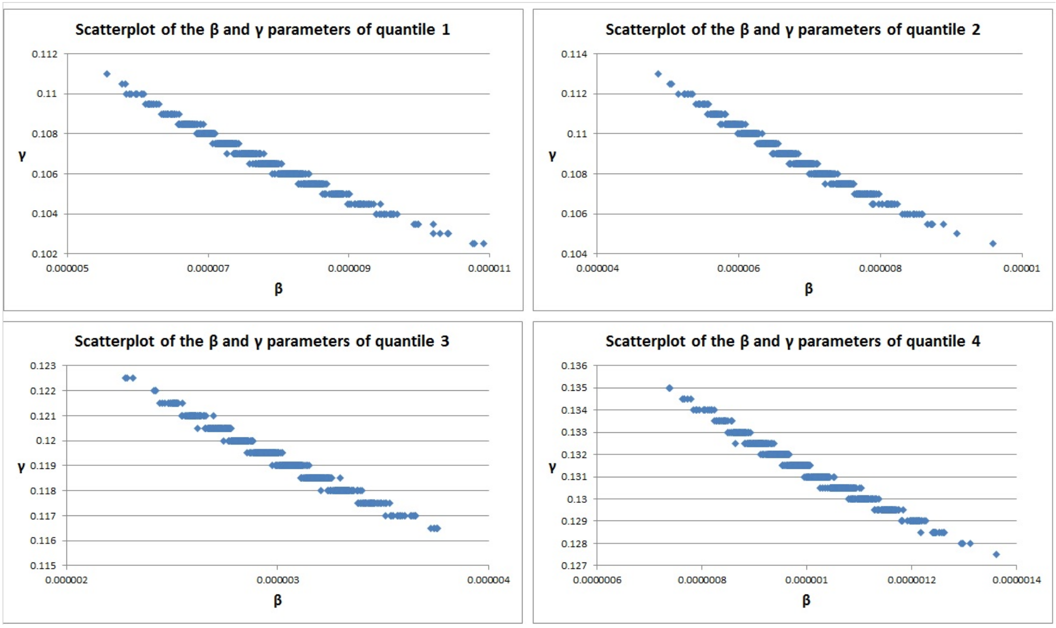

Figure 4.23. Scatter plot of the γ- and β-parameters after Bootstrapping the quantiles of the total sample 1000 times

Correlation	α/β	α/γ	β/γ
Quantile 1	-0.6368	0.6188	-0.9913
Quantile 2	-0.6299	0.6130	-0.9912
Quantile 3	-0.6046	0.5825	-0.9908
Quantile 4	-0.4964	0.4812	-0.9909

Table 4.9. The correlation between the Makeham parameters

4.2.3 Remaining life expectancy

As we have stated above, we can calculate remaining life expectancy from the 12 000 sets of Makeham parameters that we have been looking at. We are going to review the remaining life expectancy at birth, conditioned you have achieved the age of 30 and the age of 65.

The outcome is shown in tables 4.10 to 4.12 as well as the figures 4.24 to 4.32.

Total	At birth		At age 30		At age 65	
	μ	σ	μ	σ	μ	σ
1	82.56	0.1535	83.45	0.0958	86.02	0.0868
2	82.84	0.1379	83.44	0.0923	85.78	0.0841
3	83.01	0.1211	83.61	0.0857	85.54	0.0752
4	84.46	0.1160	85.00	0.0883	86.40	0.0797

Table 4.10. The expected remaining life and standard deviation of the 1000 scenarios for the total sample

Women	At birth		At age 30		At age 65	
	μ	σ	μ	σ	μ	σ
1	83.27	0.1577	84.01	0.1072	86.16	0.0949
2	83.77	0.1486	84.16	0.1022	86.10	0.0905
3	83.85	0.1387	84.45	0.0979	86.22	0.0876
4	85.73	0.1308	86.26	0.0995	87.51	0.0903

Table 4.11. The expected remaining life and standard deviation of the 1000 scenarios for the womens sample

Men	At birth		At age 30		At age 65	
	μ	σ	μ	σ	μ	σ
1	77.88	0.2912	78.65	0.1701	82.21	0.1431
2	78.43	0.3136	79.66	0.1808	83.07	0.1573
3	81.36	0.2934	82.32	0.1826	84.57	0.1668
4	83.44	0.2592	84.18	0.1918	85.75	0.1806

Table 4.12. The expected remaining life and standard deviation of the 1000 scenarios for the mens sample

At birth

The remaining life expectancy at birth differs a lot between the quantiles for all the populations and especially for the men. For the women, there is at birth a 2.5 years difference between the highest and the lowest income quantile. For men, the difference is a stunning 5.5 years. For the three populations, the lowest two or three income quantiles remain close to each other and it is with one exception only the fourth quantiles that is sticking out. For the men however, there is also the third quantile that is significantly higher than the lower two. What can also be said is that the different scenarios seems to have inherited the properties of the original sample. Furthermore, there is also quite little variation in the different samples.

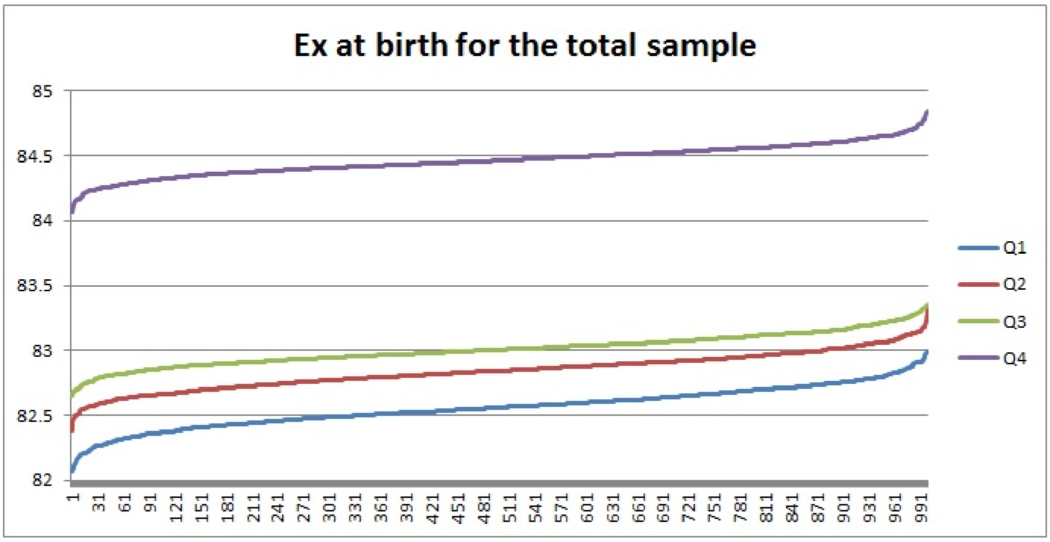

Figure 4.24. Remaining life expectancy at birth after Bootstrapping the quantiles of the total sample 1000 times

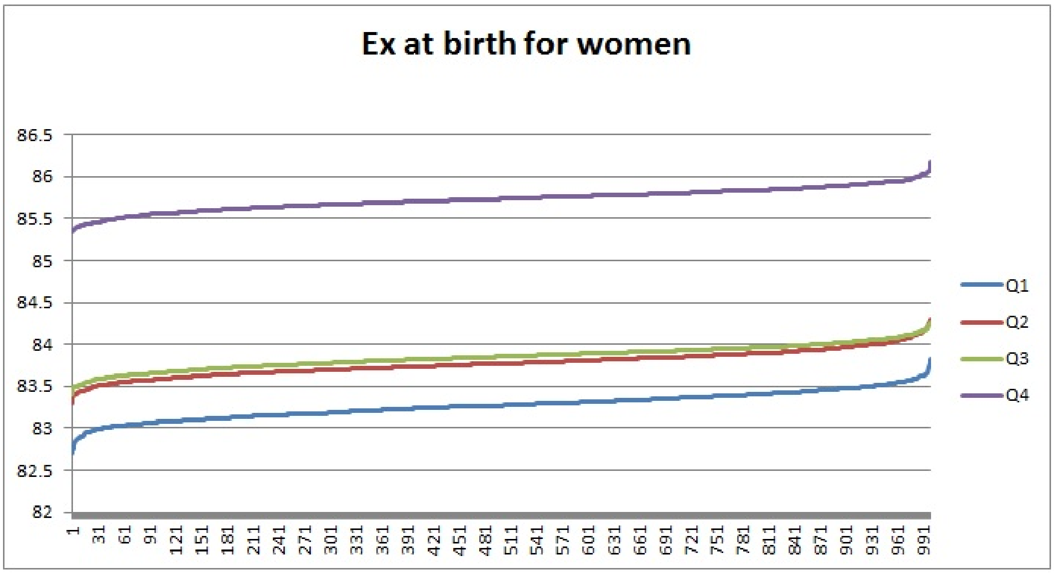

Figure 4.25. Remaining life expectancy at birth after Bootstrapping the quantiles of the womens sample 1000 times

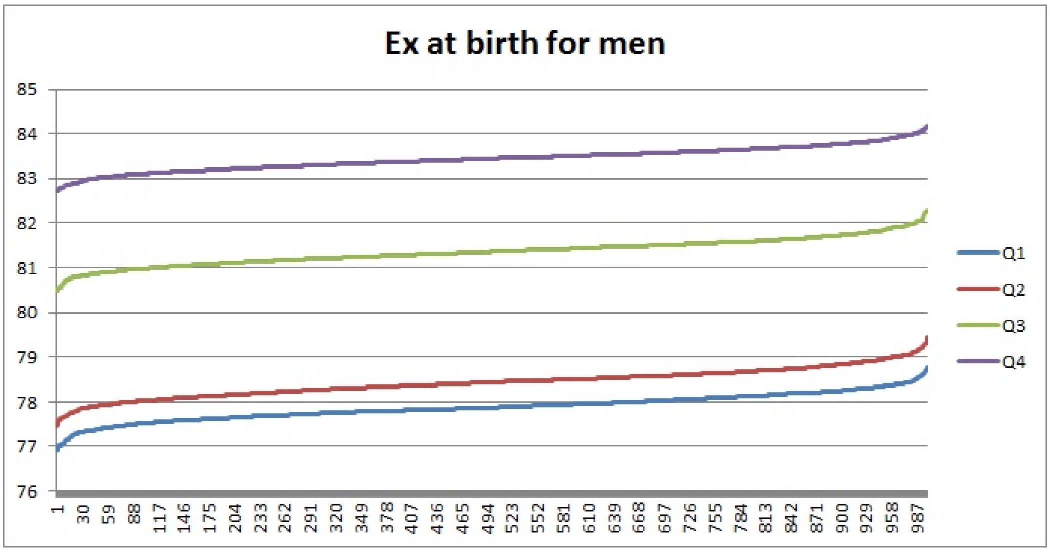

Figure 4.26. Remaining life expectancy at birth after Bootstrapping the quantiles of the mens sample 1000 times

At the age of 30

There is no discernible change between birth and the age of 30.

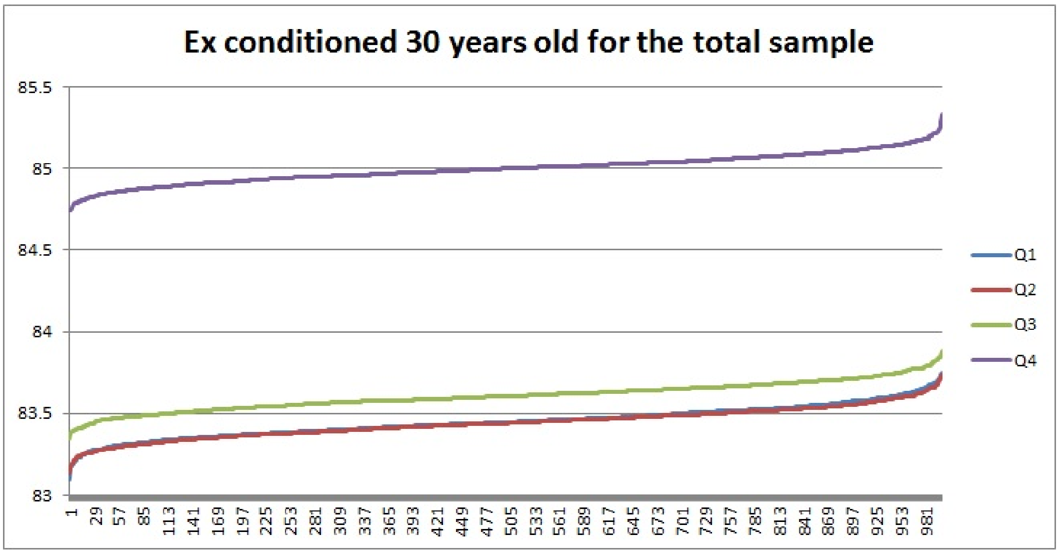

Figure 4.27. Remaining life expectancy at age 30 after Bootstrapping the quantiles of the total sample 1000 times

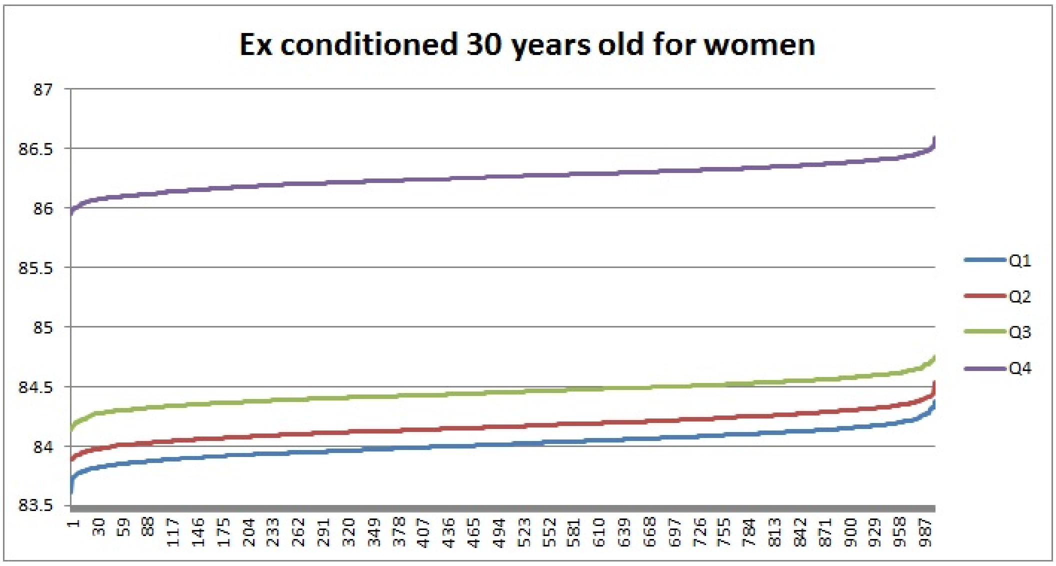

Figure 4.28. Remaining life expectancy at age 30 after Bootstrapping the quantiles of the womens sample 1000 times

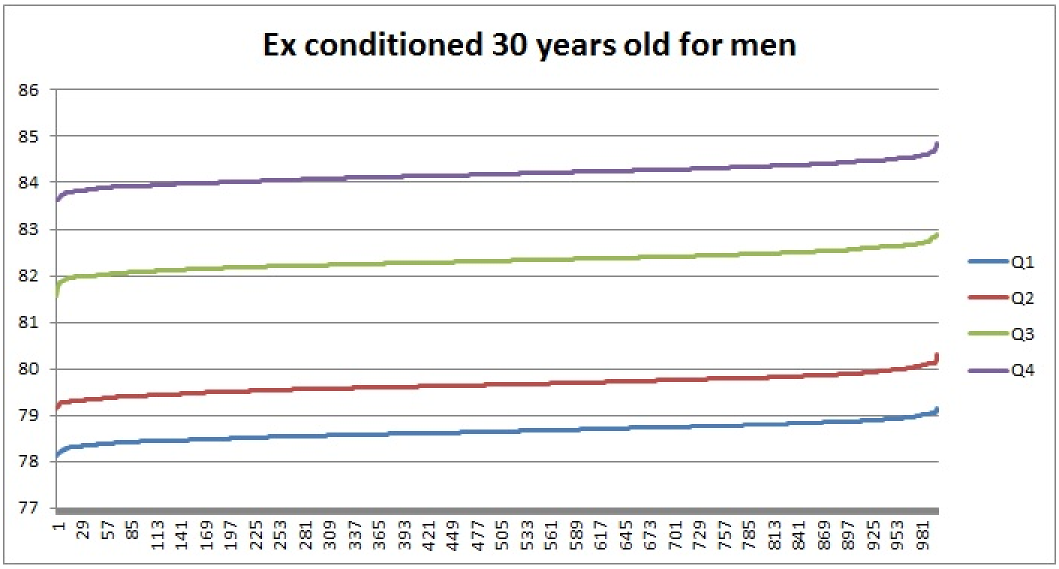

Figure 4.29. Remaining life expectancy at age 30 after Bootstrapping the quantiles of the mens sample 1000 times

At the age of 65

At the age of 65 we start seeing some change. The remaining expected lifetime has obviously increased with between two to three years, given the fact that we condition on everyone living at the age of 65. Additionally the difference in remaining expected lifetime between the quantiles have been reduced. For men you can still observe a distinct difference between all three income quantiles. For women however, the change is small difference between the three lower income groups is all but gone. The fourth quantile is still discernible though. For the total sample, there have been some change in between the quantiles, with the third quantile having the lowest remaining life expectancy.

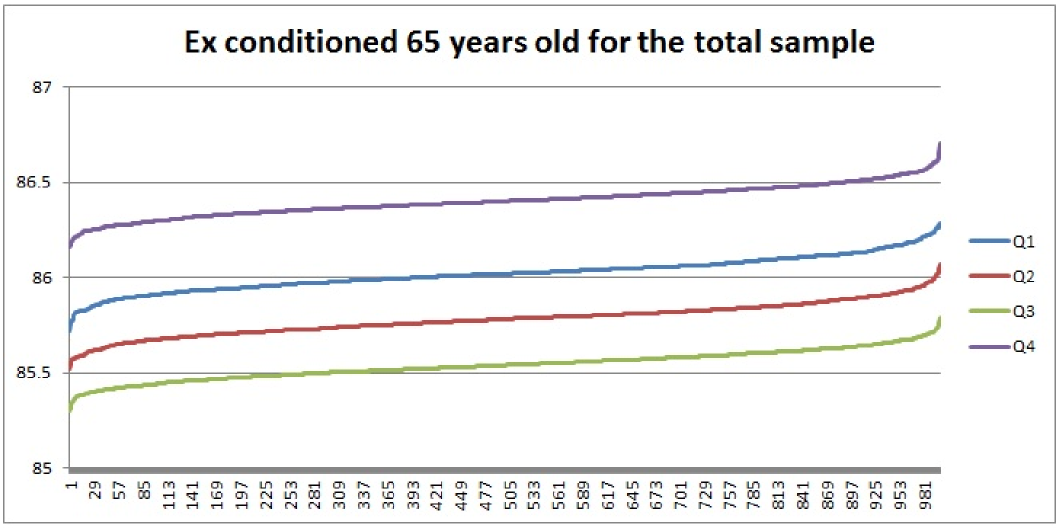

Figure 4.30. Remaining life expectancy at age 65 after Bootstrapping the quantiles of the total sample 1000 times

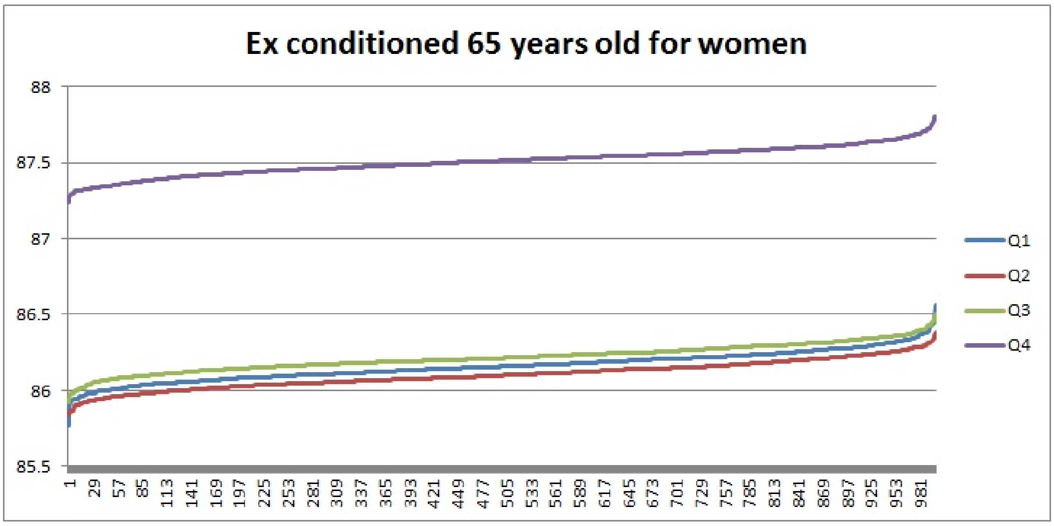

Figure 4.31. Remaining life expectancy at age 65 after Bootstrapping the quantiles of the womens sample 1000 times

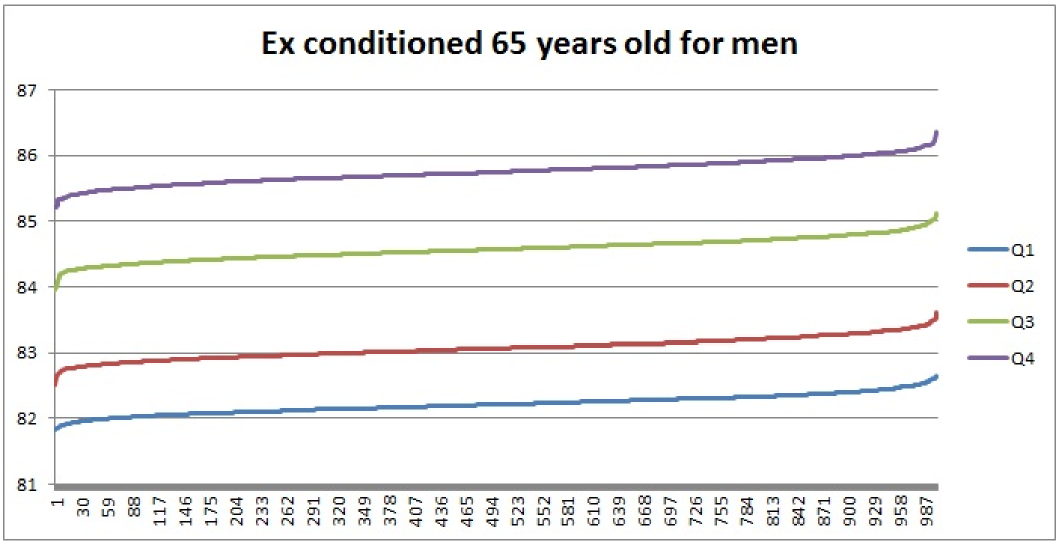

Figure 4.32. Remaining life expectancy at age 65 after Bootstrapping the quantiles of the mens sample 1000 times

4.3 Confidence intervals and histograms

So what does all this tell us about our hypothesis, that there are more parameters than just the gender that could be important to consider when doing mortality studies? As we chose to study an income parameter in this work, the hypothesis is thus that there is a significant difference in mortality between higher income groups and lower. The figure 4.33 shows the histogram of the difference between the normal distribution using the estimated mean and standard deviation calculated shown in table 4.10 above against the calculated results of the bootstrap samples for the expected remaining for the total sample.

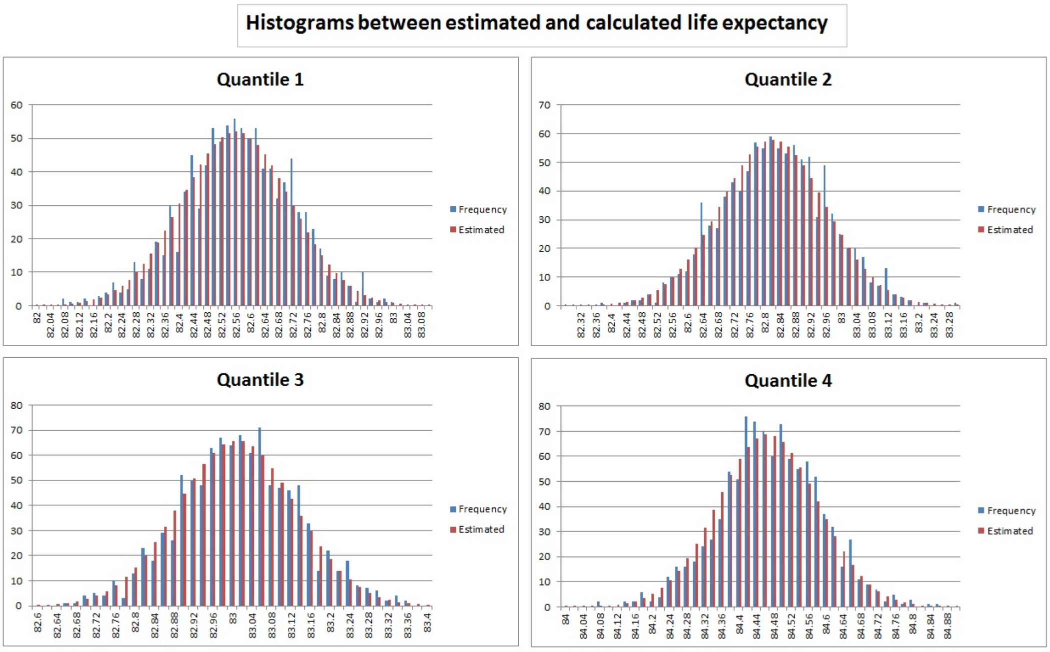

Figure 4.33. Histogram showing the difference between the normal distribution using the estimated mean and standard deviation against the calculated results of the bootstrap samples for the expected remaining for the total sample

Looking at the histograms, we seem to have a rather good fit with our normal distributions. We can then use the normal distributions to calculate confidence intervals for our quantiles, thus testing our hypothesis. Confidence intervals can be used to express the degree of uncertainty associated with a sample statistic. It is a an interval estimate and consist of a range of values that act as good estimates of the unknown parameter. In this case the mortality of the different quantiles of income groups in our set of populations. It should be remembered that the true value of the parameter is not necessarily in the computed interval of a particular sample. Considering that we are using a non parametric bootstrap on observed data that are random samples of the true population, this signifies that the confidence interval must also be random. A hypothesis test is performed with a certain level of significance, which corresponds to the confidence level. In our case, the confidence level of 0.01. To calculate the upper and lower bound of our data, we use the formula

$$\bar{x} \pm 2.576 \cdot \frac{\sigma}{\sqrt{n}} \tag{4.1}$$

where 2.576 is the z-value of the normal distribution at 0.99.

Total	Quantile	Lower limit	Mean	Upper limit
$e_{x,0}$	Quantile 1	82.16	82.56	82.95
	Quantile 2	82.48	82.84	83.20
	Quantile 3	82.70	83.01	83.32
	Quantile 4	84.17	84.46	84.76
$e_{x,30}$	Quantile 1	83.20	83.45	83.70
	Quantile 2	83.20	83.44	83.68
	Quantile 3	83.39	83.61	83.83
	Quantile 4	84.77	85.00	85.22
$e_{x,65}$	Quantile 1	85.80	86.02	86.24
	Quantile 2	85.56	85.78	86.00
	Quantile 3	85.35	85.54	86.20
	Quantile 4	86.20	86.40	86.61

Table 4.13. The confidence intervals for the expected lifetimes at birth, the age of 30 and age of 65 for the total sample

Women	Quantile	Lower limit	Mean	Upper limit
$e_{x,0}$	Quantile 1	82.86	83.27	83.68
	Quantile 2	83.39	83.77	84.16
	Quantile 3	83.49	83.85	84.20
	Quantile 4	85.39	85.73	86.06
$e_{x,30}$	Quantile 1	83.74	84.01	84.29
	Quantile 2	83.90	84.16	84.43
	Quantile 3	84.20	84.45	84.70
	Quantile 4	86.00	86.26	86.52
$e_{x,65}$	Quantile 1	85.92	86.16	86.40
	Quantile 2	85.87	86.10	86.34
	Quantile 3	85.99	86.22	86.44
	Quantile 4	86.28	87.51	87.74

Table 4.14. The confidence intervals for the expected lifetimes at birth, the age of 30 and age of 65 for the womens sample

Men	**Quantile**	**Lower limit**	**Mean**	**Upper limit**
$e_{x,0}$	Quantile 1	77.13	77.88	78.63
	Quantile 2	77.62	78.43	79.24
	Quantile 3	80.60	81.36	82.11
	Quantile 4	82.77	83.44	84.11
$e_{x,30}$	Quantile 1	78.21	78.65	79.08
	Quantile 2	79.20	79.66	80.13
	Quantile 3	81.85	82.32	82.79
	Quantile 4	83.69	84.18	84.68
$e_{x,65}$	Quantile 1	81.84	82.21	82.58
	Quantile 2	82.66	83.07	83.47
	Quantile 3	84.14	84.57	85.00
	Quantile 4	85.28	85.75	86.21

Table 4.15. The confidence intervals for the expected lifetimes at birth, the age of 30 and age of 65 for the mens sample

From the tables 4.13, 4.14 and 4.15 we can determine that the remaining expected lifetime at birth for all the three populations is lower for first than the fourth quantile, with a confidence level of 1 %. The same is true when we condition the remaining expected lifetime on the age of 30. For the condition of the age of 65, we can say that the is a statistical difference between the means, even though the confidence intervals overlap. We can see an example of this looking at figure 4.34

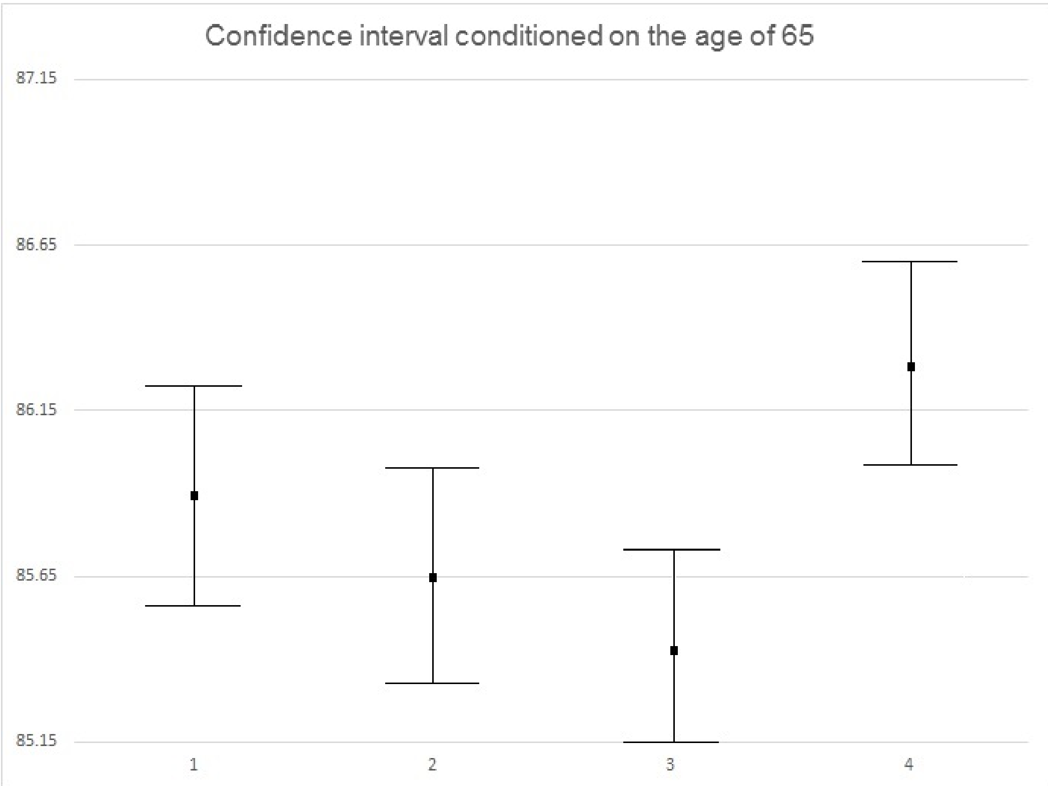

Figure 4.34. Confidence intervals for the four income quantiles of the total sample conditioned on the age of 65

Chapter 5

Conclusion and discussion

5.1 Conclusion

Starting this project, there was one main aim to answer, to look into the assumption that more parameters than just gender and age are important when doing mortality studies. I set up a hypothesis that having a higher income would affect your expected life span positively, which would have an effect for life insurance companies. Based on the results from the previous chapter, we can establish that there is a statistically significant difference in mortality between the population having an income in the lowest 25 % versus the population having an income in the top 25 %. For men at birth, the difference is almost six years, while for women the same number is 2.5 years. The relative differences lessens with rising age. Though still by the retirement age, the means of the highest quantile are still significantly higher.

5.2 Discussion

We have previously discussed how lifestyle and sickness have implication on the mortality. In this study we looked at an income parameter to see how it affects the lifespan. From the results, we now know that there is a significant discrepancy between higher and lower income groups. Higher income gives the possibility of private health insurance and stable living. Income correlates with a variety of demographical data, among others education, living and employment. The answer to why higher income leads to lower mortality probably lies within these subcategories. So how does this affect insurance companies? First of all, insurance companies uses economical mortality, which takes into account the sum at risk when weighting the Makeham parameters and calculating the commutation numbers. The sum at risk is based on the benefit owed by the company, which in turn is based on the income. The mortality assumptions is thus based on the economic risk for insurance companies. The consequence is that the weights used are based heavily on the richest, which overall increases the life span of the economic population. Consecutively the insurance companies doesn't stand much risk because of the lower mortality for the richest. However the people with the lowest income will have their pension paid out during a longer period, lowering their pension as to what could have been.

The second aim of the thesis was to analyse the spread that exist within a certain group and to understand how low the mortality could actually get within a population. Here we can examine table 5.1 to see that the difference for men is actually higher than that between the genders. The difference for the women is lower, but not notwithstanding. The longer average life expectancy of today and the nearby future derives from people living longer after retirement. At the same time evidence suggest that at the latest stages of life, the mortality remains the same as before. By analysing the spread that exist within a certain group, it would be possible to see how low the mortality curve could get, with the medicine and health care that exist today.

	Difference men	**Difference women**	**Difference gender**
$e_{x,0}$	5.6	2.5	3.9
$e_{x,30}$	5.5	2.2	3.5
$e_{x,65}$	3.5	1.3	2.6

Table 5.1. The difference in expected lifetime between the highest and lowest quantiles for men and for women and the difference between the genders

One thing to consider is the effect of using three years of data and the assumption that they are independent from each other. We know that the income will vary from year to year, which foremost should affect the salary of the deceased, as a deceased in 2009 had a lower salary than a deceased 2011. This effect should be little, as the yearly based amount have been used, which takes into account the average increased income in Sweden. As long as this population has an income increase according to the rest of the population, this varying income from year to year should be considered a minor issue. Another thing to notice in the results is that the three lower quantiles often coincided. One reason for this could be that the incomes between the different quantiles didn't really differ much. Remember table 4.2. For women, there is only 75 000 SEK between the upper limit of the first quantile and the lower limit of the third quantile. One way to have done it would be to divide the income in quantiles, and looking at the mortality in each of these quantiles instead. The problem here would be how to divide the income, and that there would probably be very little data in the later quantiles, as there would be a much lower populations as well as fewer deaths within this population. In the results there were a much larger gap in remaining estimated lifetime for men than women. We know for a fact that the data available for men is much lower than that of women, with less than a third of the population of women. This of course give a larger uncertainty in the results. But assuming the data is correct, it could have consequences. This was not really seen in the results, with the sigma for men only being slightly higher then that of women. The use of the bootstrap model on every age in the quantile groups might be a reason for this, as for every makeham parameter set, we use the bootstrap 71 times, limiting the variability.

One discussion that was going on early in the process was whether to use the Makeham model or the Lee Carter model in this thesis. It was decided to use the Makeham model for two reasons. The first reason was that neither Folksam nor KPA was using the Lee Carter model at the time, nor most other insurance companies. The second was that the Lee Carter model is based on many years of data to be able to do the trend analysis. As the data gathering for the years earlier than 2009 was difficult for different reasons, it would have been difficult to use the Lee Carter model.

Bibliography

[1] G. Andersson. (2005). Livförsäkringsmatematik. Stockholm: Svenska Försäkringsföreningen

[2] Försäkringstekniska forskningsnämnden; Sveriges Försäkringsförbund. (2007). Försäkrade i Sverige - dödlighet och livslängder, Prognoser 2007 - 2050. Stockholm: Svenska Försäkringsföreningen.

[3] Försäkringstekniska forskningsnämnden; Sveriges Försäkringsförbund. (2007). Försäkrade i Sverige - Livslängder och dödlighet, prognoser 2014 - 2070. Stockholm: Svenska Försäkringsföreningen.

[4] Ö. Hemström and L. Lundkvist. (2011). Livslängden i Sverige 2001 - 2010. Örebro: Statistiska Centralbyrån.

[5] J. Beyer, N. Keiding, W. Simonsen. (1976). The exact behavior of the maximum likelihood estimator in the pure birth process and the pure death process. Stockholm: Scandinavian Journal of Statistics, vol 3: 61-72.

[6] KPA Pension. (the 1 July 2002). http://www.kpa.se/upload/Trycksaker/ForArbetsgivare/857%20-%20KPA%20Planen.pdf the 27 September 2013.

[7] H. Lundström, Å. Nilsson, J. Qvist. (2004). Dödlighet efter utbildning, boende och civilstånd. Örebro: Statistiska Centralbyrån.

[8] S. Malmgren. https://lagen.nu/2010:110#K2P7 the 30 September 2013.

[9] S. Malmgren. https://lagen.nu/2010:110#K58 the 30 September 2013.

[10] FTN. (2006). Instruktion för rapportering till FTNs dödlighetsundersökningar, Försäkringstekniska Forskningsnämnden, Stockholm, Sverige.

[11] H. U. Gerber. (1995). Life Insurance Mathematics, Second edition. Berlin: Springer.

[12] R. C. Hill, W. E. Griffiths and G. C. Lim. (2008). Principles of Econometrics, third edition. United States of America: John Wiley & Sons, Inc.

[13] P. Kennedy. (2008). A guide to econometrics, sixth edition. United Kingdoms: Blackwell publishing.

[14] G. A. F. Seber and C. J. Wild. (1989). Nonlinear regression. United States of America. John Wiley & Sons, inc.

[15] C. L. Chiang. (1968). Introduction to Stochastic - Process in Biostatistics. United States of America. John Wiley & Sons, Inc.

[16] Pensionsmyndigheten. (the 10 November 2013). http://www.pensionsmyndigheten.se/Pensionsordlista.html#P the 14 January 2014.

[17] Regeringskansliet, Government offices of Sweden. (the 31 January 2012). http://www.government.se/sb/d/15473/a/183495 the 14 January 2014.

[18] Regeringskansliet, Government offices of Sweden. (the 19 September 2005). http://www.government.se/sb/d/5938/a/50061 the 14 January 2014.

[19] H. Hult, F. Lindskog, O. Hammarlid, C. J. Rehn. (2012). Risk and Portfolio Analysis. New York. Springer.

[20] B. Gompertz. (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 115: 513?585.

[21] W. M. Makeham. (1860). On the Law of Mortality and the Construction of Annuity Tables. J. Inst. Actuaries and Assur. Mag. 8: 301?310.

[22] R. D. Lee and L. Carter. (1992). Modeling and Forecasting the Time Series of U.S. Mortality. Journal of the American Statistical Association 87 (September): 659?671.

[23] Pensionsmyndigheten, Swedish Pensions Agency. (the first November 2013). https://www.pensionsmyndigheten.se/download/18.3ff0e0a7141eb15671117ea3/1383301922249/Unde
the 21 February 2014.

TRITA -MAT-E 2017:39
ISRN -KTH/MAT/E--17/39--SE