MATH3001: Survival Analysis

Investigating Longevity Risk in Pension Schemes

Afiba Annor

Supervisor: Dr. Leonid Bogachev

Submitted in accordance with the requirements for the degree
BSc. Mathematics

The University of Leeds
Faculty of Engineering and Physical Sciences
School of Mathematics

March 2020

Abstract

Modelling data and predictive analysis is vital for any organization, it enables one to reduce costs, make smarter business decisions and test efficiency. This study aims to provide an insight into survival analysis, particularly on its use in actuarial science, and analyse survival models to form predictions on life expectancy and investigate longevity risk using the statistical package R.

Contents

1 Introduction

1.1 Background

Survival analysis describes the analysis of data in the form of times from a “well-defined time origin up until the occurrence of some particular event” [1, chp.1]. Although mainly used within biomedical science, it has various other applications within social sciences, insurance and industrial reliability. For example,

- Time until death (literal survival time)
- Lifetime of a machine component [2]
- Period of unemployment [2]
- Time from marriage until divorce

Survival can also be considered in terms of risk. For example, suppose an individual has remained cancer-free for three years, what is the risk that they will be subject to a relapse within each unit of time [3]?

This report will focus on the applications of survival analysis in actuarial science. Hence we will discuss survival analysis in the context of the time until the death of an individual.

Before we proceed to discuss the survival function we must first understand the special features of survival data.

Definition 1.1

The survival time of a subject from a population is a sample value of a random variable $X \geq 0$ (i.e. random survival time).

1.2 Distribution

Consider a histogram of survival times. Generally, the distribution of survival times are positively skewed i.e. “the histogram will have a longer “tail” to the right of the interval that contains the largest number of observations” [1, chp.1]. An example of this can be seen in figure 1.

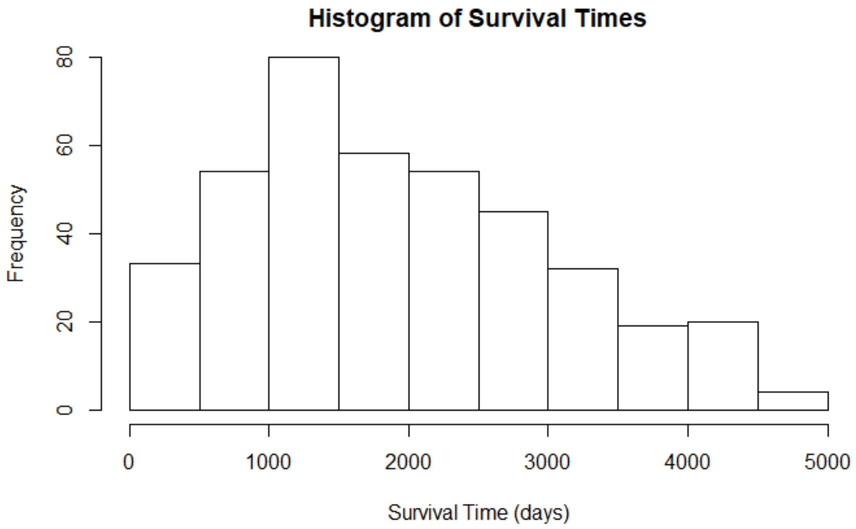


Figure 1: A histogram of the survival times of individuals from a Primary Biliary Cholangitis trial conducted between 1974 and 1984. Data sourced from R.

Here we have a histogram of the survival times of subjects for a Primary Biliary Cholangitis (PBC) trial. We can see that the data is positively skewed. Hence, survival data is usually non-symmetric. As a result, it is not appropriate to assume the data follows a normal distribution.

1.3 Censoring

One challenge faced in survival analysis is that the full lifetime may not be observed for some subjects [2, pp.15]. “The survival time of an individual is said to be censored when the end-point of interest has not been observed for that individual” [1]. This may be because the individual is still alive at the end of the study. Alternatively, the subject could have withdrawn from the study or moved location and can no longer be traced. It is also possible that the individual died due to causes unrelated to the study. For example, a subject for a PBC trial may die from a car crash.
These are all forms of right censoring. This is when “the event under study is not experienced by the last observation” [3, chp.1]. This form of censoring is most common.

Note: It can be difficult to distinguish whether or not the causes are related. Using the previous example, the car crash may be linked to PBC. Say the subject developed severe fatigue, a symptom of PBC, and therefore had a lack of awareness and concentration whilst driving, resulting in the car crash.

Left censoring is when the event of interest occurred before the observation period, “with the length of time spent in the origin state being unknown” [3, glossary]. For example, consider a study investigating the age of Leukaemia diagnosis. A subject may have the condition before the study but no record of a prior diagnosis.

Example 1.1

Suppose we are conducting a study on Leukaemia survivals. Subjects are enrolled at different times (i.e. a staggered entry) between 1^{st} March 1998 and 1^{st} June 1998. The study ends on 1^{st} February 2003.
The data is shown in the table below:

Subject	Start Date	End Date	Censoring	Comment
1	1998-03-01	1999-04-01	0	died
2	1998-04-01	2001-07-01	1	lost
3	1998-05-01	2003-02-01	1	alive
4	1998-06-01	2001-05-01	0	died

Table 1: A table showing start and end dates of individuals in a Leukaemia study in addition to which individuals are censored.

We can see the censored data more clearly in the graph below. This data is right censored as the deaths of subjects 2 and 3 have not been observed since subject 3 is alive at the end of the study and subject 2 has been lost, possibly due to migration.

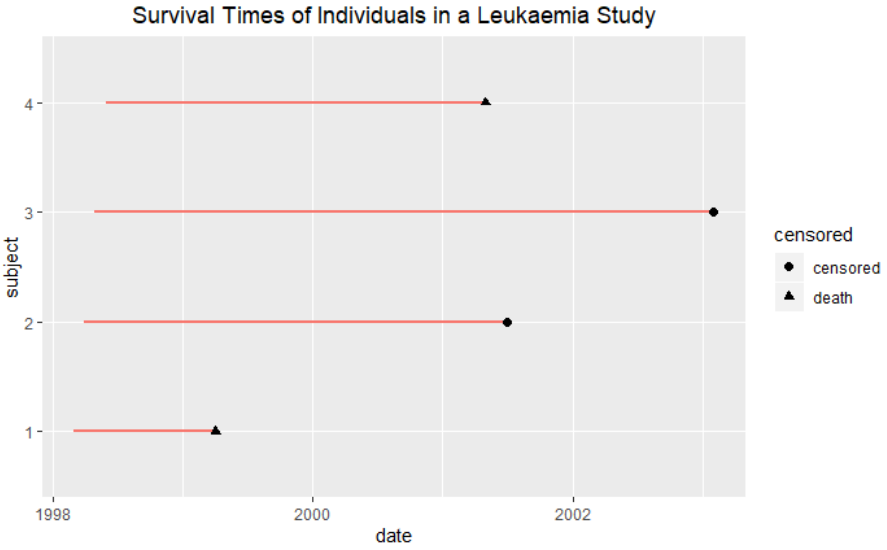


Figure 2: A chart visualising the start and end dates of the subjects represented in Figure 1.

2 The Survival Function

This chapter will illustrate the foundations of summarising survival data, using the survival function and its relations.

2.1 Assumptions

1. $X < \infty$ with $P(X < \infty) = 1$ (survival cannot continue for an indefinite period)

   Note: Theoretically, X has no upper bound (i.e. it takes values in $[0, \infty)$). However, in practical terms (e.g. actuarial applications) it may be useful to choose such a bound, say ω, then X takes values in $[0, \omega]$. An example of this can be seen in De Moivre's Law which is discussed later in table 2.

2. Each of the random survival times X_1, X_2, ... are independent and identically distributed (*i.i.d.*).

3. The population is homogeneous (i.e. the individuals within the population have the same characteristics).

Definition 2.1

The survival function of X is the probability that a new life (currently age 0) survives until age x and is defined as

$$S_0(x) = P(X \geq x) \qquad \text{where } x \geq 0 \tag{2.1}$$

For the function to be valid it must satisfy the following properties

- $S_0(0) = 1$
- $\lim_{x \to \infty} S_0(x) = 0$
- $S_0(x)$ is non-increasing

Definition 2.2

The cumulative distribution function of X is the probability that a new life (currently age 0) dies before x is defined as

$$F_0(x) = P(X < x) \qquad \text{where x} \geq 0 \tag{2.2}$$

For the function to be valid it must satisfy the following properties

- $F_0(0) = 0$
- $\lim_{x \to \infty} F_0(x) = 1$
- $F_0(x)$ is non-decreasing

Note: Since dying before age x and surviving until age x are complementary events we can state the following relationship.

$$F_0(x) = 1 - S_0(x)$$

Definition 2.3

Assume there exists a continuously differentiable survival function $S_0(x)$ for the random variable X. Then X has a continuous probability distribution function (pdf) defined as

$$f_0(x) = \frac{d}{dx}F_0(x) \tag{2.3}$$

Using the pdf we can state the integral expressions of $S_0(x)$ and $F_0(x)$

$$S_0(x) = \int_x^\infty f_0(s)ds \qquad \text{where } x \geq 0$$

$$F_0(x) = 1 - S_0(x) = \int_0^x f_0(s)ds \qquad \text{where } x \geq 0$$

Note: $F_0(x) = 1 - S_0(x) \implies S_0'(x) = -F_0'(x) = -f_0(x)$

Remark: For a life currently aged $x \geq 0$, if X is the age at death random variable. Then, given $X \geq x$, the future lifetime random variable is defined as,

$$T_x = X - x$$

The probability that a life aged $x \geq 0$ survives another t years is $S_x(t) = P(T_x \geq t)$, in actuarial science this is denoted by ${}_tp_x$. Similarly we have,

- $F_x(t) = P(T_x < t) = 1 - S_x(t)$
- $f_x(t)= \frac{d}{dt}F_x(t)$
- $S_x(t) = \int_t^\infty f_x(s)ds$

with $F_x(t)$ denoted by ${}_tq_x$ in actuarial science.

Note: $S_x(t) = \frac{S_0(x+t)}{S_0(x)}$

2.2 Mean and Median Survival Times

When considering summary measures, the median is preferred over the mean, since survival data is typically positively skewed [1, chp.2]. In this section we investigate how to calculate both measures.

Definition 2.4

The mean survival time, m, is defined as

$$m = E(T_x) = \int_0^\infty x f_0(x) dx = -\int_0^\infty x S_0'(x) dx \tag{2.4}$$

Note: The can also be given by

$$m = \int_0^\infty S_0(x) dx \tag{2.5}$$

Definition 2.5

The median survival time is defined as "the value for which 50% of the individuals in the study have longer survival times and 50% have shorter survival times" [5, pp.8]. Hence to calculate the median, denoted by β, set

$$S_0(\beta) = \frac{1}{2} \tag{2.6}$$

and rearrange to find β.

Example 2.1

Suppose we have $S_0(x) = e^{-\lambda x}$, where $\lambda > 0$, the hazard function, is a constant (we will discuss the hazard function in the next section), then by (2.5) the mean is,

$$\begin{aligned} m &= \int_0^\infty S_0(x) dx \\ &= \int_0^\infty e^{-\lambda x} dx \\ &= \frac{1}{\lambda} \end{aligned}$$

By (2.6) the median is,

$$\begin{aligned} S_0(\beta) = \frac{1}{2} &\implies e^{-\lambda\beta} = \frac{1}{2} \\ &\implies \beta = \frac{\log(2)}{\lambda} \end{aligned}$$

2.3 Hazard Function

Informally, the hazard function, also known as the hazard rate, is the conditional probability that an event occurs in a particular time interval.

Definition 2.6

For a life aged x, the hazard function (or hazard rate), $h(x)$, is defined as the probability per unit time of an event "immediately" after time x, conditioned on survival up to x, i.e.

$$h(x) = \lim_{\delta x \to 0} \frac{P(X \leq x + \delta x | X \geq x)}{\delta x} \tag{2.7}$$

In actuarial science this is denoted by μ_x

Note:
$$\begin{aligned} h(x) &= \lim_{\delta x \to 0} \frac{P(X \leq x + \delta x | X \geq x)}{\delta x} \\ &= \lim_{\delta x \to 0} \frac{P(X \geq x) - P(X \geq x + \delta x)}{P(X \geq x)\delta x} \\ &= \lim_{\delta x \to 0} \frac{S_0(x) - S_0(x + \delta x)}{S_0(x)\delta x} \\ &= -\frac{1}{S_0(x)} \lim_{\delta x \to 0} \frac{S_0(x + \delta x) - S_0(x)}{\delta x} \\ &= -\frac{1}{S_0(x)} S_0'(x) \\ &= \frac{f_0(x)}{S_0(x)} \qquad \text{(as } f_0(x) = -S_0'(x)) \end{aligned}$$

So we have

$$h(x) = \frac{f_0(x)}{S_0(x)} \tag{2.8}$$

Definition 2.7

The cumulative hazard function (i.e. the expected number of events that occur in the interval $(0, x)$), $H(x)$, is defined

$$H(x) = \int_0^x h(s)ds \qquad \text{where } x \geq 0 \tag{2.9}$$

Theorem 2.1

The cumulative hazard function can be related to the survival function via the following theorem.

$$S_0(x) = \exp(-H(x)) \qquad \text{where } x \geq 0 \tag{2.10}$$

Proof

$$\begin{aligned} h(x) &= \lim_{\delta x \to 0} \frac{P(X \leq +\delta x | X > x)}{\delta x} \\ &= -\frac{1}{S_0(x)} S_0'(x) && \text{(by previous note)} \\ &= -\frac{d}{dx}(\log S_0(x)) \end{aligned}$$

$$\begin{aligned} &\implies \log(S_0(x)) = -\textstyle\int_0^x h(s)ds = -H(x) \\ &\implies S_0(x) = \exp(-H(x)), \text{ where } x \geq 0 \end{aligned}$$

□

3 Non-Parametric Survival Estimators

In this chapter we examine various non-parametric methods of estimating survival curves. Non-parametric survival models "can be used as a check on the reasonableness of the fitted parametric curves" [4, pp.248]. For each estimator we follow notation similar to the notation described in the book Modelling Survival Data in Medical Research [1].

3.1 Actuarial Estimate of the Survival Function

The actuarial estimate of the survival function is also known as the life-table estimate. It is calculated by splitting the period of observations into a series of time intervals [1, chp.2], with times x'_j such that,

$$0 \leq x'_1 < x'_2 < ... < x'_{m-1} < x'_m \leq \infty$$

with $j = 1, 2, ..., m$

Note: Usually the number of intervals, m, is between 5 and 15 [1, chp.2].

For $j = 1, 2, ..., m$ let us define the following:

- $[x'_j, x'_{j+1})$:= the jth interval
- n_j := the number of individuals alive at the start of the jth interval
- d_j := the number of deaths in $[x'_j, x'_{j+1})$
- c_j := the number of censored survival times in $[x'_j, x'_{j+1})$

3.1.1 The Actuarial Assumption

We assume that the censoring process is such that the censored survival times occur uniformly throughout the interval $[x'_j, x'_{j+1})$ [1, chp.2].Then the average number of individuals who are at risk during the intervals is,

$$n'_j = n_j - \frac{c_j}{2}$$

Due to this assumption, we estimate that the probability of death in the jth interval is given by $\frac{d_j}{n'_j}$. It follows that the survival probability is $\frac{n'_j - d_j}{n'_j}$. The probability that a subject is alive in the kth interval, $k = 1, 2, ..., m$, is a product of the probabilities that subject survives in each of the previous intervals. Hence the actuarial estimate of the survival function is

$$S^*_0(x) = \prod_{j=1}^{k} \frac{n'_j - d_j}{n'_j} \tag{3.1}$$

for $x'_k \leq x < x'_{k+1}, k = 1, 2, ..., m$

Example 3.1

Below shows a plot of the actuarial survival probability estimate calculated from survival data of patients with Multiple Myeloma. Notice that $S_0^*(0) < 1$, this is due to the way the intervals are constructed, so in this example there are multiple deaths in the first interval.

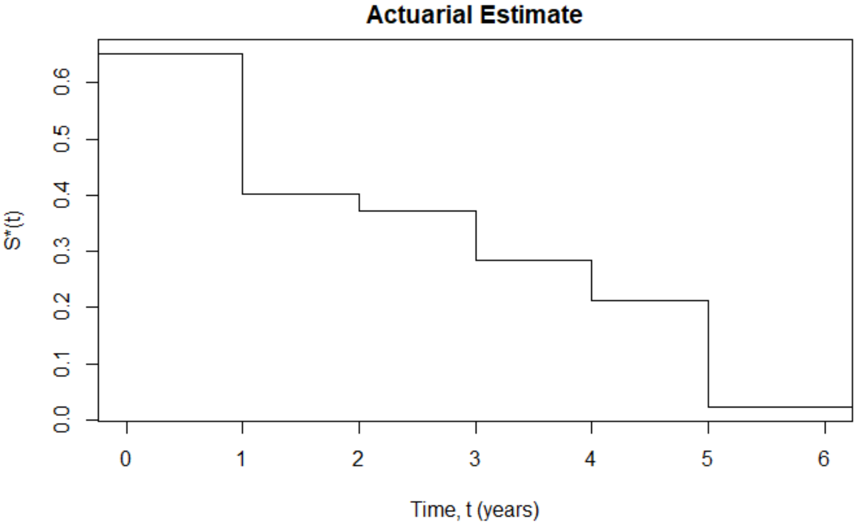


Figure 3: A chart showing the Actuarial survival probability estimate of patients with Multiple Myeloma. Data sourced from [1].

3.2 Kaplan–Meier Estimate of the Survival Function

For the Kaplan–Meier estimate we construct the time intervals similarly to as we did for the actuarial estimate. But here we construct the intervals such that there is only one death time in each interval, located at the start of the interval.
To understand this further consider the following situation illustrated in the figure below.

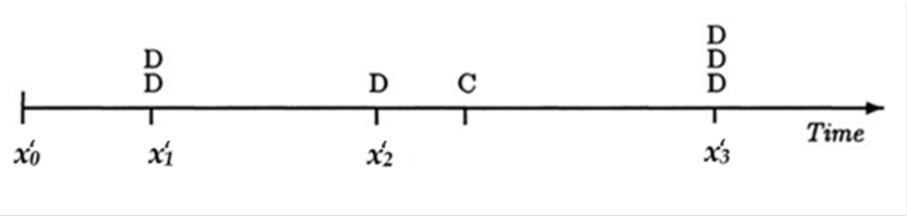


Figure 4: A construction of intervals used in the derivation of the Kaplan Meier Estimate. Adapted from [1, chp.2].

In Figure 4, D denotes a death and C denotes a censored survival time. x_0' denotes the time origin. This initial period runs from x_0' to just before x_1'. Hence, there are no deaths in the initial period. Then the first interval, which runs from x'_1 to just before x'_2, contains two death times. Similarly the second interval contains one death time and a censored time and the third interval contains three death times.

Recall the following definitions for $j = 1, 2, ..., m$:

- $n_j :=$ the number of alive individuals just before the jth interval
- $d_j :=$ the number of deaths in $[x'_j, x'_{j+1})$
- $c_j :=$ the number of censored survival times in $[x'_j, x'_{j+1})$

Since there are n_j individuals who are alive just before x'_j and d_j deaths at x'_j, the probability that an individual dies during the interval from $x'_j - \delta$ ($\delta > 0$ and is infinitesimal) to x'_j is estimated by $\frac{d_j}{n_j}$ [1, chp.2] . Hence, the probability of survival is $\frac{n_j - d_j}{n_j}$.

Note: Sometimes censored survival times occur at the same time as a death. In this case we assume the death time occurs first and the censored times occur immediately after.

Using the same argument as before, we obtain the following estimate

$$\hat{S}_0(x) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j} \tag{3.2}$$

for $x'_k \leq x < x'_{k+1}$, $k = 1, 2, ..., m$

Example 3.2

Below shows a plot of two Kaplan–Meier survival estimates for two different treatment regimes used in an ovarian cancer trial. The vertical lines on the curve indicate censored observations.

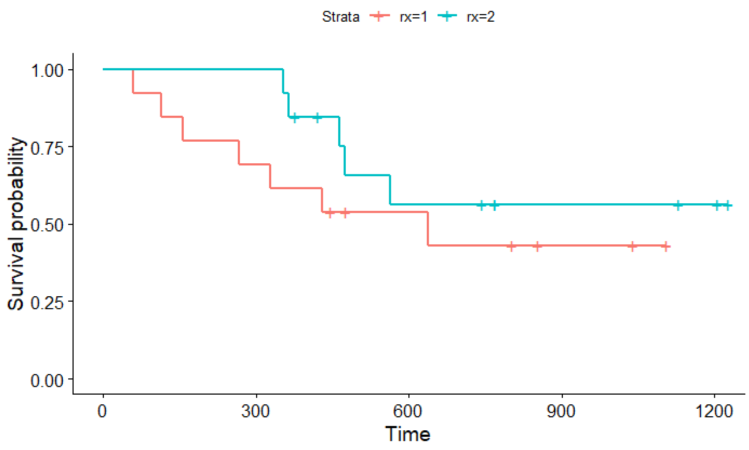


Figure 5: A chart showing the Kaplan–Meier survival probability estimate of the two treatment regimens for ovarian cancer [9].

3.3 Nelson–Aalen Estimate of the Survival Function

Recall the definitions given in section 3.1 for $j = 1, ..., m$. The Nelson–Aalen estimate can be obtained by using the estimate of the cumulative distribution function, $H(x_j) \approx \sum_{i=1}^{j} \frac{d_i}{n_i}$ [8]. This is based on individual death times, giving the Nelson–Aalen estimate as

$$\tilde{S}_0(x) = \exp(-H(x)) = \prod_{j=1}^{k} \exp(-d_j/n_j) \tag{3.3}$$

Note: The Nelson–Aalen estimate is an approximation to the Kaplen–Meier estimate. We know that by Taylor expansion we have

$$\begin{aligned} \exp(d_j/n_j) &= 1 - (d_j/n_j) + \frac{(d_j/n_j)^2}{2} - \frac{(d_j/n_j)^3}{6} + ... \\ &\approx 1 - (d_j/n_j) \qquad \text{(when } d_j/n_j \text{ is small)} \\ &= \frac{n_j - d_j}{d_j} \end{aligned}$$

Example 3.3

The plot below shows the Nelson–Aalen estimate of the survival data of patients from an Acute Myelogenous Leukemia study.

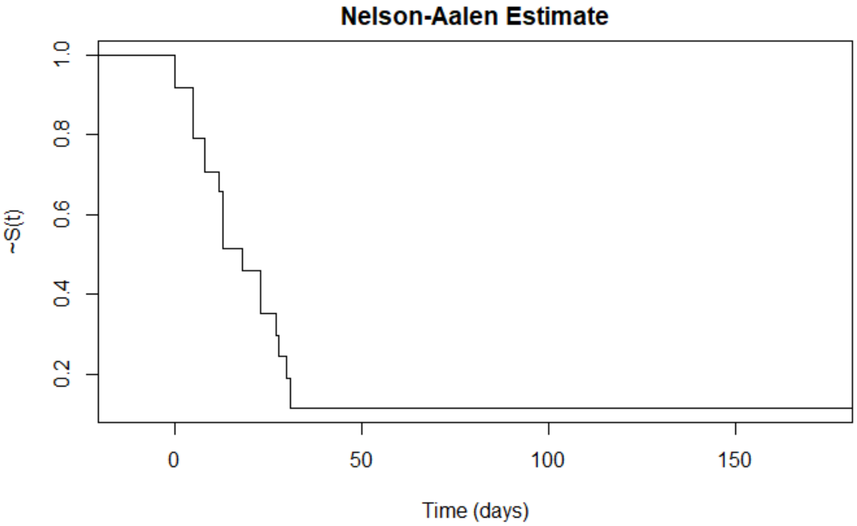


Figure 6: A chart showing the Nelson–Aalen survival probability estimates for patients from an Acute Myelogenous Leukemia study. Data sourced from R.

4 Parametric Survival Estimators

"Models in which a specific probability distribution is assumed for the survival times are known as parametric models" [1]. In this chapter we explore different parametric methods of estimating survival.

4.1 Exponential Distribution

The exponential distribution can be used if the hazard function is constant. The constant hazard rate is a unique property of the exponential distribution [5, pp.91]. This can be seen below.

Let $h(x) = \lambda$ where λ is a positive constant. Then by (2.9) and (2.10) we obtain

$$H(x) = \lambda x \implies S_0(x) = e^{-\lambda x}$$

Recall from example 2.1 that the exponential survival function has mean

$$m = \int_0^\infty S_0(x)dx = \int_0^\infty e^{-\lambda x}dx = \frac{1}{\lambda}$$

and median

$$S_0(\beta) = \frac{1}{2} \implies e^{-\lambda\beta} = \frac{1}{2} \implies \beta = \frac{\log(2)}{\lambda}$$

4.1.1 Assessment of Parametric Fit

Now how do we test if our data set fits this distribution, i.e. is our survival data such that $\acute{S}_0(x) \approx e^{-\lambda x}$?
It can be difficult to see if the data follows an exponential distribution on a plot. Hence, we perform a log transformation to obtain a linear model. So we have

$$\log(\acute{S}_0(x)) \approx -\lambda x$$

By plotting $\log(\acute{S}_0(x))$ against x, with intercept 0 and gradient $-\lambda$, a linear relationship is much easier to see.

4.2 Weibull Distribution

"The Weibull distribution is described for situations where the hazard rate is not constant but smoothly increasing or decreasing with time" [5, pp.91]. The survival function is defined by

$$S_0(x) = e^{-\lambda x^\kappa}$$

where λ and κ are constants with $\kappa > 1, \lambda > 0$.
Note: When $\kappa = 1$ we have the exponential distribution

Using (2.9) and (2.10) we obtain the hazard rate.

$$H(x) = -\log(e^{-\lambda x^\kappa}) = \lambda x^\kappa \implies h(x) = \lambda\kappa x^{\kappa-1}$$

Hence the hazard rate is time dependant.

From (2.5) and (2.6) the Weibull survival function has the following mean and median.

$$\begin{aligned} m &= \int_0^\infty S_0(x)dx \\ &= \int_0^\infty e^{-\lambda x^\kappa} dx \\ &= \kappa^{-1}\lambda^{-\frac{1}{\kappa}} \int_0^\infty x^{\frac{1}{\kappa}-1} e^{-y} dy \qquad \text{(by substituting } y = \lambda t^\kappa) \\ &= \kappa^{-1}\lambda^{-\frac{1}{\kappa}} \Gamma(\frac{1}{\kappa}) \end{aligned}$$

where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$.

$$S_0(\beta) = \frac{1}{2} \implies e^{-\lambda\beta^\kappa} = \frac{1}{2} \implies \beta = (\frac{\log(2)}{\lambda})^{\frac{1}{\kappa}}$$

4.2.1 Assessment of Parametric Fit

Once again it is difficult to see if data follows this distribution so we apply a transformation. We want to test if our survival data is such that $\acute{S}_0(x) \approx e^{-\lambda x^\kappa}$. To obtain a linear model we apply the following transformations.

$$\begin{aligned} \log \acute{S}_0(x) &= -\lambda x^\kappa \\ \log(-\log(\acute{S}_0(x))) &= \log(\lambda x^\kappa) \\ &= \log(\lambda) + \kappa \log(x) \end{aligned}$$

By plotting $\log(-\log(\acute{S}_0(x)))$ against $\log(x)$ we obtain a linear relationship with intercept $\log(\lambda)$ and gradient κ.

4.3 More Parametric Families

Table 2 provides the characteristics of exponential and Weibull distribution in a more concise format, along with De Moivre's Law and the Gompertz and Makeham distributions.

Law/Distribution	Survival Function $S_0(x)$	Hazard Rate $h(x)$	Limitations
Exponential	$e^{-\lambda x}$	λ	$\lambda > 0$
Weibull	$e^{-\lambda x^{\kappa}}$	$\lambda \kappa x^{\kappa-1}$	$x \geq 0, \kappa > 1, \lambda > 0$
De Moivre	$1 - \frac{x}{\omega}$	$\frac{1}{\omega - x}$	$0 \leq x < \omega$
Gompertz	$exp\{-\frac{\lambda}{\log(\theta)}(\theta^x - 1)\}$	$\lambda e^{\theta x}$	$\lambda > 0, \theta > 1$
Makeham	$exp\{-\gamma x - \frac{\lambda}{\log(\theta)}(\theta^x - 1)\}$	$\gamma + \lambda e^{\theta x}$	$\gamma \geq -\lambda, \theta > 1, \lambda \geq 0$

Table 2: A table of the characteristics on the following parametric estimators: Exponential, Weibull, De Moivre, Gompertz and Makeham. Adapted from [14, chp.1].

Gompertz (1825) found that in human life, for the majority of individuals (not those in the early stages of life), the hazard rate is exponential. Makeham (1860) discovered the Gompertz model could be improved by adding a constant, γ, to the hazard rate, [13, chp.2]. We will discuss the Makeham model further in chapter 7.

5 Comparison of Survival Curves

There are many instances where a series of survival curves have been developed. However, in order to see whether there is a difference between survival curves, we need to compare them. Various methods can be used to compare survival curves, including the hazard ratio and the Wilcoxon test. This chapter will explore the Log-rank test method.

5.1 Hypothesis Testing

The null hypothesis, H_0 is the criteria we are testing against, H_1 is the alternative hypothesis. H_0 can be regarded as the measure of goodness of fit.
Suppose we have survival curves $S_0^1(x)$ and $S_0^2(x)$, for $x \geq 0$. If we simply want to test whether there is a difference between these curves we conduct a two sided hypothesis test.

$$H_0 : S_0^1(x) = S_0^2(x)$$
$$H_1 : S_0^1(x) \neq S_0^2(x)$$

If we want to test whether $S_0^2(x)$ has higher survival rates that $S_0^1(x)$ then we conduct a one sided hypothesis test.

$$H_0 : S_0^1(x) = S_0^2(x)$$
$$H_1 : S_0^1(x) < S_0^2(x)$$

Similarly to test whether $S_0^1(x)$ has higher survival rates than $S_0^2(x)$.

Next we find the test-statistic. This measures how much the observed data differs from the null hypothesis [1, chp.2]. Using the distribution of the test-statistic, the corresponding p-value can be found. If the p-value is small we reject H_0. If the p-value is large we accept H_0.

5.2 The Log-rank Test

The log-rank test is the most commonly used statistical test for survival data [3, chp.4]. Consider two groups of survival data, Group 1 and Group 2. Suppose this is pooled data with r distinct death times.

$$x_1 < x_2 < ... < x_r$$

Say at time x_j, $j = 1, ..., r$, d_{1j} and d_{2j} individuals die in Group 1 and 2 respectively. We also have n_{1j} as the number of individuals at risk of death in Group 1 and n_{2j} at risk of death in Group 2. Taking $d_j = d_{1j} + d_{2j}$ and $n_j = n_{1j} + n_{2j}$, we have the following table.

Group	Number of deaths at x_j	Number surviving beyond x_j	Number at risk just before x_j
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

Table 3: A table summarising the number of individuals in various states, with time x_j as a reference, in Group 1 and Group 2 [1, chp.2].

We can test whether there is a difference in the survival experience between the two groups. Take $S_0^1(x)$ and $S_0^2(x)$ to be the survival functions of Group 1 and Group 2 respectively. Then we have

$$H_0 : S_0^1(x) = S_0^2(x)$$
$$H_1 : S_0^1(x) \neq S_0^2(x)$$

If the marginal totals in Figure 4 are fixed and H_0 is true then d_{1j} is a random variable with a *hypergeometric distribution* [1, chp.2]. Hence, the probability that there are k deaths in Group 1 is

$$P(d_{1j} = k) = \frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}, \qquad k = 0, 1, ..., min\{d_j, n_{1j}\}$$

Note: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, read "$n$ choose k", is the number of different ways k can be chosen from n.

Provided H_0 is true, the <u>mean</u> of the hypergeometric variable d_{1_j} is given by

$$E(d_{1j}) = e_{1_j} = \frac{n_{1j} d_j}{n_j} \tag{5.1}$$

Then <u>variance</u> of d_{1j} is given by

$$Var(d_{1j}) = v_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \tag{5.2}$$

5.2.1 Normal Version Test Statistic

Consider the following,

$$U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j}) \tag{5.3}$$

This is the difference between the observed numbers of deaths and the expected number of deaths in Group 1. As stated above $E(d_{1j}) = e_{1j}$, hence the mean of U_L is zero. Let the variance of U_L be denoted as below.

$$Var(U_L) = \sum_{j=1}^{r} v_{1j} = V_L \tag{5.4}$$

U_L is approximately normal, hence we obtain the following test statistic.

$$\frac{U_L}{\sqrt{V_L}} \sim N(0,1) \tag{5.5}$$

5.2.2 Chi-Squared Test Statistic

Now, the square of a standard normally distributed variable has a chi-squared distribution with one degree of freedom. Hence,

$$\frac{U_L^2}{V_L} \sim \chi_1^2 \tag{5.6}$$

Note: The chi-squared test statistic can be used to compare more than two groups of data.

6 Actuarial Concepts

In this chapter we discuss actuarial notation and introduce the actuarial concepts vital for calculating the benefit for a classical whole of life annuity policy.

6.1 Actuarial Notation

Recall from chapter 2 that we have the following notations in actuarial science.

Function	Actuarial Notation
$S_x(t)$	${}_tp_x$
$F_x(t)$	${}_tq_x$
$h(x)$	μ_x

Table 4: A table showing the actuarial notation for the functions described throughout chapter 2.

Note: Recall T_x is the future lifetime random variable. In actuarial science the hazard rate, $h(x)$, is known as the force of mortality and is denoted by μ_x.
Using actuarial notation, the force of mortality for the future lifetime variable is defined as

$$\mu_{x+t} = \lim_{\delta t \to 0} \frac{P(T_x \leq t + \delta t | T_x > t)}{\delta t}$$

Hence, we have the following relation,

$$ {}_tp_x = S_x(t) = \exp(-\int_0^t \mu_{x+s} ds) \tag{6.1}$$

Note: Let l_x denote the number of survivors aged x. Given that there are no censored times, the actuarial estimate (covered in section 3.1.1) becomes,

$$ {}_tp_x = \frac{n_j - d_j}{n_j} = \frac{l_{x+t}}{l_x} \tag{6.2}$$

where n_j and d_j are defined as in section 3.1 with the jth interval as $[x, x+t)$. So we have, $l_{x+t} = n_j - d_j$ and $l_x = n_j$.

Definition 6.1

For a life aged x, K_x is the curtate future lifetime random variable and is defined as

$$K_x = \lfloor T_x \rfloor \tag{6.3}$$

The curtate expectation of life is the mean survival time, as defined in (2.4) but for discrete random variable K_x, is defined as,

$$e_x = \mathbb{E}[K_x] = \sum_{k=0}^{\infty} kP(K_x = k)$$

Note: $e_x = \mathbb{E}[K_x] = \sum_{k=0}^{\infty} kP(K_x = k)$

$$\begin{aligned}
&= \sum_{k=0}^{\infty} k({}_kp_x -_{k+1} p_x) \\
&= \sum_{k=0}^{\infty} k({}_kp_x) - \sum_{k=0}^{\infty} k({}_{k+1}p_x) \\
&= ({}_1p_x + 2{}_2p_x + ... + n{}_np_x + ...) - ({}_2p_x + 2{}_3p_x + ... + (n-1){}_np_x + ...) \\
&= {}_1p_x + {}_2p_x + ... + {}_np_x + ... \\
&= \sum_{k=1}^{\infty} {}_kp_x
\end{aligned}$$

Hence,

$$e_x = \sum_{k=1}^{\infty} {}_kp_x \tag{6.4}$$

Definition 6.2

The central death rate of a life currently aged x after t years is defined by,

$$m_{x+t} = \frac{d_{x+t}}{E_{x+t}} \tag{6.5}$$

where d_{x+t} is the number of deaths between ages x and $x+t$, and E_{x+t} is the exposure-to-risk between ages x and $x+t$. Exposure-to-risk is an estimate of the number of individuals exposed to the risk of death in some time interval [19].

Note: The central death rate can be used to approximate the force of mortality, as $m_x \approx \mu_{x+1/2}$ [13]. In discrete time (consider the discrete random variable K_x) $\lfloor x + 1/2 \rfloor = x$ so $m_x \approx \mu_x$.

6.2 Whole of Life Insurance

A whole of life insurance pays out a benefit when the individual (i.e. the policyholder) immediately dies or on end of year of death. We will consider the end of year of death case as this is more practical since it takes time for paperwork to be completed [14, chp.3]. The premium is the amount of money the individual pays for the insurance coverage. Premiums are typically paid in regular intervals whilst the individual is alive [14, chp.3].

Recall K_x is the integer part of T_x. Hence, this is the whole number of years the policyholder survives. Since the benefit is paid at end of year of death it will be paid at $K_x + 1$.

For interest rate i per annum, we denote the following below

- Discount factor: $v = \frac{1}{1+i}$
- Discount rate: $d = 1 - v = \frac{i}{1+i}$
- Force of interest: $\delta = \log(1+i)$

6.2.1 Annuities

Definition 6.3

An n-year annuity due is the present value of a contract which pays a unit at the beginning of each year, from time $t = 0$ to time $t = n - 1$, and is denoted by,

$$\ddot{a}_{\overline{n}|} = \sum_{k=0}^{n-1} v^k = \frac{1 - v^n}{d} \tag{6.6}$$

By geometric summation, since $v < 1$ (assuming a positive interest rate).

Definition 6.4

The expected present value for a whole of life annuity due is given by

$$\ddot{a}_x = \mathbb{E}[\ddot{a}_{\overline{K_x+1}|}] = \sum_{k=0}^{\infty} \ddot{a}_{\overline{k+1}|} P(K_x = k) = \sum_{k=0}^{\infty} \ddot{a}_{\overline{k+1}|}({}_kp_x -_{k+1} p_x) \tag{6.7}$$

This can be transformed into an simpler equation using (6.6)

$$\begin{aligned}
\ddot{a}_x &= \sum_{k=0}^{\infty} \frac{1 - v^{k+1}}{d}({}_kp_x -_{k+1} p_x) \\
&= \frac{1}{d}\sum_{k=0}^{\infty}[(1 - v^{k+1})_kp_x - (1 - v^{k+1})_{k+1}p_x] \\
&= \frac{1}{d}[(1 - v)_0p_x + (1 - v^2)_1p_x + ...] - [(1 - v)_1p_x + (1 - v^2)_2p_x + ...] \\
&= \frac{1}{d}[(1 - v) + (v - v^2)_1p_x + (v^2 - v^3)_2p_x + ...] && (\text{as } {}_0p_x = 1) \\
&= \frac{1 - v}{d}[1 + v({}_1p_x) + v^2({}_2p_x) + ...] \\
&= \sum_{k=0}^{\infty} v^k{}_kp_x
\end{aligned}$$

So, we have

$$\ddot{a}_x = \sum_{k=0}^{\infty} v^k{}_kp_x \tag{6.8}$$

6.2.2 Pension Calculations

Definition 6.5

The present value of the future loss random variable is denoted by L_0 and defined as

$$L_0 = PV(\text{future outgo - future income}) \tag{6.9}$$

Note: PV denotes present value, i.e. the current value of payments made in the future.

The Equivalence Principle

The equivalence principle states that the expectation of L_0 is zero, i.e.

$$\mathbb{E}[L_0] = 0 \tag{6.10}$$

Now let us consider a **net future loss random variable**, L_0^N. Consider a life insurance/annuity contract with the following cash flows.

- Benefit, B (outgo)
- Net premium, P (income)

We then have

$$L_0^N = PV(\text{future benefit - future net premium})$$

By the equivalence principle we obtain

$$\mathbb{E}[L_0^N] = 0 \implies EPV(\text{future benefit}) - EPV(\text{future net premium}) = 0$$

Note: EPV denotes expected present value.

Using this we can calculate the benefit paid to a policyholder going on pension at age x. Suppose that the net premium, P, accumulated until the policyholder goes on pension, is denoted by M (i.e. the pension pot has value M). For a whole of life annuity contract with benefit B, payable annually until end of year of death, using (6.7) and (6.10), we have

$$\mathbb{E}[L_0^N] = B\ddot{a}_x - M = 0$$

$$\implies B = \frac{M}{\ddot{a}_x} \tag{6.11}$$

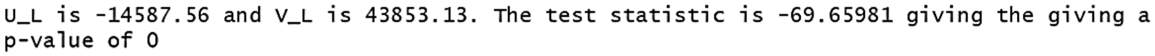
So the insurance company pays pension benefit, $B = \frac{M}{\ddot{a}_x}$, to the policyholder annually until end of year of death.

7 Investigating Longevity Risk

Over the years life expectancy has been on the rise as the number of deaths due to major diseases has fallen [17]. In fact, “global life expectancy for both sexes increased from 65.3 years in 1990 to 71.5 years in 2013” [17]. Increasing life expectancy has caused implications when it comes to retirement planning such as longevity risk. Longevity risk is “the risk that members of some reference population might live longer on average than anticipated” [16]. A study executed by MGM Advantage, now known as Canada Life, showed that in 2014, the life expectancies of males and females aged 55-64 were underestimated by 5 and 10 years respectively [10]. Underestimation of life expectancy can result in pension providers paying out more money over a longer period than expected. Hence, the whole of life contract described in section 6.2 has high longevity risk as the contract pays the policyholder for the rest of their life. In this chapter we will use the statistical package R to compare various models in order to investigate longevity risk. Throughout, we will also refer to the central death rate as the mortality/death rate.

7.1 The Data

We will be analysing the mortality rates in the United Kingdom from years 1922 to 2016, sourced from the Human Mortality Database [15]. Usually the survival experience for males and females differ. Since we have the population for each age in each year, to confirm this for our data, let us conduct a log-rank test on the data set for the year 2014. Referring to the R code given in Figure 25 in the Appendix, we obtain the following output.



```
U_L is -14587.56 and V_L is 43853.13. The test statistic is -69.65981 giving the giving a
p-value of 0
```

Figure 7: R output of the log-rank test between females and males in 2014.

Seeing as there are no censored times in this data set, we take the number of deaths as the difference between the the number at risk after each year, l_x, as the number of deaths. Using the methods described in chapter 5, the normal test statistic, as shown in (5.5), is -69.65981, hence it is so small that R computes the p-value as 0. Therefore, we can conclude that the survival experience between males and females is significantly different. Therefore, we will fit the data for models separately for males and females for further analysis.

Figures 8 and 9 show the log central death rates against age for each cohort for males and females respectively. The colored lines indicate the year, the deepest red being 1922 and the lightest purple being 2016. The sharp increase from late teens to early twenties is known as the *accident hump* [18]. In recent years, this is mainly due to increased driving accidents, particularly among males [18]. Notice that the accident hump for males is larger than the accident hump for females.

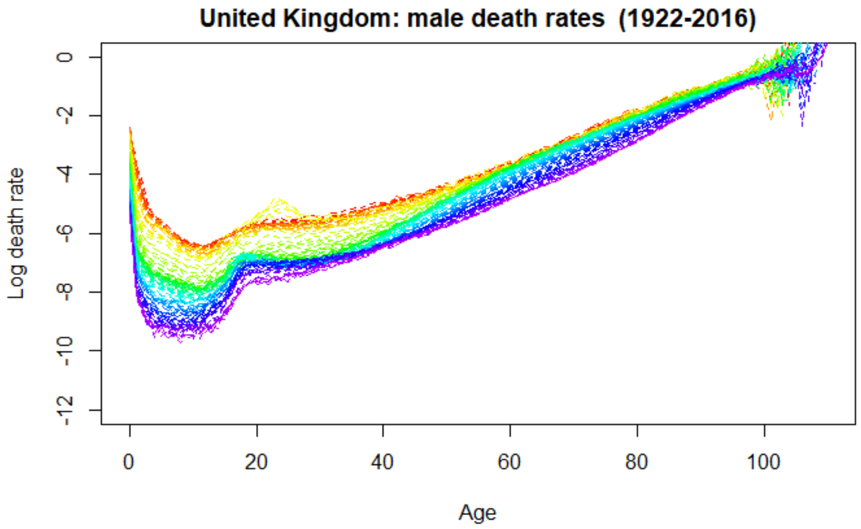


Figure 8: A chart showing the male log death rate against age for years 1922 to 2016.

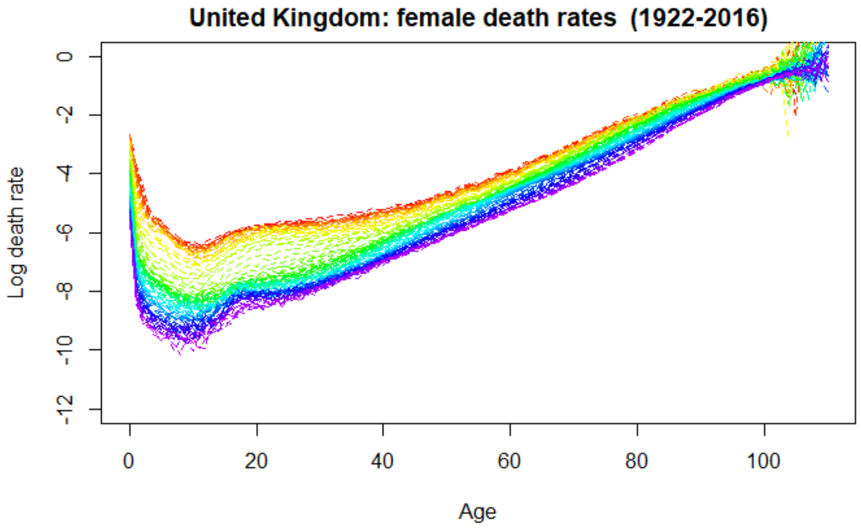


Figure 9: A chart showing the female log death rate against age for years 1922 to 2016.

Figures 10 and 11 show how the log central death rates change throughout time for each age (with ages ranging from 0 to 110). From the figures it is clear that the log central death rates decrease throughout the years. From (6.1) this implies that the survival probability is increasing at all ages. This is likely to be due to medical advancements and changes in lifestyle such as exercising and a reduction in smoking.

Note: Each coloured line represents a different age.

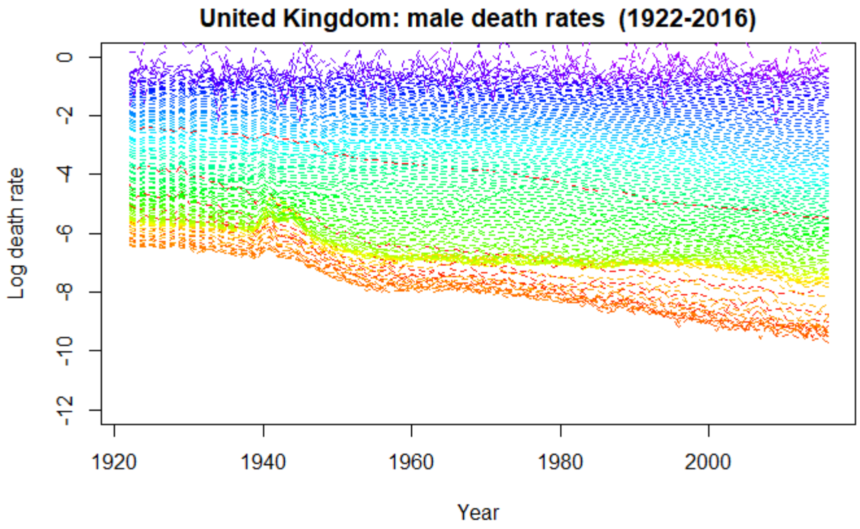


Figure 10: A chart showing the male log death rate through time for ages 0 to 110.

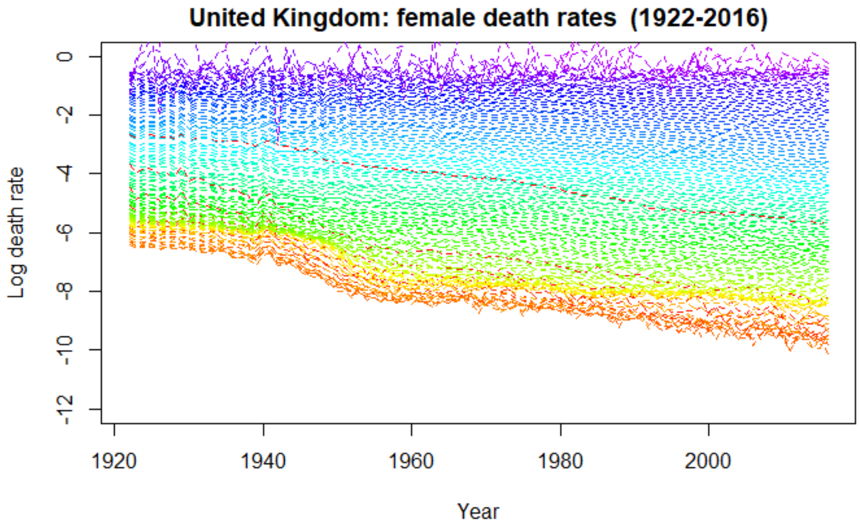


Figure 11: A chart showing the female log death rate through time for ages 0 to 110.

7.2 The Lee–Carter Model

As mentioned previously, life expectancy has been on the rise. As a result, techniques have been developed to project future mortality rates, predict life expectancy more accurately and combat longevity risk. One technique often used is the Lee–Carter model [21]. Originally, Ronald D. Lee and Lawrence Carter developed the model for U.S. mortality data from the years 1933 to 1987. However, now their approach is being applied to mortality data across various countries and time periods [22].

7.2.1 The Model

The Lee–Carter (1992) model [20] is defined as

$$\log(m_{x+t}) = a_x + b_x k_t \tag{7.1}$$

where m_{x+t} is the central death rate for a life currently aged x after t years and the parameters are represented as follows:

- a_x is the average of the log mortality rate for each age
- b_x is the deviation in the log mortality rate at each age
- k_t measures the general trend of log mortality rates through time

The paramaters k_t and b_x have the following constraints,

$$\sum_t k_t = 0, \qquad \sum_x b_x = 1$$

7.2.2 Parameter Analysis

Having fitted the Lee–Carter model we can see how the parameters change with age or time for our data set.

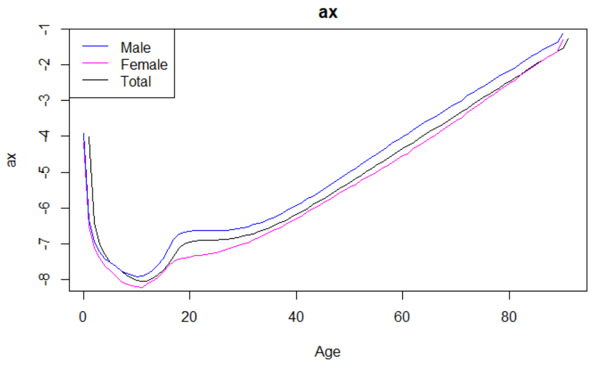


Figure 12: A chart showing the value of a_x for each age x.

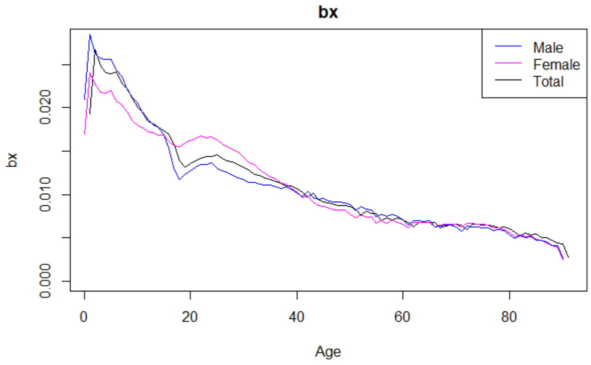


Figure 13: A chart showing the value of b_x for each age x.

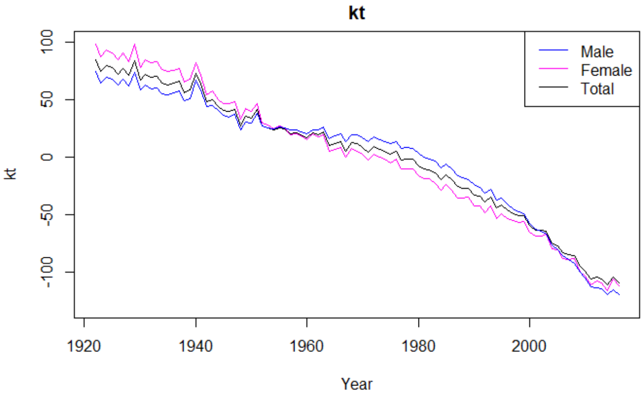


Figure 14: A chart showing the value of k_t for each year t.

The average log mortality rate, a_x, behaves as expected, similar to figures 8 and 9. Again we can see that the accident hump is larger for males. The deviation in the log mortality rate, b_x, is the greatest during the early stages of life, decreasing gradually in adult life. Also the general trend of log mortality, k_t, decreases through time as seen in figures 10 and 11.

7.2.3 Life Expectancy & Annuity Calculations

To forecast mortality rates into future years, we look at the time dependant parameter k_t. Using ARIMA models (a class of models used to forecast time series) we extrapolate the adjusted k_t to obtain future mortality rates. Hence, using R the mortality rates can be forecasted 110 years into the future with ARIMA models.

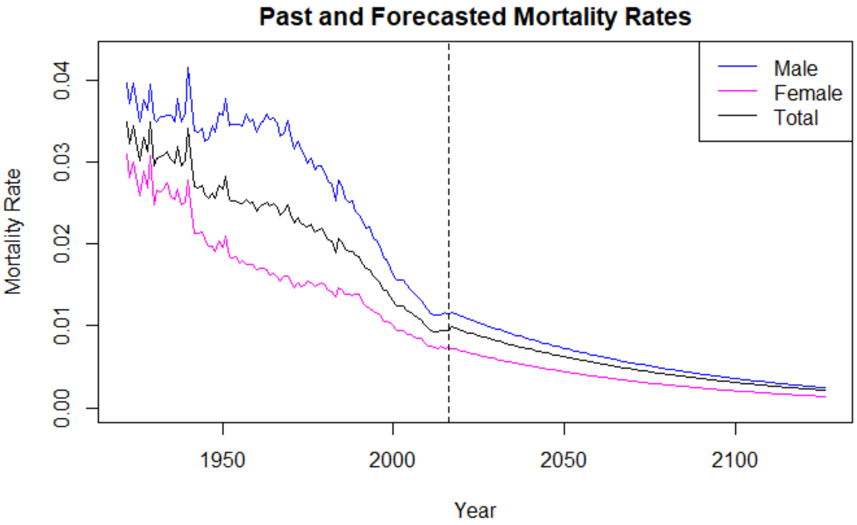


Figure 15: A chart showing the past mortality rates with the forecasted Lee–Carter mortality rates for individuals aged 65.

Next we create a life table for the males and females in R and then use (6.2) to find the survival probabilities. Using this we calculate $\ddot{a}_{65}$ and e_{65} from (6.4) and (6.8) for an interest rate of 5% per annum.

Note: "A life table is a demographic tool used to analyse death rates and calculate life expectancies at various ages" [23].

.

Cohort	1986	1990	1994	1998	2002	2006	2010	2014
e_{65}	20.78	21.05	21.30	21.55	21.80	22.04	22.27	22.50
$\ddot{a}_{65}$	13.10	13.20	13.31	13.41	13.50	13.60	13.69	13.78

Table 5: A table showing the future life expectancy and the value of the annuity for males aged 65 in different cohorts, using the Lee–Carter model.

Cohort	1986	1990	1994	1998	2002	2006	2010	2014
e_{65}	22.97	23.20	23.42	23.64	23.85	24.05	24.24	24.43
$\ddot{a}_{65}$	13.98	14.07	14.15	14.23	14.31	14.38	14.46	14.53

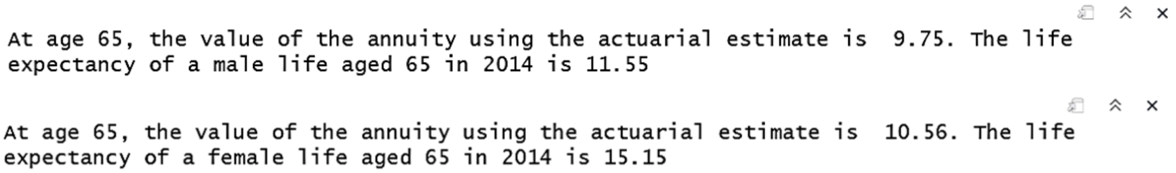
Table 6: A table showing the future life expectancy and the value of the annuity for females aged 65 in different cohorts, using the Lee–Carter model.

From the tables above we can see that in 2014, at age 65 men are expected to live another 22.5 years and women are expected to live another 24.43 years. Hence, men and women are expected to live until ages 87.50 and 89.43 respectively.

7.3 Comparisons

7.3.1 Actuarial Estimate

Recall in our data set we also have the population at each age, i.e. the number of survivors at each age, from 1922 to 2016. Hence, using (6.2) we can calculate the actuarial survival probability estimates directly from the data. Now looking at the 2014 cohort we use (6.2) to obtain ${}_tp_{65}$. Setting the interest rate $i = 5\%$ per annum, we calculate the whole of life annuity using (6.8). Using (6.4) we can calculate e_{65}, from figure 16 we observe that at age 65, men are expected to live until age 76.55 and women are expected to live until age 80.15.



```
At age 65, the value of the annuity using the actuarial estimate is  9.75. The life
expectancy of a male life aged 65 in 2014 is 11.55

At age 65, the value of the annuity using the actuarial estimate is  10.56. The life
expectancy of a female life aged 65 in 2014 is 15.15
```

Figure 16: The R output showing the future life expectancy and the value of the annuity for male and females aged 65, using the actuarial estimate.

7.3.2 Makeham Model

Recall from table 2, the mortality rate for the Makeham model is given by

$$\mu_x = \gamma + \lambda e^{\theta x}$$

For this data set we are working discrete age x (number of whole lives lived) therefore, as stated in chapter 6, the force of mortality, μ_x, can be approximated by the central death rate, m_x. Hence, $m_x \approx \gamma + \lambda e^{\theta x}$. In R we fit the model with the appropriate parameters for both males and females. For males, R fits the model with parameters,

$$\gamma = 1.1153469 \times 10^{-4} \qquad \lambda = 2.694927 \times 10^{-5} \qquad \theta = 9.445495 \times 10^{-2}$$

For females we have the following parameters,

$$\gamma = 9.794468 \times 10^{-5} \qquad \lambda = 1.047133 \times 10^{-5} \qquad \theta = 1.02319 \times 10^{-2}$$

From figure 17 we can see that the Makeham model works best for adult mortality hence it is suitable to use as we are discussing pensions with benefits being paid from age 65.

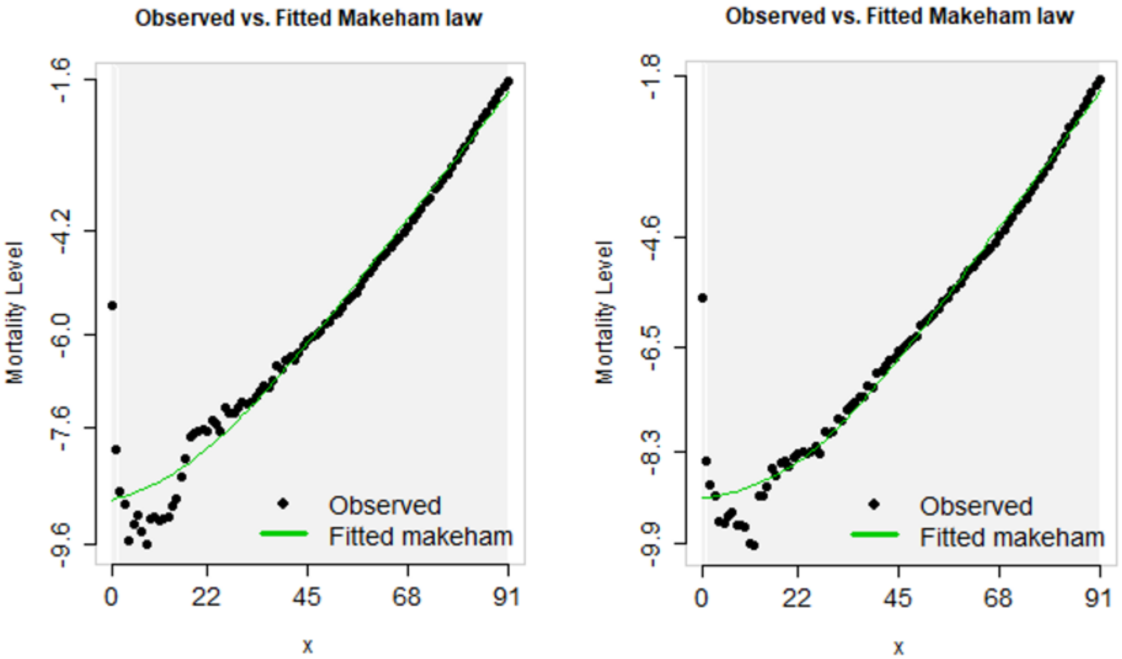
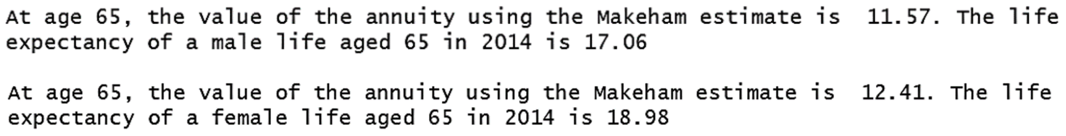


Figure 17: Charts showing the fitted Makeham model for males (left) and females (right) in 2014.

By creating a life table for the 2014 cohort in R, we can use the l_x column to calculate ${}_tp_x$ using (6.2). Again, with interest rate $i = 5\%$ per annum, $\ddot{a}_{65}$ can be calculated using (6.8). From figure 18 we observe that at age 65, men are expected to live until age 82.06 and women are expected to live until age 83.98.



```
At age 65, the value of the annuity using the Makeham estimate is  11.57. The life
expectancy of a male life aged 65 in 2014 is 17.06

At age 65, the value of the annuity using the Makeham estimate is  12.41. The life
expectancy of a female life aged 65 in 2014 is 18.98
```

Figure 18: The R output showing the future life expectancy and the value of the annuity for male and females aged 65, using the Makeham model.

7.3.3 Analysis

Nationally, lives aged 55-64 in 2014, are expected to live until age 86 and 89 for men and women respectively [10]. We have chosen to analyse the default retirement age of 65 as chosen in many longevity risk studies, such as the paper, Longevity Risk and Annuity Pricing with the Lee–Carter Model [12] and the journal, Longevity Bonds: Financial Engineering, Valuation and Hedging [16]. Given that 65 is close to the 55-64 age range, we will compare the model results to the actual life expectancy ages of 86 for males and 89 for females.

Model	Gender	Expected age of death (years)	$\ddot{a}_{65}$
Actuarial	Male	76.55	9.75
	Female	80.15	10.56
Makeham	Male	82.06	11.57
	Female	83.98	12.41
Lee–Carter	Male	87.50	13.78
	Female	89.43	14.53

Table 7: A table summarizing $\ddot{a}_{65}$ and the expected age of death for males and females aged 65 using the Actuarial, Makeham and Lee–Carter models.

Notice, from table 7, that the Makeham and actuarial model largely underestimate life expectancy. The actuarial model underestimates life expectancy by 9.45 years for males and 8.85 years for females.

The Makeham model produces better estimates than the actuarial model, with the greatest improvement for males. However, it still considerably underestimates life expectancy with male life expectancy underestimated by 3.94 years and female life expectancy underestimated by 5.02 years.

Now looking at the 2014 cohort for the Lee–Carter model, we can see that the life expectancies are the closest to the actual life expectancies compared to the previous two models. The estimates are slightly overestimated by 1.27 years and 0.24 years for males and females respectively. Hence, the Lee–Carter model provides better estimations of life expectancy.

The Lee–Carter model gives a larger annuity value than the Makeham and actuarial models. Hence, using (6.11) the annual benefit the insurance company pays the policyholder (the pension) is therefore smaller. As a result, the insurance company will pay out less money per year as the policyholder is expected to live longer. Therefore, the longevity risk is reduced.

8 Discussion

Throughout this study we have discussed various principles within survival analysis, starting with introducing the survival function itself and then exploring other concepts such as parametric estimators, non-parametric estimators and the log-rank test.
We have also seen the applications of survival analysis in actuarial science, specifically investigating longevity risk. As seen in chapter 7, the Lee–Carter model estimates life expectancy more accurately than the actuarial and the Makeham estimate and leads to insurance companies paying out smaller annuities each year for a whole of life annuity. However, many techniques have been used to further improve the Lee–Carter model. One being the Delwarde, Denuit & Eilers (2007) (DDE) model [24]. Delwarde, Denuit & Eilers noticed that there are irregularities in the parameter b_x, for example consider the jump in b_x from age 46 to age 47 in the figure below. To solve this, the DDE model smoothens b_x, as shown in the figure below.

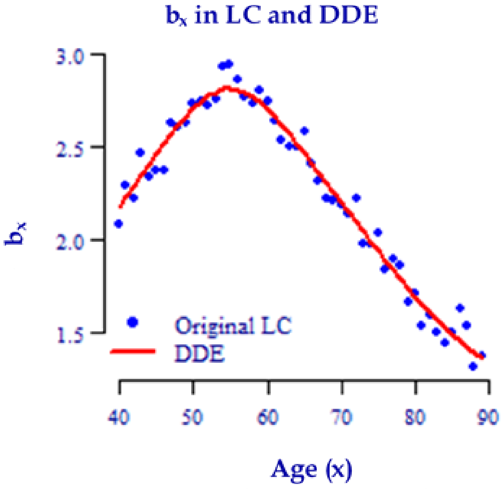


Figure 19: A plot of the b_x estimates for the Lee–Carter model (LC) and the DDE model. Adapted from [25].

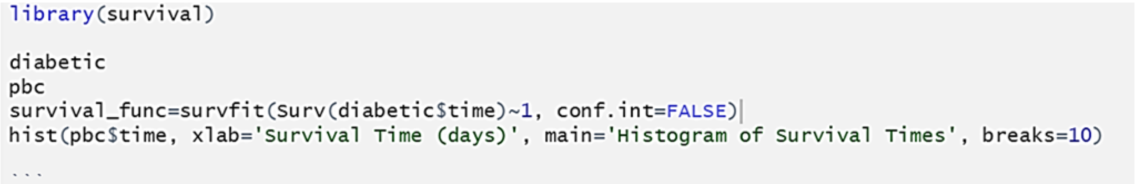
The Currie-Richards (CR) model [12] goes on to smooth the parameter k_t also. Further analysis can be conducted to understand how the mortality rate projections of each these models in the 'Lee–Carter family' affect financial risk calculations.

There are other areas of study to consider further. As discussed previously, life expectancy is increasing. Is it possible for human life to be unlimited? “In western countries and Japan and after age 110 the risk of dying is constant and is about 47% per year” [26]. Therefore, there is no maximum age for human life. One could consider longevity risk particularly for very old age and how this affects policyholders and insurance companies when it comes to pensions.

Although life expectancy is generally on the rise, certain events can counteract this such as epidemics, climate change and war. For example, in figure 11 the log death rate increases much more rapidly for young adult males around the early 1940s (the yellow lines). This is likely to be due to male soldiers dying in battle during World War II (1939-1945).

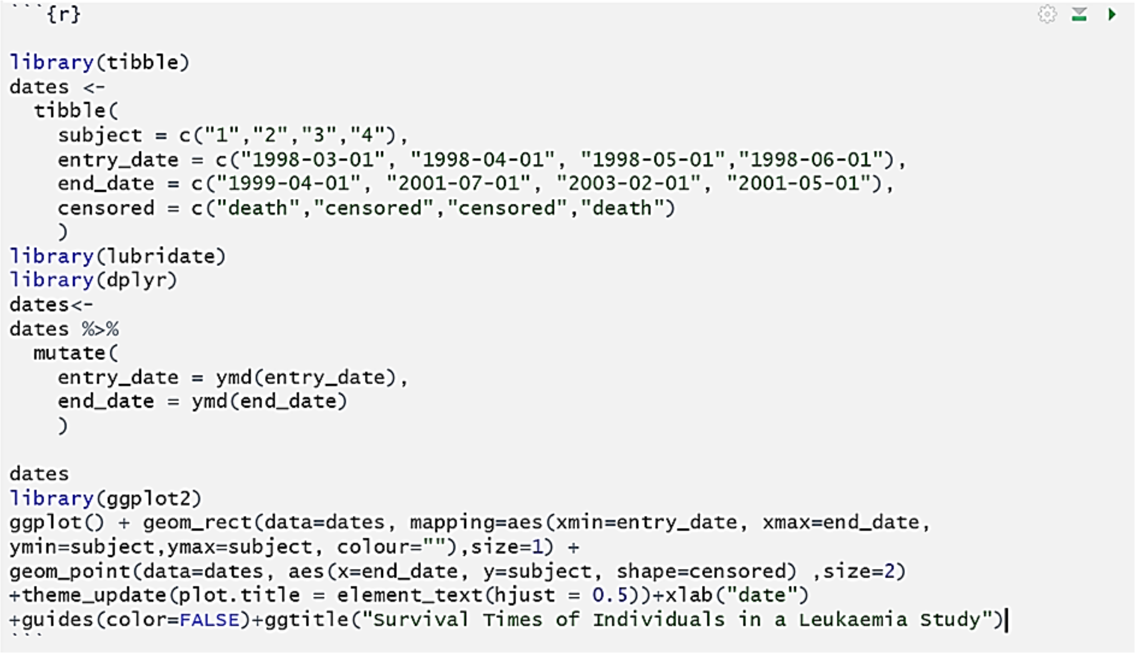
As a result, life expectancy cannot always be predicted accurately as various factors can drastically affect the mortality rate. Despite this, the Lee–Carter model, and other models in the Lee–Carter family, can be used to make more accurate predictions in general, reducing longevity risk.

9 Appendix



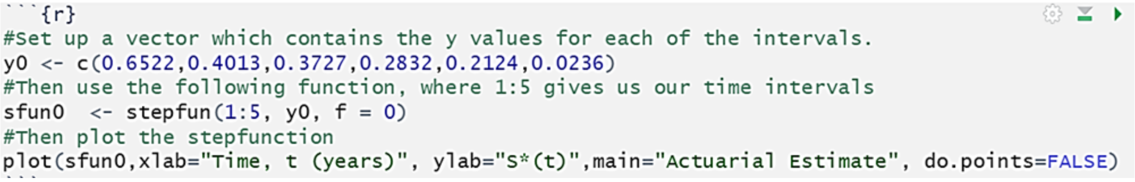
~~~
library(survival)

diabetic
pbc
survival_func=survfit(Surv(diabetic$time)~1, conf.int=FALSE)
hist(pbc$time, xlab='Survival Time (days)', main='Histogram of Survival Times', breaks=10)

```
~~~

Figure 20: R code for Figure 1.



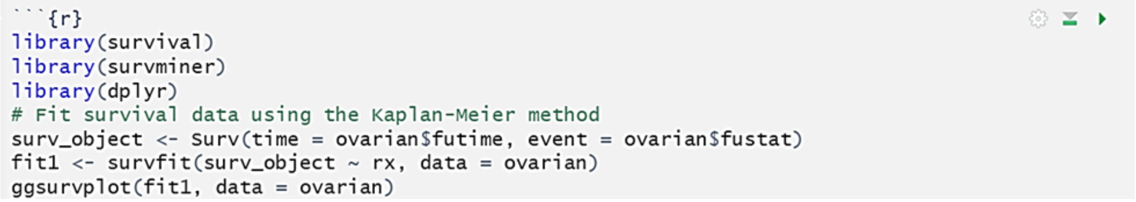
~~~
```{r}

library(tibble)
dates <-
  tibble(
    subject = c("1","2","3","4"),
    entry_date = c("1998-03-01", "1998-04-01", "1998-05-01","1998-06-01"),
    end_date = c("1999-04-01", "2001-07-01", "2003-02-01", "2001-05-01"),
    censored = c("death","censored","censored","death")
    )
library(lubridate)
library(dplyr)
dates<-
dates %>%
  mutate(
    entry_date = ymd(entry_date),
    end_date = ymd(end_date)
    )

dates
library(ggplot2)
ggplot() + geom_rect(data=dates, mapping=aes(xmin=entry_date, xmax=end_date,
ymin=subject,ymax=subject, colour=""),size=1) +
geom_point(data=dates, aes(x=end_date, y=subject, shape=censored) ,size=2)
+theme_update(plot.title = element_text(hjust = 0.5))+xlab("date")
+guides(color=FALSE)+ggtitle("Survival Times of Individuals in a Leukaemia Study")
```
~~~

Figure 21: R for code Figure 2.



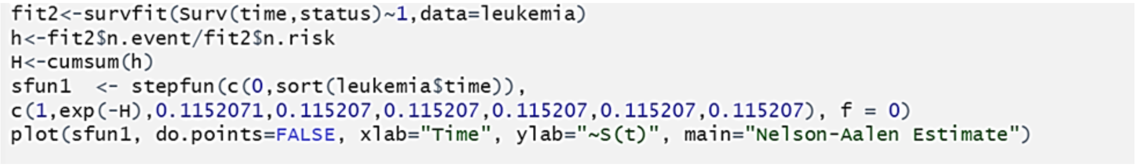
~~~
```{r}
#Set up a vector which contains the y values for each of the intervals.
y0 <- c(0.6522,0.4013,0.3727,0.2832,0.2124,0.0236)
#Then use the following function, where 1:5 gives us our time intervals
sfun0  <- stepfun(1:5, y0, f = 0)
#Then plot the stepfunction
plot(sfun0,xlab="Time, t (years)", ylab="S*(t)",main="Actuarial Estimate", do.points=FALSE)
```
~~~

Figure 22: R for code Figure 3.



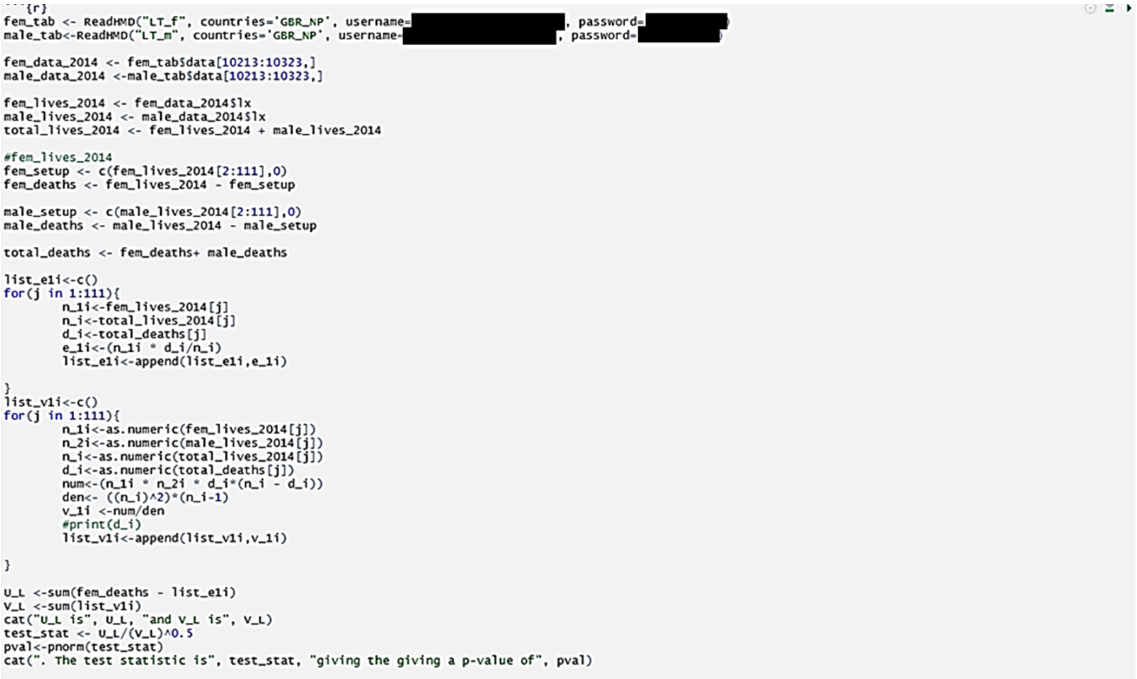
~~~
```{r}
library(survival)
library(survminer)
library(dplyr)
# Fit survival data using the Kaplan-Meier method
surv_object <- Surv(time = ovarian$futime, event = ovarian$fustat)
fit1 <- survfit(surv_object ~ rx, data = ovarian)
ggsurvplot(fit1, data = ovarian)
~~~

Figure 23: Figure 5.



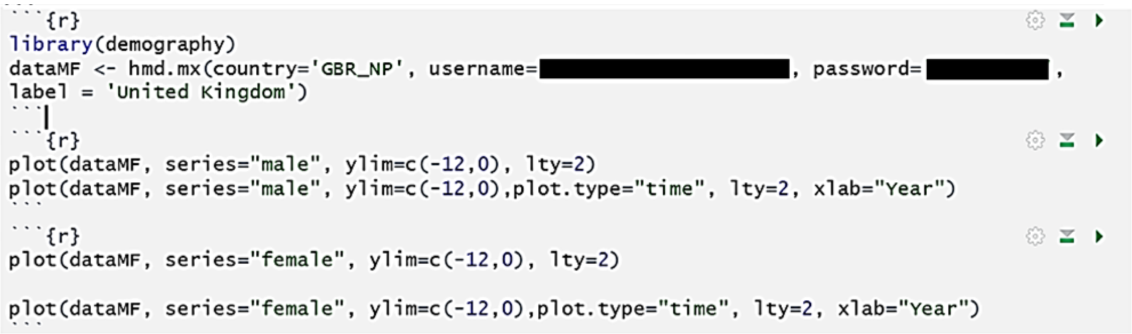
```
fit2<-survfit(Surv(time,status)~1,data=leukemia)
h<-fit2$n.event/fit2$n.risk
H<-cumsum(h)
sfun1  <- stepfun(c(0,sort(leukemia$time)),
c(1,exp(-H),0.1152071,0.115207,0.115207,0.115207,0.115207,0.115207), f = 0)
plot(sfun1, do.points=FALSE, xlab="Time", ylab="~S(t)", main="Nelson-Aalen Estimate")
```

Figure 24: R code for Figure 6 [9].



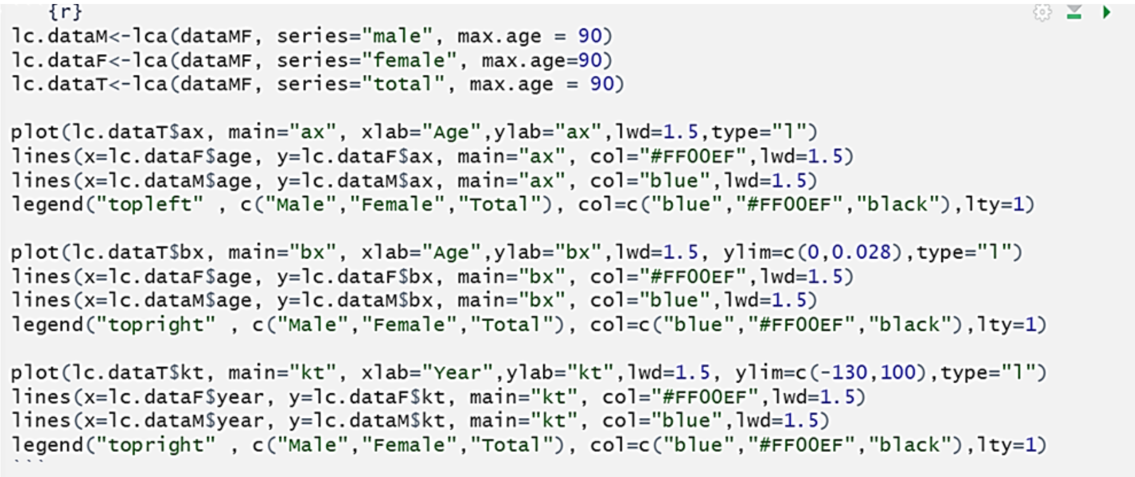
~~~
```{r}
fem_tab <- ReadHMD("LT_f", countries='GBR_NP', username=[redacted], password=[redacted])
male_tab<-ReadHMD("LT_m", countries='GBR_NP', username=[redacted], password=[redacted])

fem_data_2014 <- fem_tab$data[10213:10323,]
male_data_2014 <-male_tab$data[10213:10323,]

fem_lives_2014 <- fem_data_2014$lx
male_lives_2014 <- male_data_2014$lx
total_lives_2014 <- fem_lives_2014 + male_lives_2014

#fem_lives_2014
fem_setup <- c(fem_lives_2014[2:111],0)
fem_deaths <- fem_lives_2014 - fem_setup

male_setup <- c(male_lives_2014[2:111],0)
male_deaths <- male_lives_2014 - male_setup

total_deaths <- fem_deaths+ male_deaths

list_e1i<-c()
for(j in 1:111){
        n_1i<-fem_lives_2014[j]
        n_i<-total_lives_2014[j]
        d_i<-total_deaths[j]
        e_1i<-(n_1i * d_i/n_i)
        list_e1i<-append(list_e1i,e_1i)

}
list_v1i<-c()
for(j in 1:111){
        n_1i<-as.numeric(fem_lives_2014[j])
        n_2i<-as.numeric(male_lives_2014[j])
        n_i<-as.numeric(total_lives_2014[j])
        d_i<-as.numeric(total_deaths[j])
        num<-(n_1i * n_2i * d_i*(n_i - d_i))
        den<- ((n_i)^2)*(n_i-1)
        v_1i <-num/den
        #print(d_i)
        list_v1i<-append(list_v1i,v_1i)

}

U_L <-sum(fem_deaths - list_e1i)
V_L <-sum(list_v1i)
cat("U_L is", U_L, "and V_L is", V_L)
test_stat <- U_L/(V_L)^0.5
pval<-pnorm(test_stat)
cat(". The test statistic is", test_stat, "giving the giving a p-value of", pval)
~~~

Figure 25: R code for Figure 7.



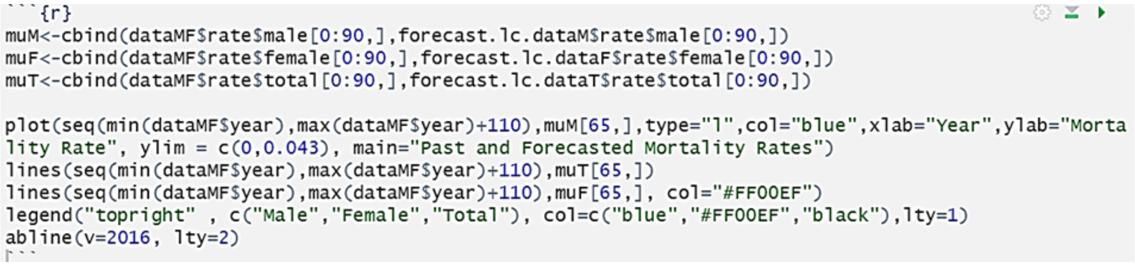
~~~
```{r}
library(demography)
dataMF <- hmd.mx(country='GBR_NP', username=[redacted], password=[redacted],
label = 'United Kingdom')
```
```{r}
plot(dataMF, series="male", ylim=c(-12,0), lty=2)
plot(dataMF, series="male", ylim=c(-12,0),plot.type="time", lty=2, xlab="Year")
```
```{r}
plot(dataMF, series="female", ylim=c(-12,0), lty=2)

plot(dataMF, series="female", ylim=c(-12,0),plot.type="time", lty=2, xlab="Year")
```
~~~

Figure 26: R code for Figure 8, 9, 10 and 11.



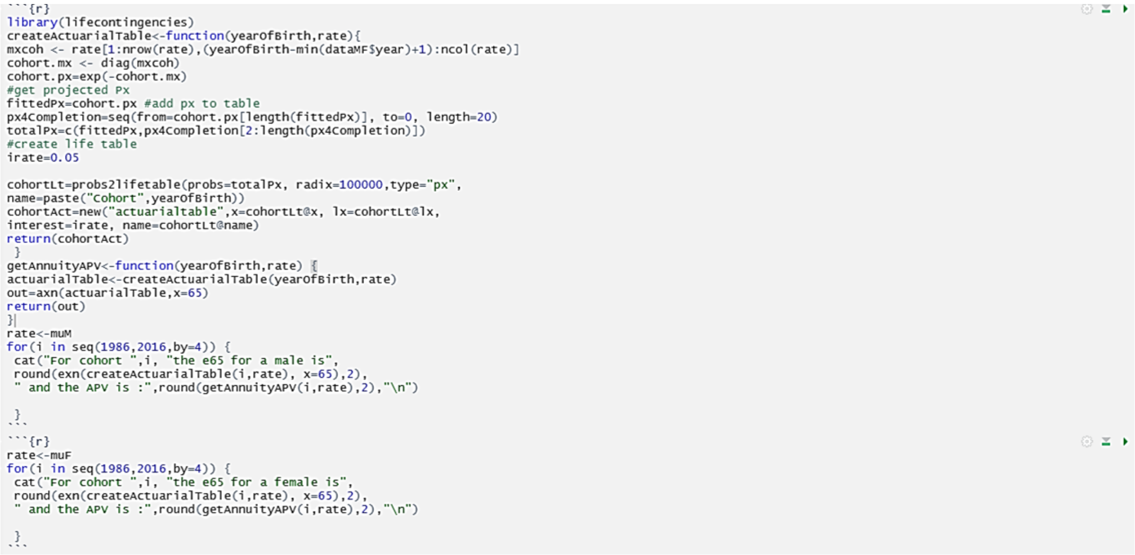
~~~
```{r}
lc.dataM<-lca(dataMF, series="male", max.age = 90)
lc.dataF<-lca(dataMF, series="female", max.age=90)
lc.dataT<-lca(dataMF, series="total", max.age = 90)

plot(lc.dataT$ax, main="ax", xlab="Age",ylab="ax",lwd=1.5,type="l")
lines(x=lc.dataF$age, y=lc.dataF$ax, main="ax", col="#FF00EF",lwd=1.5)
lines(x=lc.dataM$age, y=lc.dataM$ax, main="ax", col="blue",lwd=1.5)
legend("topleft" , c("Male","Female","Total"), col=c("blue","#FF00EF","black"),lty=1)

plot(lc.dataT$bx, main="bx", xlab="Age",ylab="bx",lwd=1.5, ylim=c(0,0.028),type="l")
lines(x=lc.dataF$age, y=lc.dataF$bx, main="bx", col="#FF00EF",lwd=1.5)
lines(x=lc.dataM$age, y=lc.dataM$bx, main="bx", col="blue",lwd=1.5)
legend("topright" , c("Male","Female","Total"), col=c("blue","#FF00EF","black"),lty=1)

plot(lc.dataT$kt, main="kt", xlab="Year",ylab="kt",lwd=1.5, ylim=c(-130,100),type="l")
lines(x=lc.dataF$year, y=lc.dataF$kt, main="kt", col="#FF00EF",lwd=1.5)
lines(x=lc.dataM$year, y=lc.dataM$kt, main="kt", col="blue",lwd=1.5)
legend("topright" , c("Male","Female","Total"), col=c("blue","#FF00EF","black"),lty=1)
```
~~~

Figure 27: R code for Figure 12, 13 and 14. Adapted from [6].



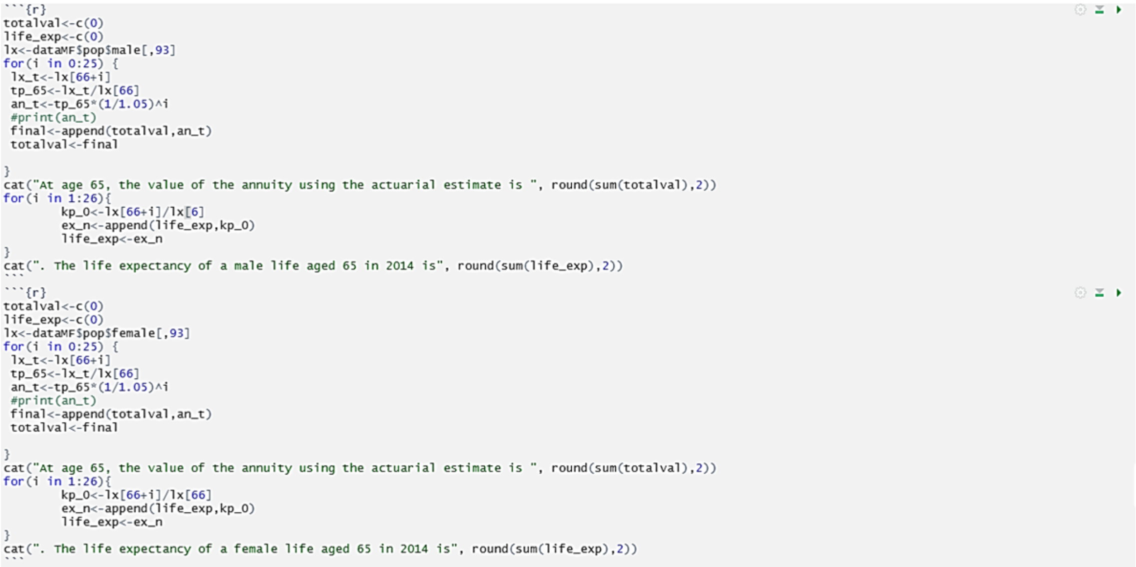
~~~
```{r}
muM<-cbind(dataMF$rate$male[0:90,],forecast.lc.dataM$rate$male[0:90,])
muF<-cbind(dataMF$rate$female[0:90,],forecast.lc.dataF$rate$female[0:90,])
muT<-cbind(dataMF$rate$total[0:90,],forecast.lc.dataT$rate$total[0:90,])

plot(seq(min(dataMF$year),max(dataMF$year)+110),muM[65,],type="l",col="blue",xlab="Year",ylab="Morta
lity Rate", ylim = c(0,0.043), main="Past and Forecasted Mortality Rates")
lines(seq(min(dataMF$year),max(dataMF$year)+110),muT[65,])
lines(seq(min(dataMF$year),max(dataMF$year)+110),muF[65,], col="#FF00EF")
legend("topright" , c("Male","Female","Total"), col=c("blue","#FF00EF","black"),lty=1)
abline(v=2016, lty=2)
```
~~~

Figure 28: R code for Figure 15. Adapted from [6].



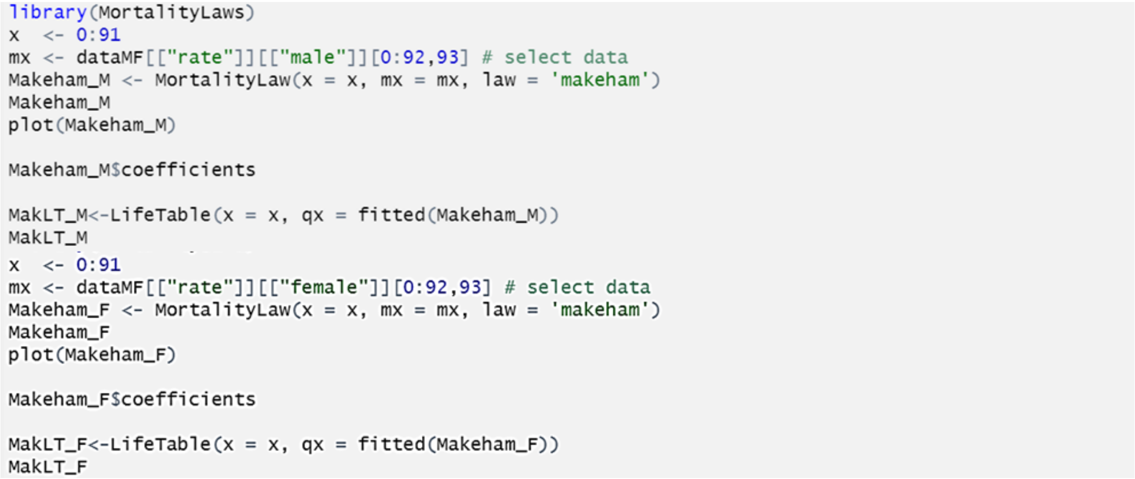
~~~
```{r}
library(lifecontingencies)
createActuarialTable<-function(yearOfBirth,rate){
mxcoh <- rate[1:nrow(rate),(yearOfBirth-min(dataMF$year)+1):ncol(rate)]
cohort.mx <- diag(mxcoh)
cohort.px=exp(-cohort.mx)
#get projected Px
fittedPx=cohort.px #add px to table
px4Completion=seq(from=cohort.px[length(fittedPx)], to=0, length=20)
totalPx=c(fittedPx,px4Completion[2:length(px4Completion)])
#create life table
irate=0.05

cohortLt=probs2lifetable(probs=totalPx, radix=100000,type="px",
name=paste("Cohort",yearOfBirth))
cohortAct=new("actuarialtable",x=cohortLt@x, lx=cohortLt@lx,
interest=irate, name=cohortLt@name)
return(cohortAct)
 }
getAnnuityAPV<-function(yearOfBirth,rate) {
actuarialTable<-createActuarialTable(yearOfBirth,rate)
out=axn(actuarialTable,x=65)
return(out)
}
rate<-muM
for(i in seq(1986,2016,by=4)) {
 cat("For cohort ",i, "the e65 for a male is",
 round(exn(createActuarialTable(i,rate), x=65),2),
 " and the APV is :",round(getAnnuityAPV(i,rate),2),"\n")

 }
```
```{r}
rate<-muF
for(i in seq(1986,2016,by=4)) {
 cat("For cohort ",i, "the e65 for a female is",
 round(exn(createActuarialTable(i,rate), x=65),2),
 " and the APV is :",round(getAnnuityAPV(i,rate),2),"\n")

 }
```
~~~

Figure 29: R code for Table 5 and Table 6. Adapted from [6].



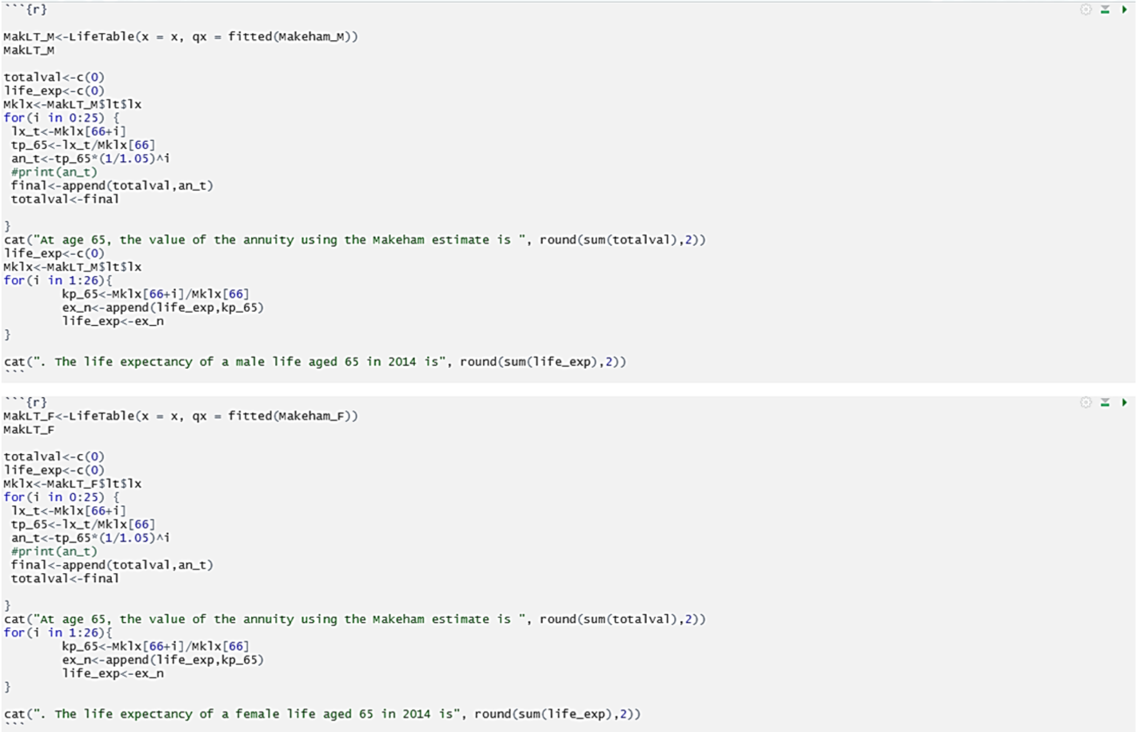
~~~
```{r}
totalval<-c(0)
life_exp<-c(0)
lx<-dataMF$pop$male[,93]
for(i in 0:25) {
 lx_t<-lx[66+i]
 tp_65<-lx_t/lx[66]
 an_t<-tp_65*(1/1.05)^i
 #print(an_t)
 final<-append(totalval,an_t)
 totalval<-final

}
cat("At age 65, the value of the annuity using the actuarial estimate is ", round(sum(totalval),2))
for(i in 1:26){
        kp_0<-lx[66+i]/lx[6]
        ex_n<-append(life_exp,kp_0)
        life_exp<-ex_n
}
cat(". The life expectancy of a male life aged 65 in 2014 is", round(sum(life_exp),2))
```
```{r}
totalval<-c(0)
life_exp<-c(0)
lx<-dataMF$pop$female[,93]
for(i in 0:25) {
 lx_t<-lx[66+i]
 tp_65<-lx_t/lx[66]
 an_t<-tp_65*(1/1.05)^i
 #print(an_t)
 final<-append(totalval,an_t)
 totalval<-final

}
cat("At age 65, the value of the annuity using the actuarial estimate is ", round(sum(totalval),2))
for(i in 1:26){
        kp_0<-lx[66+i]/lx[66]
        ex_n<-append(life_exp,kp_0)
        life_exp<-ex_n
}
cat(". The life expectancy of a female life aged 65 in 2014 is", round(sum(life_exp),2))
```
~~~

Figure 30: R code for Figure 16.



````
library(MortalityLaws)
x  <- 0:91
mx <- dataMF[["rate"]][["male"]][0:92,93] # select data
Makeham_M <- MortalityLaw(x = x, mx = mx, law = 'makeham')
Makeham_M
plot(Makeham_M)

Makeham_M$coefficients

MakLT_M<-LifeTable(x = x, qx = fitted(Makeham_M))
MakLT_M
x  <- 0:91
mx <- dataMF[["rate"]][["female"]][0:92,93] # select data
Makeham_F <- MortalityLaw(x = x, mx = mx, law = 'makeham')
Makeham_F
plot(Makeham_F)

Makeham_F$coefficients

MakLT_F<-LifeTable(x = x, qx = fitted(Makeham_F))
MakLT_F
````

Figure 31: R code for Figure 17.



~~~
```{r}

MakLT_M<-LifeTable(x = x, qx = fitted(Makeham_M))
MakLT_M

totalval<-c(0)
life_exp<-c(0)
Mklx<-MakLT_M$lt$lx
for(i in 0:25) {
 lx_t<-Mklx[66+i]
 tp_65<-lx_t/Mklx[66]
 an_t<-tp_65*(1/1.05)^i
 #print(an_t)
 final<-append(totalval,an_t)
 totalval<-final

}
cat("At age 65, the value of the annuity using the Makeham estimate is ", round(sum(totalval),2))
life_exp<-c(0)
Mklx<-MakLT_M$lt$lx
for(i in 1:26){
        kp_65<-Mklx[66+i]/Mklx[66]
        ex_n<-append(life_exp,kp_65)
        life_exp<-ex_n
}

cat(". The life expectancy of a male life aged 65 in 2014 is", round(sum(life_exp),2))
```
~~~

~~~
```{r}
MakLT_F<-LifeTable(x = x, qx = fitted(Makeham_F))
MakLT_F

totalval<-c(0)
life_exp<-c(0)
Mklx<-MakLT_F$lt$lx
for(i in 0:25) {
 lx_t<-Mklx[66+i]
 tp_65<-lx_t/Mklx[66]
 an_t<-tp_65*(1/1.05)^i
 #print(an_t)
 final<-append(totalval,an_t)
 totalval<-final

}
cat("At age 65, the value of the annuity using the Makeham estimate is ", round(sum(totalval),2))
for(i in 1:26){
        kp_65<-Mklx[66+i]/Mklx[66]
        ex_n<-append(life_exp,kp_65)
        life_exp<-ex_n
}

cat(". The life expectancy of a female life aged 65 in 2014 is", round(sum(life_exp),2))
```
~~~

Figure 32: R code for Figure 18.

References

[1] Collett, D. *Modelling Survival Data in Medical Research Third edition.* Boca Raton: CRC Press, Taylor & Francis Group, 2015. [Book].

[2] Cox, D.R., *Analysis of Survival Data.* First edition. CRC Press, 2018. [Book].

[3] Mills, M., *Introducing Survival Analysis and Event History Analysis.* London: Sage, 2011. [Book].

[4] Richards, S.J., *A Handbook of Parametric Survival Models for Actuarial Use.* Scandinavian Actuarial Journal, 2012, no. 4: 233–257. [Article] [http://www.tandfonline.com/doi/abs/10.1080/03461238.2010.506688.] [Accessed 6 November 2019].

[5] Machin, D., Cheung, Y.B., and Parmar, M.K.B., *Survival Analysis: a Practical Approach Second edition.* Chichester: Wiley, 2006. [Book].

[6] Spedicato, A.G., [Article] [https://cran.r-project.org/web/packages/lifecontingencies /vignettes/mortality_projection.pdf] [Accessed 23 January 2020].

[7] Germán, R., *Generalised Linear Models*, Princeton University Lecture Notes, 2016. [Website] [https://data.princeton.edu/wws509/notes/c7s1] [Accessed 23 January 2020] .

[8] Germán R., *Non-Parametric Estimation in Survival Models*, Princeton University Lecture Notes, 2005. [Website] [https://data.princeton.edu/wws509/notes/c7s1] [Accessed 28 January 2020].

[9] Schütte D., Survival analysis in R for beginners, DataCamp, 2019. [Website] [https://www.datacamp.com/community/tutorials/survival-analysis-R] [Accessed 30 January 2020].

[10] Institute and Faculty of Actuaries, 2015. [Article] [https://www.actuaries.org.uk/system/files/field/document/Policy%20summary% 20-%20Longevity%20Risk%20A4%20V02% 20WEB.PDF] [Accessed 1 February 2020].

[11] Charpentier A., *Computational Actuarial Science with R*, 2014, [Book].

[12] Richards S. J. and Currie I. D., Longevity Risk and Annuity Pricing with the Lee-Carter Model, 2009. [Article] [https://www.actuaries.org.uk/system/files/documents/pdf/facsm20090216.pdf] [Accessed 1 February 2020].

[13] Thatcher A. R., Kannisto V., & Vaupel J. W., *The Force of Mortality at Ages 80 to 120*, Odense University Press, 1998. [Book].

[14] Murphy, G., MATH3510: Actuarial Mathematics 1, University of Leeds Lecture Notes, [Website] [Accessed 11 February 2020].

[15] Human Mortality Database. [Website] [http://www.mortality.org/] [Accessed 24 January 2020].

[16] Blake, D. et al., *Longevity Bonds: Financial Engineering, Valuation and Hedging*, Journal of Risk and Insurance 73, 647-672, American Risk and Insurance Foundation, 2006. [Book].

[17] The Institute for Health Metrics and Evaluation, 2014. [Article] [http://www.healthdata.org/news-release/life-expectancy-increases-globally-death-toll-falls-major-diseases] [Accessed 14 February 2020].

[18] Glossary, Club Vita. [Website] [https://clubvita.net/glossary/accident-hump] [Accessed 14 February 2020].

[19] Wilmoth, J.R. et al, Methods Protocol for the Human Mortality Database, 2019. [Article] [https://www.mortality.org/Public/Docs/MethodsProtocol.pdf] [Accessed 23 February 2020].

[20] Lee, R., & Carter, L., (1992), Modeling and Forecasting U. S. Mortality, Journal of the American Statistical Association, 87(419), 659-671. [Article] [Accessed 6 March 2020].

[21] Nigri, A., Levantesi S., Marino M., Scognamiglio S., and Perla, F., A Deep Learning Integrated Lee–Carter Model, 2019. [Article] [https://www.mdpi.com/2227-9091/7/1/33/pdf.] [Accessed 6 March 2020].

[22] Girosi, F., King, G., Understanding the Lee-Carter Mortality Forecasting Method, 2007. [Article] [https://gking.harvard.edu/files/gking/files/lc.pdf] [Accessed 6 March 2020].

[23] Guide to calculating national life tables, Office for National Statistics, 2019. [Website] [https://www.ons.gov.uk/peoplepopulationa ndcommunity/healthandsocialcare/healthandlifeexpectancies/methodologies/guidetocalculatingnationallifetables] [Accessed 6 March 2020].

[24] Delwarde, A., Denuit, M., & Eilers, P., Smoothing the Lee–Carter and Poisson log-bilinear models for mortality forecasting: A penalized log-likelihood approach, Statistical Modelling, 7(1), 29–48, 2007. [Article] [https://doi.org/10.1177/1471082X0600700103] [Accessed 7 March 2020].

[25] The Lee–Carter Family, Longevitas. [Website] [https://www.longevitas.co.uk/site/informationmatrix/thelee carterfamily.html] [Accessed 7 March 2020].

[26] Rootzén, H., Zholud, D., Human life is unlimited – but short. Extremes 20, 713–728, 2017. [Article] [https://doi.org/10.1007/s10687-017-0305-5] [Accessed 8 March 2020].