PART ONE

ANALYSING PORTFOLIO MORTALITY

1
Introduction

1.1 Survival Data

This part of the book is about the statistical analysis and modelling of survival data. The purpose we usually have in mind is the pricing or valuation of some insurance contract whose payments are *contingent* on the death or survival of an individual. So, our starting point is the question: what form does survival data take?

1.1.1 Examples of Survival Data

Consider the following two examples:

(i) On 1 January 2014, Mr Brown took out a life insurance policy. The premium he paid took into account his age on that date (he was exactly 31 years and three months old) and the fact that he had never smoked cigarettes. On 19 April 2017 (the date when this is being written) Mr Brown was still alive.

(ii) On 23 September 2013, Ms Green reached her 60th birthday and retired. She used her pensions savings on that date to purchase an annuity. Unfortunately, her health was poor and the annual amount of annuity was higher than normal for that reason. The annuity ceased when she died on 3 April 2016.

These observations, typical of what may be extracted from the files of an insurance company or pension scheme, illustrate the raw material of survival analysis, as actuaries practise it. We can list some features, all of which may be relevant to the subsequent analysis:

- There are three *timescales* in each example, namely age, calendar time and the duration since the life insurance policy or annuity commenced.

- Our observations began only when the insurance policy or annuity commenced. Before that time we had no reason to know of Mr Brown's or Ms Green's existence. All we know now is that they were alive on the relevant commencement dates.
- Observation of Mr Brown ceased when this account was written on 19 April 2017, at which time he was still alive. We know that he will die after 19 April 2017, but we do not know when.
- Observation of Ms Green ceased because she died while under observation (after 23 September 2013 but before 19 April 2017).
- In both cases, additional information was available that influenced the price of the financial contract: age, gender, Mr Brown's non-smoking status and Ms Green's poor health. Clearly, these data influenced the pricing because they tell us something about a person's chances of dying sooner rather than later.

1.1.2 Individual Life Histories

The key features of life history data can be summarised as follows:

- The age at starting observation, the date of starting observation and the reason for starting observation.
- The age at ending observation, the date of ending observation and the reason for ending observation.
- Any additional information, such as gender, benefit amount or health status.

1.1.3 Grouped Survival Data

One main purpose of this book is to describe statistical models of mortality that use, directly, data like the examples above. This is a destination, not a starting point. We will soon introduce the idea of representing the future lifetime of an indvidual as a non-negative random variable T. Ordinary statistical analysis proceeds by observing some number n of observations $t_1, t_2, \ldots, t_n$ drawn from the distribution of T. A key assumption is that these are independent and identically distributed (i.i.d.). In the case of Mr Brown and Ms Green, we have no reason to doubt independence, but they are clearly not identically distributed.

So we take a step back, and ask how we can define statistics derived from the life histories described above that are plausibly i.i.d.. One way is to group data according to qualities that advance homogeneity and reduce heterogeneity. For example, we could group data by the following qualities:

- age
- gender
- policy size (sum assured or annuity payment)
- type of insurance policy
- calendar time
- duration since taking out insurance policy
- smoking status
- occupation
- medical history.

Another way is to propose a statistical model which incorporates directly any important sources of heterogeneity, for example as covariates in a regression model. In Chapter 7 we discuss the relative merits of these two approaches.

1.2 Software

Throughout this book we will illustrate basic model-fitting with the freely available R software package. This is both a programming language and a statistical analysis package, and it has become a standard for academic and scientific work. R is free to download and use; basic instructions for downloading and installing R can be found in Appendix A. Partly because it is free of charge, R comes with no warranties. However, support is available in a number of online forums.

Many actuaries in commerce use Microsoft Excel®, and they may ask why we do not use this (or any other spreadsheet) for model-fitting. The answer is twofold. First, R has many advantages, not least the vast libraries of scientific functions to call upon which mean we can often fit complex models with a few lines of code. Second, there are some important limits to Excel, especially when it comes to fitting projection models like those in Part Two. Some of these limits are rather subtle, so it is important that an analyst is aware of Excel's limitations.

The first issue is that at the time of writing Excel's standard Solver feature will not work with more than 200 variables (that is, parameters which have to be optimised in order to fit the model). This is a problem for a number of important stochastic projection models in Part Two. One option is to use only models with fewer than 200 parameters, but this would allow software limitations to dictate what the analyst can do.

Another issue is that Excel's Solver function will often claim an optimal solution has been found when this is not the case. If the Solver is re-run several times in succession, it often finds a better-fitting set of parameters on the second and third attempts. It is therefore important that the analyst re-runs the Solver a few times until no further change is found. Even then, we have come across examples where R found a better-fitting set of parameters, which the Solver agreed was a better fit, but which the Solver could not find on its own.

One option would be to consider one of the commercially supported alternative plug-ins for Excel's Solver, although analysts would need to check that it was indeed capable of finding the solutions that Excel cannot.

Whatever the analyst does, it is important not to rely uncritically on a single software implementation without some form of checking.

1.3 Grouped Counts

Consider Table 1.1, which shows the mortality-experience data for the UK pension scheme in the Case Study (see Section 1.10 for a fuller description). It shows the number of deaths and time lived in ten-year age bands for males and females combined. The main advantage of the data format in Table 1.1 is its simplicity. The entire human age span is represented by just 11 data points (age bands), and a reasonably well-specified statistical model can be fitted in just four R statements (more on this in Section 1.5). We call the data in Table 1.1 *grouped data*, because there is no information on individuals. (It is likely that information on individuals was collected, but then aggregated. The analyst might not have access to the data originally collected, only to some summarised form.) A natural and intuitive measure of mortality in each age band is the ratio of the number of deaths to the total time lived, which is shown in the last column of Table 1.1. We call quantities of this form *mortality ratios*.

1.4 What Mortality Ratio Should We Analyse?

Suppose in a mortality analysis we want to calculate mortality ratios, as in the rightmost column of Table 1.1. The numerator for the mortality ratio is obvious: it is the number of deaths which have occurred. However, we have two fundamental choices for the denominator:

- the number of lives (which is not shown in Table 1.1), or
- the time lived by those lives.

Table 1.1 *High-level mortality data for Case Study (see Section 1.10); time lived and deaths in 2007–2012.*

Age interval	Time lived, t (years)	Deaths, d	Mortality ratio ($d/t \times 1000$)
[0, 10)	71.9	0	0
[10, 20)	449.0	2	4.5
[20, 30)	163.9	0	0
[30, 40)	121.7	3	24.6
[40, 50)	893.1	6	6.7
[50, 60)	5,079.3	48	9.5
[60, 70)	32,546.7	278	8.5
[70, 80)	21,155.9	510	24.1
[80, 90)	10,606.7	866	81.6
[90, 100)	1,751.5	363	207.3
[100, ∞)	23.1	11	475.7
All ages	72,862.7	2,087	28.6

The distinction arises because some of the individuals in the study may not have been present for the whole period 2007–2012. For example, consider someone who retired on 1 January 2009. Such a person would contribute one to the total number of lives, but a maximum of four out of a possible six years of time lived while a member of the scheme. The methods needed to analyse these alternative formulations will clearly be different.

If we use the number of lives as the denominator, we are calculating the proportion dying. For example, suppose a total of 3,500 individuals were pensioners aged between ages 70 and 80. Then the mortality ratio, which is $510 \div 3500 = 0.1457$, is the proportion of members between ages 70 and 80 who died during the six calendar years 2007–2012. The proportion dying during a single calendar year might then be estimated by $0.1457 \div 6 = 0.0243$. It is natural to suppose that this estimates the probability of dying during a single year. Such probabilities are denoted by q ($0 \leq q \leq 1$).

As it stands, this may not be a very good or reliable estimate. It takes no account of persons who, as mentioned above, were not under observation throughout all of 2007–2012, or who passed from one age band to the next during 2007–2012. Adjustments would have to be made to allow for these, and other, anomalies. Nevertheless, this analysis of mortality ratios based on "number of lives" has been very common in actuarial work, perhaps motivated by the fact that the probabilities being estimated are precisely the probabilities of the life table.

The alternative, which we will advocate in this book, is to use the time lived as the denominator. In detail, for each individual we record the time at which they entered an age group and the time when they left it, and the difference is the *survival time* during which they were alive and in that age group. Then the sum of all survival times in an age group is the total lime lived, shown in the second column of Table 1.1. Analysis based on time lived has certain advantages. Potentially important from a statistical point of view is that it avoids losing information on who died and when. We will illustrate this in the following example adapted from Richards (2008).

Consider two small groups of pension scheme members, A and B, each with four lives. Over the course of a full calendar year one life dies in each group. The proportion dying is the same in each group: $\hat{q}_A = \hat{q}_B = 1/4$ (we use the circumflex to denote an estimate of some true-but-unknown quantity; thus, $\hat{q}$ is an estimate of q). Analysis of the proportion dying does not distinguish between the mortality experience of groups A and B.

Let us denote mortality ratios based on time lived by m. Suppose that the death in group A occurred at the end of January. The total time lived in group A was therefore $3\frac{1}{12}$ years ($= 1+1+1+\frac{1}{12}$), and the ratio of deaths to time lived is thus $\hat{m}_A = 1 \div 3\frac{1}{12} = \frac{12}{37}$. In contrast, suppose that the death in group B occurred at the end of November. Then the total time lived in group B was $3\frac{11}{12}$ years ($= 1+1+1+\frac{11}{12}$), and the mortality ratio for group B is $\hat{m}_B = 1 \div 3\frac{11}{12} = \frac{12}{47}$. Thus, using the time lived as the denominator enables us to distinguish a genuine difference between the two mortality experiences. Using the number of lives leads us to overlook this difference; the information on the time actually lived is discarded.

We do not need to worry if we need q-type probabilities (that is, a life table) for specific kinds of work. As we will see later, we can derive any actuarial quantity we need having estimated m-type mortality ratios.

Let us develop the example further. Suppose that in group A one of the three surviving individuals leaves the scheme at the end of August. The reason might be resigning from employment (if an active member accruing benefits), or a trivial commutation (if a pensioner member). Using the number of lives, we now have a major problem in calculating $\hat{q}_A$, because we will not know if the departed individual dies or not in the last third of the year. If they did, then we should have $\hat{q}_A = 2/4$; if they did not, then we should have $\hat{q}_A = 1/4$, but we do not know. We will be forced to complicate our analysis on a number-of-lives basis by making some additional assumptions. Unfortunately, the assumptions which are easiest to implement are seldom justified in practice. In contrast, using time lived, the adjustment is trivial and no further assumptions are required.

The total time lived is now simply $2\frac{3}{4}$ years ($= 1+1+\frac{8}{12}+\frac{1}{12}$), and the mortality ratio is $\hat{m}_A = 1 \div 2\frac{3}{4} = \frac{4}{11}$.

This example exhibits the other advantage of using time lived instead of the number of lives – it is better able to handle real-world data where individuals enter and leave observation for various reasons, at times that are not under the control of the analyst.

The mortality ratio q is referred to as the *initial rate of mortality*, while m is referred to as the *central rate of mortality* (see Section 3.6). When used in the denominator, the number of lives is called the *initial exposed-to-risk* (sometimes denoted by E), while the time lived is called the *central exposed-to-risk* (sometimes denoted by E^c). Having set out some reasons for preferring mortality ratios based on time lived, the next section demonstrates how to fit a model to grouped counts.

1.5 Fitting a Model to Grouped Counts

One recurring feature in this book is that quantities closely related to Poisson random variables and, later on, Poisson processes arise naturally in survival models. Why this is so will ultimately be explained in Chapter 17, but for now we shall just accept that the data in Table 1.1 appear to be suitable for modelling as a Poisson random variable from age band (30, 40] upwards. For reasons we explain in Section 1.6, we exclude data below age 30 as having too few observed deaths.

We can build a statistical model for the data in Table 1.1 in just four R commands:

```
vExposures = c(121.7, 893.1, 5079.3, 32546.7, 21155.9,
               10606.7, 1751.5, 23.1)
vDeaths = c(3, 6, 48, 278, 510, 866, 363, 11)
oModelOne = glm(vDeaths ~ 1, offset=log(vExposures),
                family=poisson)
summary(oModelOne)
```

We shall explain what each of these four commands does.

- We first put the times lived and deaths into two separate vectors of equal length. The R function `c()` concatenates objects (here scalar values) into a vector. It can be handy to begin the variable names with a `v` as a reminder that they are vectors, not scalars.

- We next fit the Poisson model as a generalised linear model (GLM; see Section 10.7) using R's `glm()` function. We specify the deaths as the response variable, and we have to provide the exposures as an offset. We also specify a distribution for the response variable with the `family` argument. The results of the model are placed in the new model object, `oModelOne`. It can be handy to begin such variable names with an `o` as a reminder that it is a complex object, rather than a simple scalar or vector.
- Last, we inspect the model object using R's `summary()` function.

Part of what we will see in the output is the following:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.5444     0.0219  -161.8   <2e-16 ***
```

The above model has fitted a single constant parameter applying across all ages, which R calls the "intercept". The default behaviour in R is to operate on a logarithmic scale, so the parameter labelled `(Intercept)` is $\log \hat{m}$. Thus, the fitted model is $\hat{m} = \exp(-3.5444) = 0.02889$. This is simply the mortality ratio: the total number of deaths (2,085) divided by the total time lived (72,178.0 years). What this tells us is that the mortality ratio is not just an intuitive measure of mortality, but emerges as the estimate in a probabilistic model. In this context it has sampling properties, such as the standard error and the p-value, and these are provided automatically in the R output.

We can do better. Table 1.1 suggests that mortality ratios increase sharply with increasing age. A better model would therefore allow the Poisson parameter to vary by each age group. We can do this by running the following R commands:

```
vExposures = c(121.7, 893.1, 5079.3, 32546.7, 21155.9,
               10606.7, 1751.5, 23.1)
vDeaths = c(3, 6, 48, 278, 510, 866, 363, 11)
vAgeBand = factor(c(35, 45, 55, 65, 75, 85, 95, 105))
oModelTwo = glm(vDeaths ~ -1 + vAgeBand,
                offset=log(vExposures), family=poisson)
summary(oModelTwo)
```

We shall explain the two new features:

- The age bands are labelled with the age at the mid-point of each band. The `factor()` command ensures that the age bands are to be treated as factor levels, rather than values for regression. In other words, a mortality rate will be fitted to each age band separately.